Επιχειρηματική Ευφυΐα & Εξόρυξη Δεδομένων

Ευστάθιος Γ. Κύρκος

Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα
www.kallipos.gr
HEALLINK
Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών
Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

ΕΥΣΤΑΘΙΟΣ Γ. ΚΥΡΚΟΣ
Αναπληρωτής Καθηγητής ΑΤΕΙΘ

Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων

Ανακάλυψη Γνώσης για Λήψη Επιχειρηματικών Αποφάσεων

Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα
www.kallipos.gr

Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων

Συγγραφή

Ευστάθιος Γ. Κύρκος

Κριτικός αναγνώστης

Παναγιώτης Συμεωνίδης

Συντελεστές έκδοσης

Τεχνική Επεξεργασία: Σπυρίδων Παπαβασιλείου

ISBN: 978-960-603-109-0

ΣΥΝΔΕΣΜΟΣ ΕΛΛΗΝΙΚΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΒΙΒΛΙΟΘΗΚΩΝ

Εθνικό Μετσόβιο Πολυτεχνείο

Ηρώων Πολυτεχνείου 9, 15780 Ζωγράφου

www.kallipos.gr

Στον πατέρα μου Γιώργο

Περιεχόμενα

Πίνακας συντομεύσεων-ακρωνύμια 14

Εισαγωγή 16

1 Εισαγωγή στην Επιχειρηματική Ευφυΐα 23

1.1 Η Επιχειρηματική Ευφυΐα 23

1.2 Γιατί Επιχειρηματική Ευφυΐα; 25

1.2.1 Λήψη Επιχειρηματικών Αποφάσεων σε συνθήκες αβεβαιότητας 25

1.2.2 Οι προκλήσεις της παγκοσμιοποίησης 26

1.2.3 Η οικονομική κρίση και οι νέες κανονιστικές διατάξεις 26

1.2.4 Διαθεσιμότητα δεδομένων 27

1.2.5 Νέες τεχνολογίες και μέθοδοι ανάλυσης 27

1.3 Δομικά Επίπεδα Συστημάτων Επιχειρηματικής Ευφυΐας 28

1.3.1 Πηγές Δεδομένων 28

1.3.2 Αποθήκες Δεδομένων 28

1.3.3 Διερεύνηση Δεδομένων 29

1.3.4 Εξόρυξη Δεδομένων 29

1.3.5 Βελτιστοποίηση 30

1.3.6 Λήψη απόφασης 30

1.4 Οφέλη και Περιορισμοί της Επιχειρηματικής Ευφυΐας 30

1.4.1 Οφέλη της Επιχειρηματικής Ευφυΐας 30

1.4.2 Περιορισμοί της Επιχειρηματικής Ευφυΐας 31

1.5 Η Επιχειρηματική Ευφυΐα στην Πράξη 32

1.5.1 Διοίκηση Επιχειρησιακής Απόδοσης 32

1.5.2 Χρηματοοικονομική ανάλυση και διαχείριση 32

1.5.3 Πωλήσεις 33

1.5.4 Marketing 33

1.5.5 Διαχείριση Εφοδιαστικής Αλυσίδας. 33

1.5.6 Διαχείριση Ανθρωπίνων Πόρων 33

1.5.7 Χρηματοπιστωτικός τομέας 34

1.6 Πάροχοι λογισμικού και υπηρεσιών Επιχειρηματικής Ευφυΐας 34

1.6.1 SAS 34

1.6.2 IBM 35

1.6.3 ORACLE 35

1.6.4 SAP 36

1.6.5 Microsoft 36

1.6.6 Qlik 37

Βιβλιογραφία/Αναφορές 38

2 Συστήματα Υποστήριξης Αποφάσεων 39
2.1 Λήψη Αποφάσεων 40
2.1.1 Λογικές Αποφάσεις 40
2.1.2 Φάσεις στη Λήψη Αποφάσεων 41
2.1.3 Είδη Αποφάσεων 42
2.1.4 Διοικητικά Στελέχη και Λήψη Αποφάσεων 43
2.1.5 Λήψη αποφάσεων και πληροφοριακά συστήματα. 44
2.2 Συστήματα Υποστήριξης Αποφάσεων 46
2.2.1 Ορισμός. 46
2.2.2 Ειδικά χαρακτηριστικά και χρησιμότητα των ΣΥΑ 46
2.2.3 Σύγκριση Πληροφοριακών Συστημάτων Διοίκησης και Συστημάτων Υποστήριξης Αποφάσεων 47
2.2.4 Δομή ΣΥΑ 49
2.2.5 Σύστημα Διαχείρισης Βάσης Μοντέλων 50
2.2.6 Συστήματα Υποστήριξης Ομαδικών Αποφάσεων 50
2.2.7 Συστήματα Υποστήριξης Διοίκησης 54
2.2.8 Ευφυή Συστήματα Υποστήριξης Αποφάσεων 55
Αναφορές / Βιβλιογραφία 57
3 Μοντελοποίηση Προβλημάτων 58
3.1 Η έννοια του μοντέλου 59
3.2 Κατηγορίες μοντέλων 60
3.3 Βεβαιότητα, Αβεβαιότητα, Ρίσκο. 60
3.4 Ανάλυση Αποφάσεων 61
3.4.1 Διαγράμματα Επιρροής 62
3.4.2 Πίνακες Αποφάσεων. 63
3.5 Συστατικά μέρη Μαθηματικών Μοντέλων 64
3.6 Μαθηματική Βελτιστοποίηση και Γραμμικός Προγραμματισμός. 65
3.7 Αναλύσεις what – if και αναζήτησης στόχου 67
3.8 Ανάλυση Ευαισθησίας 70
3.9 Ευρετικές Μέθοδοι - Γενετικοί Αλγόριθμοι 71
3.9.1 Γενετικοί Αλγόριθμοι 72
3.10 Προσομοίωση 74
Βιβλιογραφία / Αναφορές 76
4 Πολυδιάστατη Ανάλυση και Αποθήκες Δεδομένων 77
4.1 Εισαγωγή - Ορισμός 78
4.2 Αρχιτεκτονική Αποθήκης Δεδομένων 79
4.3 OLTP και OLAP 80
4.4 Σχεδιασμός Αποθήκης Δεδομένων 81

4.5 Πολυδιάστατο Μοντέλο Δεδομένων - Κύβοι 85
4.5.1 Κυβοειδή 86
4.6 Ιεραρχίες Εννοιών 88
4.7 Πράξεις OLAP 89
4.7.1 Συναθροιστική Άνοδος 89
4.7.2 Αναλυτική Κάθοδος 90
4.7.3 Οριζόντιος Τεμαχισμός 91
4.7.4 Κάθετος Τεμαχισμός 91
4.7.5 Περιστροφή 91
4.8 Καθοδηγούμενη Διερεύνηση 92
4.9 Πρατήρια Δεδομένων 92
4.10 Εξαγωγή Μετασχηματισμός Φόρτωση 93
4.11 Μεταδεδομένα 96
4.12 Πεδία εφαρμογής και οφέλη των Αποθηκών Δεδομένων 97
Βιβλιογραφία / Αναφορές 99
Κριτήρια Αξιολόγησης 100
Άσκηση Υπολογισμών 4.1 100
Λύση 100
Άσκηση Υπολογισμών 4.2 100
Λύση 101
Άσκηση Υπολογισμών 4.3 101
Λύση 101
5 Οπτική και Διερευνητική Ανάλυση Δεδομένων 103
5.1 Οπτικοποίηση 104
5.2 Διερευνητική Ανάλυση Δεδομένων 106
5.3 Ταξινόμηση μεθόδων οπτικοποίησης δεδομένων 107
5.4 Τεχνικές Απεικόνισης Δεδομένων 108
5.4.1 Τυπικές 109
5.4.1.1 Γραφήματα Γραμμής 109
5.4.1.2 Ραβδογράμματα 109
5.4.1.3 Γραφήματα Πίτας 110
5.4.1.4 Διαγράμματα Διασποράς 110
5.4.2 Γεωμετρικού Μετασχηματισμού 111
5.4.2.1 Πίνακας Διαγραμμάτων Διασποράς (Scatter plot Matrix) 111
5.4.2.2 Διαγράμματα Παράλληλων Συντεταγμένων 111
5.4.2.3 HyperSlice 113
5.4.3 Εικονογραφικές 113
5.4.3.1 Πρόσωπα Chernoff 113

5.4.3.2 Εικόνες Stick Figure 114
5.4.3.3 Διαγράμματα Αστέρων 115
5.4.3.4 Τεχνική Shape Coding 116
5.4.4 Τεχνικές Εικονοστοιχείων 117
5.4.4.1 Επαναληπτικών Προτύπων (Recursive Pattern) 117
5.4.4.2 Κυκλικών Τομέων 118
5.4.5 Τεχνικές Στοίβας 118
5.4.5.1 Dimensional Stacking 118
5.4.5.2 Worlds within Worlds 119
5.4.5.3 Δενδροχάρτες 120
5.5 Μελέτη περίπτωσης. Αναγνώριση απάτης με γραφικά μέσα 121
5.6 Ταμπλό 122
Βιβλιογραφία / Αναφορές 124
6 Εξόρυξη Γνώσης από Δεδομένα 126
6.1 Εισαγωγή 127
6.2 Εξόρυξη Δεδομένων - Ορισμός. 128
6.3 Στάδια της Διαδικασίας Ανακάλυψης Γνώσης 129
6.4 Εργασίες Εξόρυξης Δεδομένων 132
6.5 Η ΕΔ στη σύγχρονη επιχείρηση 134
6.5.1 Πωλήσεις και Διαφήμιση 135
6.5.2 Ηλεκτρονικό Εμπόριο 137
6.5.3 Τράπεζες 138
6.5.4 Ασφάλειες 139
6.5.5 Χρηματιστήριο 140
6.5.6 Τηλεπικοινωνίες 140
6.5.7 Λογιστική - Ελεγκτική 141
6.5.8 Εξόρυξη Κειμένου 142
6.5.9 Ανάλυση Κοινωνικών Δικτύων 143
6.5.10 Ενσωμάτωση Της Εξόρυξης Δεδομένων στις Επιχειρηματικές Διαδικασίες 145
6.5.11 Εξόρυξη Επιχειρηματικών Διαδικασιών 146
Βιβλιογραφία / Αναφορές 147
7 Προεπεξεργασία Δεδομένων 149
7.1 Η αναγκαιότητα της προεπεξεργασίας δεδομένων 150
7.2 Χαμένες Τιμές 151
7.3 Θορυβώδη Δεδομένα 154
7.4 Κανονικοποίηση 155
7.5 Κατασκευή νέων πεδίων 157
7.6 Μείωση Διαστάσεων και Επιλογή Χαρακτηριστικών 158

7.6.1 Filters 159
7.6.1.1 t Στατιστικό Τεστ (t-test) και Ανάλυση Διακύμανσης (Analysis of Variance – ANOVA) 159
7.6.1.2 Πρόσθια Επιλογή και Οπίσθια Εξάλειψη (Forward Selection and Backward Elimination) 160
7.6.1.3 Χρήση ενός αλγορίθμου ως φίλτρου για άλλον αλγόριθμο 161
7.6.1.4 Correlation-Based Feature Selection (CFS) 161
7.6.2 Wrappers 162
7.6.3 Ανάλυση Κυρίων Συνιστωσών (Principal Components Analysis) 162
7.6.4 Επιλογή Χαρακτηριστικών – Συμπεράσματα 165
7.7 Διακριτοποίηση 165
Βιβλιογραφία / Αναφορές 169
Κριτήρια Αξιολόγησης 171
Άσκηση Υπολογισμών 7.1 171
Λύση 171
Άσκηση Υπολογισμών 7.2 172
Λύση 172
Άσκηση Εφαρμογής 7.3 172
Λύση 172
Άσκηση Εφαρμογής 7.4 173
Λύση 173
8 Κανόνες Συσχέτισης 175
8.1 Εισαγωγή 176
8.2 Ορισμοί 176
8.3 Εξόρυξη Κανόνων Συσχέτισης 178
8.3.1 Εντοπισμός συχνών στοιχειοσυνόλων – Ο αλγόριθμος Apriori 178
8.3.2 Δημιουργία Κανόνων Συσχέτισης από τα συχνά στοιχειοσύνολα 182
8.4 Πρόσθετα κριτήρια αποτίμησης των κανόνων. 183
8.5 Εξόρυξη Πολυδιάστατων Κανόνων Συσχέτισης 184
8.6 Εξόρυξη Κανόνων Συσχέτισης με διαφορετικά επίπεδα γενίκευσης 185
8.7 Κανόνες Συσχέτισης με πεδία συνεχών τιμών 186
8.8 Εξόρυξη Κανόνων Συσχέτισης βασισμένη σε περιορισμούς 188
8.9 Μελέτη Περίπτωσης – Μοντελοποίηση Αποφάσεων Εξωτερικών Ελεγκτών με χρήση Κανόνων Συσχέτισης 190
Βιβλιογραφία/Αναφορές 195
Κριτήρια Αξιολόγησης 197
Άσκηση Υπολογισμών 8.1 197
Άσκηση Υπολογισμών 8.2 197
Άσκηση Υπολογισμών 8.3 198
Άσκηση Εφαρμογής 8.4 198

9 Κατηγοριοποίηση 201
9.1 Εισαγωγή 202
9.2 Επαγωγικοί Αλγόριθμοι και Μοντέλα 203
9.3 Στάδια κατηγοριοποίησης 203
9.4 Υπερπροσαρμογή μοντέλων 204
9.5 Κριτήρια αξιολόγησης μεθόδων κατηγοριοποίησης 206
9.6 Προεπεξεργασία δεδομένων για κατηγοριοποίηση 207
9.7 Δένδρα Αποφάσεων 208
9.7.1 Εισαγωγή στα Δένδρα Αποφάσεων 208
9.7.2 Δένδρα Αποφάσεων ID3 209
9.7.3 Δένδρα Αποφάσεων C4.5 211
9.7.4 Δένδρα CART 211
9.7.5 Κλάδεμα 212
9.7.6 Δημιουργία κανόνων από Δένδρα Αποφάσεων 212
9.7.7 Πλεονεκτήματα και Μειονεκτήματα των Δένδρων Αποφάσεων 212
9.8 Κατηγοριοποίηση με Νευρωνικά Δίκτυα 213
9.8.1 Νευρώνες και Συνδέσεις 213
9.8.2 Δομή MLP 214
9.8.3 Εκπαίδευση Δικτύου 215
9.8.4 Θέματα μοντελοποίησης με νευρωνικά δίκτυα 217
9.8.5 Πλεονεκτήματα και μειονεκτήματα των Νευρωνικών Δικτύων 217
9.9 Μπαϋεσιανοί Κατηγοριοποιητές 218
9.9.1 Αφελείς Μπαϋεσιανοί Κατηγοριοποιητές 218
9.9.2 Μπαϋεσιανά Δίκτυα 219
9.9.3 Πλεονεκτήματα και μειονεκτήματα των Μπαϋεσιανών Δικτύων 221
9.10 Μελέτη περίπτωσης – Εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων με χρήση μεθόδων κατηγοριοποίησης. 221
Βιβλιογραφία / Αναφορές 225
Κριτήρια Αξιολόγησης 227
Άσκηση Υπολογισμών 9.1 227
Λύση 227
Άσκηση Υπολογισμών 9.2 228
Λύση 228
Άσκηση Εφαρμογής 9.3 228
Λύση 229
Άσκηση Εφαρμογής 9.4 229
Λύση 230
10 Εναλλακτικές Μέθοδοι και ειδικά θέματα Κατηγοριοποίησης 231

10.1 Μηχανές Διανυσμάτων Υποστήριξης 232
10.2 k-Πλησιέστεροι Γείτονες 236
10.3 Παλινδρόμηση 238
10.3.1 Απλή Γραμμική Παλινδρόμηση 238
10.3.2 Πολλαπλή Γραμμική Παλινδρόμηση 240
10.3.3 Πολυωνυμική Παλινδρόμηση 241
10.3.4 Λογιστική (ή Λογαριθμική) Παλινδρόμηση 242
10.4 Σύνθετοι Κατηγοριοποιητές 243
10.5 Επικύρωση Κατηγοριοποιητών 246
10.6 Ανισοκατανομή κλάσεων και κόστος σφάλματος 247
10.7 Επιδόσεις ανά κλάση 249
10.8 Καμπύλες ROC 250
10.9 Μελέτη Περίπτωσης – Πρόβλεψη τύπου εξωτερικού ελεγκτή με χρήση μεθόδων κατηγοριοποίησης 252
Βιβλιογραφία/Αναφορές 255
Κριτήρια Αξιολόγησης 257
Άσκηση Υπολογισμών 10.1 257
Λύση 257
Άσκηση Υπολογισμών 10.2 257
Λύση 257
Άσκηση Εφαρμογής 10.3 258
Λύση 258
Άσκηση Εφαρμογής 10.4 258
Λύση 259
Άσκηση Εφαρμογής 10.5 259
Λύση 260
11 Ανάλυση Συστάδων 261
11.1 Εισαγωγή 262
11.2 Ομοιότητα και απόσταση 262
11.2.1 Απόσταση με αριθμητικά γνωρίσματα 263
11.2.2 Απόσταση με δυαδικά γνωρίσματα 264
11.2.3 Απόσταση με ονομαστικά γνωρίσματα 265
11.2.4 Απόσταση με διατακτικά γνωρίσματα 266
11.2.5 Απόσταση με μεικτών τύπων γνωρίσματα 267
11.3 Κατηγορίες Μεθόδων ΑΣ 267
11.4 Ιεραρχική Ανάλυση Συστάδων 268
11.4.1 Δενδρογράμματα 269
11.4.2 Ιεραρχική Συσσωρευτική Ανάλυση Συστάδων 270

11.4.3 Απλή Σύνδεση....270
11.4.4 Πλήρης Σύνδεση....271
11.4.5 Σύνδεση Μέσου Όρου....271
11.4.6 Απόσταση Μέσων Σημείων (centroids)....272
11.4.7 Μέθοδος Ward....272
11.5 Διαχωριστική Ανάλυση Συστάδων....273
11.5.1 Η μέθοδος k-Means....274
11.5.2 Λοιποί αλγόριθμοι Διαιρετικής Ανάλυσης Συστάδων....276
11.5.2.1 k-Medoids....276
11.5.2.2 CLARA....276
11.6 Αυτοοργανούμενοι Χάρτες....277
11.6.1 Δομή AOX....277
11.6.2 Εκπαίδευση AOX....278
11.7 Επιχειρηματικές Εφαρμογές της Ανάλυσης Συστάδων....280
Βιβλιογραφία/Αναφορές....282
Κριτήρια Αξιολόγησης....284
Άσκηση Υπολογισμών 11.1....284
Λύση....284
Άσκηση Εφαρμογής 11.2....285
Λύση....285
Άσκηση Εφαρμογής 11.3....286
Λύση....286
Άσκηση Εφαρμογής 11.4....287
Λύση....287
12 Διαχείριση Έργων Επιχειρηματικής Ευφυΐας....288
12.1 Ανάπτυξη Συστημάτων Επιχειρηματικής Ευφυΐας....289
12.2 Ο Κύκλος Ζωής Ανάπτυξης Συστήματος Επιχειρηματικής Ευφυΐας.....289
12.2.1 Αιτιολόγηση Έργου....291
12.2.2 Οργάνωση Έργου....292
12.2.3 Ανάλυση απαιτήσεων του έργου....293
12.2.4 Σχεδιασμός....296
12.2.5 Υλοποίηση....299
12.2.6 Εφαρμογή....300
12.2.7 Αξιολόγηση....301
12.2.8 Η Επιχειρηματική Ευφυΐα Ως Υπηρεσία....302
12.3 Παράγοντες Επιτυχίας σε Έργα Επιχειρηματικής Ευφυΐας....303
Βιβλιογραφία/Αναφορές....310
13 Οδηγός WEKA....312

13.1 Εισαγωγή **313**
13.2 Προεπεξεργασία **315**
13.3 Κατηγοριοποίηση **319**
13.4 Ανάλυση Συστάδων **323**
13.5 Κανόνες Συσχέτισης **326**
13.6 Επιλογή Χαρακτηριστικών **328**
13.8 Οπτικοποίηση **329**
13.9 Άλλες πηγές για το WEKA **331**
13.10 Ελεύθερα λογισμικά Επιχειρηματικής Ευφυΐας και Εξόρυξης Δεδομένων **332**
13.11 Πηγές για ελεύθερα σύνολα δεδομένων **333**
Βιβλιογραφία/Αναφορές **335**
Πίνακας Όρων **336**

Πίνακας συντομεύσεων-ακρωνύμια

Ακρωνύμιο	Ανάλυση
ACFE	Association of Certified Fraud Examiners
AICPA	American Institute of Certified Public Accountants
AUC	Area Under Curve
BPM	Business Process Mining
CPM	Corporate Performance Management
CPT	Conditional Probability Table
CRM	Customer Relationship Management
CWM	Common Warehouse Metamodel
DM	Data Mining
DSA	Data Staging Area
DW	Data Warehouse
EDA	Exploratory Data Analysis
ERP	Enterprise Resources Planning
ESS	Executive Support Systems
ETL	Extract, Transform, Load
FAME	Financial Analysis Made Easy
GDSS	Group Decision Support Systems
GNP	Genetic Network Programming
GPS	Global Positioning System
GSS	Group Support Systems
IBC	Instance Based Classifiers
IDSS	Intelligent Decision Support Systems
KDD	Knowledge Discovery in Databases
k-NN	k-Nearest Neighbors
KPI	Key Performance Indicators
MDL	Model Definition Language
OIM	Open Information Model
OLAP	On Line Analytical Processing
OLTP	On Line Transaction Processing
PCA	Principal Components Analysis
PCF	Process Classification Framework
PSO	Particle Swarm Optimization
RBF	Radial Base Function
RFID	Radio Frequency Identification
ROC	Receiver Operating Characteristics
ROI	Return On Investment
ROTA	Return on Total Assets
SAS	Statement of Auditing Standards
SCM	Supply Chain Management Systems
SDLC	System Development Life Cycle
SOM	Self Organizing Maps
SVM	Support Vector Machines
SVR	Support Vector Regression
VLDB	Very Large Data Base
ΑΓΒΔ	Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων

ΑΔ	Αποθήκες Δεδομένων
ΑΚΣ	Ανάλυση Κυρίων Συνιστωσών
ΑΜΚ	Αφελείς Μπαϋεσιανοί Κατηγοριοποιητές
ΑΟΧ	Αυτοοργανούμενοι Χάρτες
ΑΣ	Ανάλυση Συστάδων
ΓΑ	Γενετικοί Αλγόριθμοι
ΓΠ	Γραμμικός Προγραμματισμός
ΔΑΔ	Διερευνητική Ανάλυση Δεδομένων
ΔΕ	Διαγράμματα Επιρροής
ΔΕΑ	Διοίκηση Επιχειρησιακής Απόδοσης
ΕΔ	Εξόρυξη Δεδομένων
ΕΕ	Επιχειρηματική Ευφυΐα
ΕΕΩΥ	Επιχειρηματική Ευφυΐα Ως Υπηρεσία
ΕΜ	Ευρετικές Μέθοδοι
ΕΜΦ	Εξαγωγή, Μετασχηματισμός, Φόρτωση
ΕΣΥΑ	Ευφυή Συστήματα Υποστήριξης Αποφάσεων
ΚΓΜ	Καθοδηγούμενο από τα Γεγονότα Μάρκετινγκ
ΚΔΕ	Κύριοι Δείκτες Επίδοσης
ΚΖΑΣ	Κύκλος Ζωής Ανάπτυξης Συστήματος
ΚΠ	Κέρδος Πληροφορίας
ΚΠΕ	Κρίσιμοι Παράγοντες Επιτυχίας
ΜΔΥ	Μηχανές Διανυσμάτων Υποστήριξης
ΠΔ	Πρατήρια Δεδομένων
ΠΣΔ	Πληροφοριακά Συστήματα Διοίκησης
ΣΒΔ	Σχεσιακές Βάσεις Δεδομένων
ΣΔΒΜ	Σύστημα Διαχείρισης Βάσης Μοντέλων
ΣΕΕ	Συστήματα Επιχειρηματικής Ευφυΐας
ΣΥΑ	Συστήματα Υποστήριξης Αποφάσεων
ΣΥΔ	Συστήματα Υποστήριξης Διοίκησης
ΣΥΟ	Συστήματα Υποστήριξης Ομάδων
ΣΥΟΑ	Συστήματα Υποστήριξης Ομαδικών Αποφάσεων
ΥΔ	Υποσύστημα Μεταδεδομένων

Εισαγωγή

Η λήψη αποφάσεων είναι μια από τις σημαντικότερες ευθύνες των διοικητικών στελεχών μιας επιχείρησης. Το γεγονός αυτό αναγνωρίζεται ρητά στη σύγχρονη επιστήμη της διοίκησης επιχειρήσεων. Σύμφωνα με τον Mintzberg, η λήψη αποφάσεων είναι ένας από τους τρεις βασικούς ρόλους που επιτελεί η διοίκηση ενός οργανισμού. Οι επιχειρηματικές αποφάσεις ανήκουν στην κατηγορία των λεγόμενων «λογικών αποφάσεων». Ωστόσο, όπως κατέδειξε ο Simon, οι ανθρώπινες αποφάσεις είναι μερικώς λογικές λόγω υπαρκτών περιορισμών. Η μερικότητα της πληροφόρησης, οι ατέλειες των μεθόδων επεξεργασίας, οι περιορισμένες ανθρώπινες αντιληπτικές δυνατότητες και η χρονική πίεση θέτουν όρια στην ικανότητα των επιχειρηματικών στελεχών να λάβουν λογικές αποφάσεις. Ταυτόχρονα, οι επιχειρηματικές αποφάσεις χαρακτηρίζονται από ορισμένο βαθμό αβεβαιότητας. Συνήθως ο βαθμός αβεβαιότητας αυξάνεται καθώς μεταβαίνουμε από το λειτουργικό προς το τακτικό και το στρατηγικό επίπεδο. Εδικά στο στρατηγικό επίπεδο, η περιπλοκότητα των συνθηκών και το πλήθος των πιθανών λύσεων επιβάλλει τη λήψη ως επί το πλείστον αδόμητων αποφάσεων.

Από τα παραπάνω καθίσταται σαφές ότι η λήψη επιχειρηματικών αποφάσεων είναι ένα δύσκολο καθήκον. Στη σημερινή εποχή όμως ο βαθμός δυσκολίας αυξάνεται περαιτέρω λόγω των ειδικών συνθηκών οι οποίες επικρατούν στο σύγχρονο επιχειρηματικό περιβάλλον. Η παγκοσμιοποίηση προκάλεσε την ανάπτυξη παγκόσμιων αγορών. Οι επιχειρήσεις πλέον είναι υποχρεωμένες να δραστηριοποιούνται και να ανταγωνίζονται σε παγκόσμια κλίμακα. Με τον περιορισμό των συνοριακών δασμών και την κατάργηση προστατευτικών μέτρων νέοι ανταγωνιστές εισέρχονται σε εγχώριες αγορές. Το αποτέλεσμα είναι η αύξηση του ανταγωνισμού τόσο ποσοτικά όσο και ποιοτικά. Η απορρύθμιση κανονιστικών διατάξεων επιτρέπει στις επιχειρήσεις μεγαλύτερη ευελιξία κινήσεων, με αποτέλεσμα να αυξάνεται το πλήθος των εναλλακτικών λύσεων. Η γεωγραφική διασπορά των επιχειρήσεων αυξάνει την πολυπλοκότητα τους και καθιστά δυσκολότερη την παρακολούθηση και τη διοίκηση τους. Ο ρυθμός λειτουργίας έχει εντατικοποιηθεί, με αποτέλεσμα οι αποφάσεις να λαμβάνονται υπό την πίεση του χρόνου. Επιπροσθέτως, η πρόσφατη οικονομική κρίση, η οποία έχει προκαλέσει τη χρεοκοπία επιχειρήσεων και την επιβράδυνση του ρυθμού ανάπτυξης των οικονομιών, συμβάλλει στην αύξηση των δυσκολιών αλλά και των προκλήσεων. Οι παραπάνω παράγοντες συνθέτουν ένα επιχειρηματικό περιβάλλον ιδιαίτερα περίπλοκο και ταχέως μεταβαλλόμενο. Σε αυτές τις συνθήκες η αναβάθμιση των διοικητικών πρακτικών και η βελτίωση των διαδικασιών λήψης αποφάσεων αποτελεί αδήριτη ανάγκη.

Τα παλαιότερα χρόνια η λήψη αποφάσεων θεωρούνταν περισσότερο ως μια τέχνη, ως ένα σύνολο προσωπικών ικανοτήτων που αναπτύχθηκαν μέσω της εμπειρίας με την πάροδο του χρόνου. Στη σημερινή εποχή η προσέγγιση αυτή δεν επαρκεί. Τα σύγχρονα διοικητικά στελέχη, πέρα από τις προσωπικές τους ικανότητες, οφείλουν να αξιοποιούν τις δυνατότητες που τους προσφέρουν οι νέες τεχνολογίες. Ειδικότερα, οι τεχνολογίες της πληροφορικής, οι οποίες έχουν εφαρμοστεί πλέον σε ευρύτατο βαθμό στις επιχειρήσεις, παρέχουν πρωτόγνωρες δυνατότητες για την ανάκτηση πληροφοριών, καθώς και για την επεξεργασία τους και την εξαγωγή συμπερασμάτων. Η παροχή κατάλληλης πληροφόρησης αποτελεί καθοριστικό παράγοντα για τη λήψη επιτυχημένων αποφάσεων. Κατάλληλη πληροφόρηση σημαίνει ότι δίνεται η σωστή πληροφορία στο σωστό άτομο την αναγκαία χρονική στιγμή. Όπως αναφέρθηκε και προηγουμένως, η μερικότητα της πληροφόρησης αποτελεί έναν από τους βασικούς ανασχετικούς παράγοντες στη λήψη λογικών αποφάσεων. Με την παροχή της πληρέστερης δυνατής πληροφόρησης αυτός ο ανασχετικός παράγοντας περιορίζεται. Επίσης, η παροχή αυξημένης πληροφόρησης οδηγεί στην καλύτερη κατανόηση του προβλήματος και κατ' επέκταση στη μείωση της αβεβαιότητας και στον περιορισμό του ρίσκου. Στον σύγχρονο επιχειρηματικό κόσμο οι ρυθμοί λειτουργίας έχουν επιταχυνθεί. Τα στελέχη των επιχειρήσεων εργάζονται υπό συνεχή πίεση χρόνου. Για τον λόγο αυτό, απαιτούν ποιοτική πληροφόρηση την κατάλληλη χρονική στιγμή. Η παροχή πλήρους και έγκαιρης πληροφόρησης έχει ως συνέπεια τη βελτίωση των αποφάσεων. Βελτιωμένες αποφάσεις και κατ' επέκταση βελτιωμένο μάνατζμεντ μπορούν να αυξήσουν τις επιδόσεις της επιχείρησης και να της εξασφαλίσουν το ανταγωνιστικό πλεονέκτημα.

Τα σύγχρονα διοικητικά στελέχη, στην προσπάθεια τους να αντλήσουν πληροφόρηση και να αναλύσουν τα δεδομένα, χρησιμοποιούν πληροφοριακά συστήματα. Οι σημερινές επιχειρήσεις διαθέτουν πληροφοριακά συστήματα παρακολούθησης συναλλαγών, όπως συστήματα Σχεδιασμού Επιχειρησιακών Πόρων, συστήματα Διαχείρισης Εφοδιαστικής Αλυσίδας και συστήματα Διαχείρισης Σχέσεων Πελατών. Όλα αυτά τα συστήματα καταγράφουν καθημερινά σε σχεσιακές βάσεις τεράστιους όγκους δεδομένων που αφορούν τις δραστηριότητες της επιχείρησης. Τα συστήματα παρακολούθησης συναλλαγών, τα οποία έχουν καταγεγραμμένες όλες τις συναλλαγές της επιχείρησης, αποτελούν την κύρια πηγή δεδομένων. Όποιος αναζητά εσωτερική πληροφόρηση με τη μέγιστη δυνατή λεπτομέρεια, θα πρέπει να ανατρέξει σε αυτά. Πέραν όμως των συστημάτων

παρακολούθησης συναλλαγών, υπάρχουν πρόσθετες πηγές δεδομένων. Οι επιχειρησιακοί διαδικτυακοί σέρβερς καταγράφουν πληροφορίες όπως σχόλια πελατών για τα προϊόντα της επιχείρησης και το ρεύμα κλικ των επισκεπτών της ιστοθέσης. Επίσης, πολύτιμες πληροφορίες προέρχονται από εξωτερικές πηγές. Τρίτοι φορείς, όπως κρατικές υπηρεσίες, μέσα ενημέρωσης, τράπεζες και άλλες επιχειρήσεις, μπορεί να προσφέρουν σημαντική πληροφόρηση. Μια πρόσθετη και διαρκώς αυξανόμενη δεξαμενή δεδομένων είναι το Web 2.0. Ιστοθέσεις κοινωνικής δικτύωσης, blogs, wikis και γενικώς ιστοθέσεις, το περιεχόμενο των οποίων παράγεται από τους χρήστες του διαδικτύου, επιτρέπουν την ελεύθερη έκφραση των ανθρώπων και την καταγραφή των απόψεων τους. Κατάλληλη επεξεργασία των στοιχείων αυτών μπορεί να αποκαλύψει καταναλωτικές τάσεις και επιχειρηματικές ευκαιρίες.

Από τα παραπάνω, καθίσταται σαφές ότι η σύγχρονη επιχείρηση έχει στη διάθεση της μια εξαιρετικά μεγάλη πληθώρα, αλλά και ποικιλία δεδομένων. Τα δεδομένα αυτά, με κατάλληλη επεξεργασία, μπορούν να αποτελέσουν μια πολύτιμη πηγή πληροφόρησης, αναγκαία για τη λήψη βελτιωμένων αποφάσεων. Ωστόσο, τα δεδομένα αυτά, τα οποία προέρχονται από μια πανσπερμία διαφορετικών πληροφοριακών συστημάτων, είναι ακατάλληλα για τελική επεξεργασία και εξαγωγή συμπερασμάτων. Καταρχάς, τα δεδομένα είναι διασκορπισμένα και αποθηκευμένα σε διαφορετικά πηγαία πληροφοριακά συστήματα. Η αποτελεσματική ανάκτηση των δεδομένων από τα πηγαία συστήματα κατά τη διάρκεια της τελικής επεξεργασίας, κατά κανόνα, δεν είναι εφικτή. Για τον λόγο αυτό, απαιτείται η εκ των προτέρων συγκέντρωση και αποθήκευση τους σε ένα ενιαίο πληροφοριακό σύστημα. Κατά δεύτερον, τα πηγαία δεδομένα χαρακτηρίζονται από μια σειρά προβλημάτων. Ένα πολύ συνηθισμένο πρόβλημα είναι η ύπαρξη σφαλμάτων τα οποία, σύμφωνα με μελέτες, μπορεί να ανέρχονται στο ύψος του 5%. Ένα άλλο πολύ συνηθισμένο πρόβλημα είναι η ύπαρξη χαμένων τιμών, τιμών δηλαδή οι οποίες απουσιάζουν από τις βάσεις δεδομένων. Σύνηθες φαινόμενο είναι και η ύπαρξη ασυνεπειών. Για παράδειγμα, σε δύο διαφορετικά συστήματα μπορεί να χρησιμοποιείται διαφορετικός κωδικός για το ίδιο προϊόν ή να παρουσιάζεται διαφορετική διεύθυνση για τον ίδιο πελάτη ή να χρησιμοποιείται διαφορετικό όνομα πεδίου για την ίδια πληροφορία. Σημειώνεται ότι η ύπαρξη τέτοιων προβλημάτων είναι ο κανόνας για τα δεδομένα του πραγματικού κόσμου. Τα δεδομένα που υποφέρουν από τέτοια προβλήματα χαρακτηρίζονται «ακάθαρτα». Η χρήση ακάθαρτων δεδομένων για τη διεξαγωγή αναλύσεων μπορεί να αποπροσανατολίσει την αναλυτική διαδικασία και να οδηγήσει σε εσφαλμένα ευρήματα και συμπεράσματα. Μια αγγλική έκφραση που αποτυπώνει αυτό το γεγονός και που συναντάται συχνά στη βιβλιογραφία της Επιχειρηματικής Ευφυΐας είναι η ρήση «garbage in, garbage out». Εξαιτίας της ύπαρξης προβλημάτων, τα πηγαία δεδομένα πρέπει όχι μόνο να συγκεντρωθούν, αλλά και να υποστούν μια διαδικασία ‘καθαρισμού’, ώστε να απαλλαγούν από τα προβλήματα τους. Ένα άλλο χαρακτηριστικό των πηγαίων δεδομένων είναι ότι ο βαθμός λεπτομέρειας τους είναι υπερβολικά μεγάλος. Τα αναλυτικά στοιχεία των καθημερινών συναλλαγών της επιχείρησης είναι ογκώδη και δύσχρηστα. Κατά κανόνα, για τη διεξαγωγή αναλύσεων απαιτείται η ύπαρξη συγκεντρωτικών και ουσιωδέστερων πληροφοριών. Επίσης, σε πολλές περιπτώσεις, για τη διεξαγωγή αναλύσεων απαιτείται και η ύπαρξη ιστορικής πληροφορίας προηγούμενων ετών, ώστε να είναι δυνατή η εκτέλεση συγκρίσεων. Όμως τα δεδομένα των συστημάτων παρακολούθησης συναλλαγών έχουν βραχύ χρονικό ορίζοντα, ενώ οι συγκρίσεις σε ορισμένες περιπτώσεις μπορεί να αφορούν στοιχεία με βάθος δεκαετίας. Για όλους τους παραπάνω λόγους δεν ενδείκνυται η χρήση των δεδομένων των πηγαίων πληροφοριακών συστημάτων για τη διεξαγωγή αναλύσεων. Αντιθέτως, απαιτείται η χρήση ενοποιημένων, συγκεντρωτικών, καθαρών και ιστορικών δεδομένων, τα οποία θα είναι αποθηκευμένα σε ένα ενιαίο και ανεξάρτητο πληροφοριακό σύστημα.

Η χρήση εξειδικευμένων πληροφοριακών συστημάτων για την επεξεργασία των δεδομένων και για την υποβοήθηση της λήψης αποφάσεων δεν είναι ένα πρόσφατο φαινόμενο. Ήδη από τη δεκαετία του 1970 αναπτύχθηκαν και χρησιμοποιήθηκαν τα Συστήματα Υποστήριξης Αποφάσεων. Τα συστήματα αυτά στηριζόταν κυρίως στη χρήση μαθηματικών μοντέλων, τα οποία προέρχονταν κυρίως από τον χώρο της Επιχειρησιακής Έρευνας και της Στατιστικής. Η μοντελοποίηση οικονομικών προβλημάτων αποτέλεσε για πολλά χρόνια μια πολύ δημοφιλή τακτική. Τεχνικές όπως οι Πίνακες Αποφάσεων και η μέθοδος του Γραμμικού Προγραμματισμού εφαρμόστηκαν κατά κόρο για τη λήψη επιχειρηματικών αποφάσεων. Οι τεχνικές αυτές εξακολουθούν να χρησιμοποιούνται σε ευρεία κλίμακα και σήμερα. Με την πάροδο όμως του χρόνου και την εξέλιξη της τεχνολογίας, αναπτύχθηκαν πρόσθετες τεχνολογίες και μεθοδολογίες για την επεξεργασία των δεδομένων και τη λήψη αποφάσεων. Οι Αποθήκες Δεδομένων είναι εξειδικευμένες βάσεις δεδομένων, στις οποίες αποθηκεύονται σε συγκεντρωτική και ολοκληρωμένη μορφή τα δεδομένα των πηγαίων συστημάτων. Οι Αποθήκες Δεδομένων βρίσκονται στο επίκεντρο των σύγχρονων συστημάτων Επιχειρηματικής Ευφυΐας. Οι θεματικός προσανατολισμός τους επιτρέπει την παροχή πληροφόρησης, στοχευμένης σε συγκεκριμένα ζητήματα ενδιαφέροντος. Με τις Αποθήκες Δεδομένων σχετίζονται και τα συστήματα Αναλυτικής Επεξεργασίας Δεδομένων (On Line Analytical Processing (OLAP). Τα συστήματα OLAP επιτρέπουν στον χρήστη να προβάλει τα δεδο-

μένα με διαφορετικό τρόπο και με διαφορετικό επίπεδο γενίκευσης. Ειδικότερα, ο χρήστης έχει τη δυνατότητα να υποβάλει ελεύθερα μη προκαθορισμένες ερωτήσεις, επικεντρώνοντας σε ζητήματα που τον ενδιαφέρουν και αυξομειώνοντας τον βαθμό γενίκευσης. Βασικό χαρακτηριστικό των συστημάτων OLAP είναι η δυνατότητα ταχείας πρόσβασης σε μεγάλους όγκους δεδομένων.

Τα τελευταία χρόνια έχει αναπτυχθεί ένας νέος κλάδος της επιστήμης της Πληροφορικής, η Εξόρυξη Δεδομένων. Η Εξόρυξη Δεδομένων αποτέλεσε μια απάντηση στο πρόβλημα της υπερσυσσώρευσης δεδομένων και στην ανάγκη επεξεργασίας τους για την ανακάλυψη χρήσιμης γνώσης. Σήμερα ο ρυθμός καταγραφής και αποθήκευσης δεδομένων είναι καταιγιστικός. Σύμφωνα με στοιχεία που αναφέρονται στην ιστοθέση της IBM, καθημερινά παράγονται 2,5 πεντάκις εκατομμύρια bytes και το 90% των αποθηκευμένων δεδομένων έχουν παραχθεί την τελευταία διετία. Παράγοντες που συμβάλλουν στην έξαρση αυτού του φαινομένου είναι η παραγωγή φθηνού, εξαιρετικά ισχυρού και ποικιλόμορφου υλικού υπολογιστών και άλλων παρεμφερών συσκευών, η καθολική διάχυση και ενσωμάτωση των νέων τεχνολογιών πληροφορικής στη σύγχρονη κοινωνία, η ευρύτατη εξάπλωση και χρήση του Διαδικτύου και τέλος, η εκρηκτική ανάπτυξη του Web 2.0, δηλαδή υπηρεσιών και ιστοθέσεων, το περιεχόμενο των οποίων συντάσσεται από τους χρήστες. Ο ανθρώπινος νους έχει περιορισμένες αναλυτικές δυνατότητες, ανεπαρκείς για την αντιμετώπιση του μεγάλου όγκου των δεδομένων. Η επεξεργασία των δεδομένων αυτών χωρίς εξειδικευμένα εργαλεία, αν δεν είναι αδύνατη, είναι αργή, ακριβή και εν πολλοίς υποκειμενική. Προηγούμενοι επιστημονικοί κλάδοι, όπως η Στατιστική και η Μηχανική Μάθηση, δεν λαμβάνουν μέριμνα για την αντιμετώπιση του προβλήματος του πολύ μεγάλου όγκου των δεδομένων, ενώ ο κλάδος των Βάσεων Δεδομένων, ο οποίος είναι και ο κατ' εξοχήν αρμόδιος για την τήρηση μεγάλου όγκου δεδομένων, δεν είναι προσανατολισμένος στην ανάλυση τους. Η Εξόρυξη Δεδομένων αντλεί μεθοδολογίες από όλους τους επιστημονικούς κλάδους που αναφέρθηκαν παραπάνω, καθώς και από άλλους, όπως η Οπτικοποίηση, και στοχεύει στην ανακάλυψη όχι προφανούς και εν δυνάμει χρήσιμης γνώσης, η οποία είναι κρυμμένη στα δεδομένα. Η Εξόρυξη Δεδομένων διαφοροποιείται από προηγούμενους σχετικούς επιστημονικούς κλάδους με διάφορους τρόπους. Κατ' αρχήν, ασχολείται με την επεξεργασία μεγάλου όγκου δεδομένων, δίνοντας απαντήσεις σε σχετικά προβλήματα. Δεύτερον, ακολουθεί μια ολιστική προσέγγιση και παρέχει μεθοδολογίες για όλα τα στάδια της ανακάλυψης γνώσης, από την αρχική συγκέντρωση και προεπεξεργασία των δεδομένων μέχρι και την οπτικοποίηση των προτύπων και την τελική αξιολόγηση τους. Αντιμετωπίζονται προβλήματα όπως οι χαμένες τιμές, ο θόρυβος, ο κατάλληλος μετασχηματισμός των δεδομένων κλπ. Τρίτον, οι μέθοδοι της επεξεργασίας των δεδομένων δεν προέρχονται μόνο από τη Στατιστική. Η Εξόρυξη Δεδομένων κάνει ευρύτατη χρήση μεθόδων που προέρχονται από τη Μηχανική Μάθηση και την Αναγνώριση Προτύπων. Έρευνες έχουν αποδείξει ότι οι νέες αυτές μέθοδοι μπορούν να δώσουν καλύτερα αποτελέσματα από τις παραδοσιακές στατιστικές μεθόδους. Επίσης, η ανάλυση Κανόνων Συσχέτισης είναι μια νέα μέθοδος επεξεργασίας η οποία προέρχεται απ' ευθείας από την Εξόρυξη Δεδομένων. Τέταρτον, πολλές από τις παραπάνω μεθόδους δεν απαιτούν την εκ των προτέρων διατύπωση υποθέσεων. Αντιθέτως, τα μοντέλα προκύπτουν απευθείας από τα δεδομένα με κατάλληλη επεξεργασία. Τέλος, οι νέες μέθοδοι δίνουν τη δυνατότητα προγνωστικής ανάλυσης, δηλαδή την επεξεργασία ιστορικών στοιχείων και τη διατύπωση προβλέψεων για το μέλλον.

Η ανάγκη των επιχειρήσεων για βελτιωμένη πληροφόρηση και αναβάθμιση των διαδικασιών λήψης αποφάσεων, η διαθεσιμότητα εξαντλητικών και ποικίλου περιεχομένου δεδομένων, και η ανάπτυξη νέων τεχνολογιών και μεθοδολογιών για την ανάλυση των δεδομένων, αποτέλεσαν τα εφαλτήρια για την ανάπτυξη των συστημάτων Επιχειρηματικής Ευφυΐας. Τα συστήματα Επιχειρηματικής Ευφυΐας είναι εξειδικευμένα πληροφοριακά συστήματα, τα οποία προσφέρουν ποιοτική πληροφορία, βασισμένη σε ποιοτικά και συγκεντρωτικά δεδομένα. Τα δεδομένα συνδυάζονται με λογισμικό, που υλοποιεί και αλγορίθμους Εξόρυξης Δεδομένων, και είναι ικανό να διεξάγει υψηλού επιπέδου αναλύσεις. Η βελτίωση της ποιότητας της πληροφορίας οφείλεται στις δυνατότητες αυτών των συστημάτων, τα οποία προσφέρουν ποιοτικά δεδομένα και επιτρέπουν την ταχύτερη πρόσβαση στην πληροφορία, την ευκολότερη υποβολή ερωτημάτων στο σύστημα και τη σύνταξη αναφορών, καθώς και την προχωρημένη ανάλυση των δεδομένων. Οι τελικοί αποδέκτες του προϊόντος των συστημάτων Ε.Ε., οι οποίοι πολλές φορές αναφέρονται στην βιβλιογραφία ως «εργάτες γνώσης», τροφοδοτούνται έγκαιρα με γνώση που χρησιμοποιούν για τη λήψη αποφάσεων.

Τα συστήματα Επιχειρηματικής Ευφυΐας βρίσκονται τον τελευταίο καιρό στο επίκεντρο του ενδιαφέροντος του επιχειρηματικού κόσμου. Σύμφωνα με μελέτες που πραγματοποιούν οίκοι ερευνών και συμβουλευτικών υπηρεσιών, η Επιχειρηματική Ευφυΐα συγκαταλέγεται στις κορυφαίες θέσεις των τεχνολογικών προτεραιοτήτων των μεγαλύτερων επιχειρήσεων παγκοσμίως. Ως αποτέλεσμα του ενδιαφέροντος των επιχειρήσεων, έχει αναπτυχθεί μια αγορά σχετικών συστημάτων και λογισμικού με κύκλο εργασιών της τάξης δεκάδων δισεκατομμυρίων δολαρίων. Οι κορυφαίες επιχειρήσεις πληροφορικής, όπως η Oracle, η IBM, η Microsoft και η SAP, δραστηριοποιούνται ενεργά και πρωταγωνιστούν στον χώρο, ενώ ταυτόχρονα μια πλειάδα εξειδικευμέ-

νων επιχειρήσεων όπως η Qlik και η Tableau διεκδικούν δυναμικά σημαντικά μερίδια της νέας αυτής αγοράς. Τα συστήματα Επιχειρηματικής Ευφυΐας που προσφέρουν οι κατασκευαστές λογισμικού, επιτρέπουν στους οργανισμούς να μαθαίνουν, να αντιλαμβάνονται καταστάσεις, να σκέφτονται αφαιρετικά, να προβλέπουν τάσεις και μελλοντικά συμβάντα, να σχεδιάζουν και να καινοτομούν. Η παραγόμενη πληροφορία μετουσιώνεται σε γνώση, η οποία αξιοποιείται από τα διοικητικά στελέχη, ώστε να δρομολογηθούν κατάλληλες δράσεις, που θα οδηγήσουν στον καθορισμό και την επίτευξη επιχειρηματικών στόχων με τρόπο αποτελεσματικό και αποδοτικό. Όπως χαρακτηριστικά λέγεται, η γνώση αποτελεί το πολυτιμότερο κεφάλαιο των σύγχρονων επιχειρήσεων. Ορισμένα από τα κυριότερα πεδία εφαρμογής των μεθοδολογιών Επιχειρηματικής Ευφυΐας είναι η Διοίκηση Επιχειρησιακής Απόδοσης, η χρηματοοικονομική ανάλυση και διαχείριση, οι πωλήσεις, το μάρκετινγκ, η διαχείριση της εφοδιαστικής αλυσίδας και η διαχείριση των ανθρωπίνων πόρων. Σημαντικά και εξειδικευμένα πεδία εφαρμογής βρίσκει η Επιχειρηματική Ευφυΐα στις επιχειρήσεις του χρηματοπιστωτικού κλάδου.

Η ανάπτυξη συστημάτων Επιχειρηματικής Ευφυΐας αποτελεί μια πολύ ισχυρή και πολύ σύγχρονη τάση εφαρμογής τεχνολογιών πληροφορικής στις σημερινές επιχειρήσεις. Τα συστήματα αυτά μπορούν να παίξουν καθοριστικό ρόλο στην ποιοτική αναβάθμιση των διαδικασιών λήψης αποφάσεων. Τα σύγχρονα στελέχη επιχειρήσεων οφείλουν να είναι σε θέση να αξιοποιούν τις δυνατότητες που τους προσφέρουν αυτά τα συστήματα. Προϋπόθεση τέτοιων ικανοτήτων είναι η βαθιά γνώση των χαρακτηριστικών και του τρόπου λειτουργίας των συστημάτων Επιχειρηματικής Ευφυΐας, η κατανόηση των μεθόδων, των δυνατοτήτων και των περιορισμών τους, καθώς επίσης και η ανάπτυξη σχετικών δεξιοτήτων. Στελέχη με κατάλληλες γνώσεις και δεξιότητες μπορούν να αποτελέσουν την κινητήρια δύναμη για την ταχύτερη εφαρμογή της Επιχειρηματικής Ευφυΐας στις ελληνικές επιχειρήσεις. Το παρόν σύγγραμμα αποτελεί μια συμβολή στην προσπάθεια εκπαίδευσης επιχειρηματικών στελεχών στο αντικείμενο της Επιχειρηματικής Ευφυΐας.

Επιδίωξη του γράφοντος είναι η κάλυψη όλων των σημαντικών πλευρών και ζητημάτων που άπτονται του αντικειμένου της Επιχειρηματικής Ευφυΐας. Στόχος είναι η εκπαίδευση στελεχών ικανών να πρωτοστατήσουν στην ανάπτυξη, εφαρμογή και χρήση συστημάτων Επιχειρηματικής Ευφυΐας. Για τον λόγο αυτό, η προσέγγιση του αντικειμένου είναι πολύπλευρη και περιλαμβάνει θεωρητικά ζητήματα λήψης αποφάσεων, αλγορίθμους ανάλυσης, ζητήματα σχεδιασμού, ανάπτυξης και εφαρμογής συστημάτων Επιχειρηματικής Ευφυΐας, παρουσίαση συγκεκριμένου λογισμικού Εξόρυξης Δεδομένων κλπ. Μεγάλο μέρος του συγγράμματος ασχολείται με αλγορίθμους και μεθόδους ανάλυσης, κυρίως με σύγχρονες μεθόδους Εξόρυξης Δεδομένων. Οι μεθοδολογίες της Εξόρυξης Δεδομένων αποτελούν τα πιο σύγχρονα και περίτεχνα εργαλεία ανάλυσης δεδομένων και ανακάλυψης γνώσης. Ωστόσο, ο συγγραφέας δεν παραλείπει να συμπεριλάβει παραδοσιακές μεθόδους, όπως οι Πίνακες Αποφάσεων και ο Γραμμικός Προγραμματισμός. Η χρήση αυτών των μεθόδων για τη λήψη αποφάσεων αποτελεί πάγια και ευρύτατα διαδεδομένη τακτική στις επιχειρήσεις εδώ και δεκαετίες. Ένα άλλο χαρακτηριστικό του συγγράμματος είναι ότι δεν περιορίζεται και δεν προσαρμόζεται σε κάποιο συγκεκριμένο λογισμικό Επιχειρηματικής Ευφυΐας. Αντιθέτως, η παρουσίαση των συστημάτων, αλγορίθμων κλπ. είναι όσο το δυνατόν πιο γενικευμένη. Με τον τρόπο αυτόν, ο αναγνώστης αποκτά ένα γενικό υπόβαθρο γνώσεων, το οποίο θα του επιτρέψει να κατανοήσει εύκολα, και να χειριστεί λογισμικά συγκεκριμένων κατασκευαστών.

Το σύγγραμμα απευθύνεται πρωτίστως σε φοιτητές τμημάτων Ανώτατων Εκπαιδευτικών Ιδρυμάτων οικονομικής κατεύθυνσης. Το γεγονός αυτό επιβάλλει μια ειδική προσέγγιση του αντικειμένου. Καταρχάς επιλέχθηκε μια αρκετά εμβριθής παρουσίαση των αλγορίθμων και μεθόδων. Ο συγγραφέας συμφωνεί απολύτως με τη ρήση των Witten και Frank ότι «τα μοντέλα είναι τόσο καλά όσο είναι οι χρήστες τους». Είναι ιδιαίτερα σημαντικό να αντιλαμβάνονται οι χρήστες λογισμικών τον τρόπο λειτουργίας των αλγορίθμων και κατασκευής των μοντέλων. Η γνώση αυτή τους επιτρέπει να κατανοήσουν τις δυνατότητες και τους περιορισμούς της κάθε μεθόδου, να επιλέξουν την κατάλληλη μέθοδο για την εργασία που θέλουν να εκτελέσουν, καθώς επίσης να ρυθμίσουν αποτελεσματικά τις παραμέτρους του εκάστοτε αλγορίθμου. Ωστόσο, επιλέχθηκε να αποφευχθεί η εξαντλητική παράθεση λεπτομερειών των αλγορίθμων, η ανάπτυξη τεχνικών ζητημάτων και η συστηματική παρουσίαση των αλγορίθμων σε μορφή ψευδοκώδικα, καθώς θεωρήθηκε ότι η ενασχόληση με τέτοια θέματα προσιδιάζει περισσότερο σε καθαρόαιμα συγγράμματα πληροφορικής. Επίσης, στο Διαδίκτυο διατίθεται πλήθος αξιόλογων βιβλιοθηκών με μεθόδους Εξόρυξης Δεδομένων, όμως η χρήση τους απαιτεί ικανότητες προγραμματισμού. Το λογισμικό WEKA, το οποίο επιλέχθηκε να παρουσιαστεί στο τελευταίο κεφάλαιο, διαθέτει γραφικό περιβάλλον εργασίας, και κατά συνέπεια μπορεί να χρησιμοποιηθεί από τελικούς χρήστες, χωρίς να είναι απαραίτητη η συγγραφή κώδικα.

Το τελικό ζητούμενο είναι η εκπαίδευση στελεχών που θα έχουν ικανότητες άμεσης και πρακτικής εφαρμογής των σύγχρονων αναλυτικών μεθοδολογιών και των συστημάτων Επιχειρηματικής Ευφυΐας στη σύγχρονη επιχείρηση. Για την εξυπηρέτηση αυτού του στόχου επιστρατεύονται διάφορα συγγραφικά μέσα. Στο σύγ-

γραμμα περιλαμβάνεται μεγάλος αριθμός σχημάτων, τα οποία βοηθούν στην καλύτερη κατανόηση των θεμάτων που παρουσιάζονται. Σε πολλές περιπτώσεις, τα σχήματα αναφέρονται σε συγκεκριμένα παραδείγματα. Η χρήση παραδειγμάτων είναι αρκετά συχνή. Η παρουσίαση των μεθόδων συνοδεύεται από την παράθεση των πλεονεκτημάτων και μειονεκτημάτων τους, ώστε ο χρήστης, όταν θα κληθεί να εκτελέσει μια αναλυτική εργασία, να μπορεί να επιλέξει την πλέον κατάλληλη μέθοδο, αλλά και να αξιολογήσει τις δυνατότητες και τα όρια της. Σε πολλές περιπτώσεις παρατίθενται μελέτες περίπτωσης. Για παράδειγμα, στο κεφάλαιο 9, περιλαμβάνεται μελέτη περίπτωσης, όπου εφαρμόζονται τεχνικές κατηγοριοποίησης για τον εντοπισμό χειραγωγημένων χρηματοοικονομικών καταστάσεων. Οι μελέτες περίπτωσης εξοικειώνουν τον αναγνώστη με ρεαλιστικά και ολοκληρωμένα σενάρια εφαρμογής των αναλυτικών μεθόδων για την επίτευξη ορισμένων στόχων. Τα πρακτικά κεφάλαια περιλαμβάνουν στο τέλος τους λυμένες ασκήσεις, οι οποίες βοηθούν τον αναγνώστη να κατανοήσει καλύτερα τις μεθόδους, και να αποκτήσει δεξιότητες εφαρμογής τους. Για την απόκτηση δεξιοτήτων πρακτικής εφαρμογής των μεθόδων, μπορεί να χρησιμοποιηθεί και το λογισμικό WEKA, το οποίο παρουσιάζεται στο κεφάλαιο 13. Το WEKA είναι ένα από τα πλέον αναγνωρισμένα λογισμικά Εξόρυξης Δεδομένων και επιπλέον διατίθεται ελεύθερα. Οι αναγνώστες μπορούν να προμηθευτούν και να εγκαταστήσουν ελεύθερα το λογισμικό και να το χρησιμοποιήσουν για εξάσκηση, διεξαγωγή πειραμάτων, διεξαγωγή έρευνας κλπ. Στο τέλος του δέκατου τρίτου Κεφαλαίου, παρατίθεται κατάλογος πρόσθετων λογισμικών εξόρυξης δεδομένων, καθώς επίσης και κατάλογος δημοσίως διαθέσιμων συνόλων δεδομένων, τα οποία μπορούν να χρησιμοποιηθούν για εξάσκηση. Τέλος, ιδιαίτερης πρακτικής αξίας είναι το κεφάλαιο 12, το οποίο ασχολείται με ζητήματα ανάπτυξης συστημάτων Επιχειρηματικής Ευφυΐας. Το συγκεκριμένο κεφάλαιο μπορεί να αποτελέσει χρήσιμο βοήθημα για επιχειρήσεις και ομάδες εργασίας, οι οποίες εμπλέκονται σε έργα Επιχειρηματικής Ευφυΐας.

Μερικές χρήσιμες επισημάνσεις για τον εκπαιδευτικό, ο οποίος θα επιθυμήσει να αξιοποιήσει το παρόν σύγγραμμα ως διδακτικό βοήθημα, είναι οι ακόλουθες. Το βιβλίο είναι σχεδιασμένο για να χρησιμοποιηθεί ως κύριο σύγγραμμα για τη διδασκαλία του αντικειμένου της Επιχειρηματικής Ευφυΐας και της εφαρμογής της Εξόρυξης Δεδομένων για τη λήψη επιχειρηματικών αποφάσεων. Κατά συνέπεια επιχειρεί να καλύψει πλήρως αυτό το γνωστικό αντικείμενο. Το βιβλίο θεωρείται κατάλληλο για τη διδασκαλία μαθημάτων σε προχωρημένο προπτυχιακό ή και σε μεταπτυχιακό επίπεδο. Είναι οργανωμένο σε δεκατρία κεφάλαια, τα οποία αντιστοιχούν σε δεκατρείς διδακτικές εβδομάδες, χρονικό διάστημα τυπικό για την ολοκλήρωση του διδακτικού έργου εντός ενός ακαδημαϊκού εξαμήνου. Τα κεφάλαια έχουν σχεδιαστεί με τέτοιο τρόπο ώστε να είναι κατά το δυνατόν αυτόνομα. Κατά συνέπεια, κάθε κεφάλαιο μπορεί να διαβαστεί και να αξιοποιηθεί ξεχωριστά. Σε περίπτωση που για την κατανόηση ορισμένων θεμάτων απαιτείται προηγούμενη γνώση, αυτό επισημαίνεται στο εισαγωγικό σημείωμα του εκάστοτε κεφαλαίου και παρατίθενται παραπομπές σε άλλα κεφάλαια του βιβλίου ή σε πρόσθετες πηγές. Στα εισαγωγικά σημειώματα των κεφαλαίων παρατίθενται επίσης άλλα συγγράμματα και λοιπές πηγές για πρόσθετη μελέτη. Αναφέρθηκε και προηγουμένως ότι η παρουσίαση των μεθόδων και αλγορίθμων δεν είναι εξαντλητική, εμβαθύνει όμως σε σημαντικό βαθμό. Κατά συνέπεια, επιλεγμένα κεφάλαια μπορούν να χρησιμοποιηθούν ως συμπληρωματικό υλικό για διδασκαλία τεχνικών Εξόρυξης Δεδομένων ή Μηχανικής Μάθησης. Στο τέλος αρκετών κεφαλαίων περιλαμβάνονται λυμένες ασκήσεις. Επίσης, το βιβλίο συνοδεύεται από ηλεκτρονικές παρουσιάσεις, οι οποίες θα διευκολύνουν το εκπαιδευτικό έργο.

Για τον υποψήφιο αναγνώστη – φοιτητή θα θέλαμε να σημειώσουμε τα εξής. Επιθυμία και ελπίδα του γράφοντος είναι ότι το παρόν σύγγραμμα θα κεντρίσει το ενδιαφέρον του για το αντικείμενο της Επιχειρηματικής Ευφυΐας και της εφαρμοσμένης Εξόρυξης Δεδομένων. Καταβλήθηκε σημαντική προσπάθεια, ώστε σύνθετες έννοιες να παρουσιαστούν με απλό και κατανοητό τρόπο. Η εκτεταμένη χρήση σχημάτων διευκολύνει την κατανόηση των εννοιών και μεθόδων. Αποφεύχθηκε ηθελημένα η παράθεση υπερβολικών λεπτομερειών και η χρήση εξεζητημένων μαθηματικών. Το λογισμικό WEKA μπορεί να αποτελέσει χρησιμότατο εργαλείο για την εκπαίδευση και την εξοικείωση του με τις τεχνικές Εξόρυξης Δεδομένων, και την εφαρμογή τους για την επίτευξη συγκεκριμένων στόχων. Τα σύνολα δεδομένων που καταγράφονται στο σχετικό παράρτημα μπορούν να χρησιμοποιηθούν για την πρακτική διεξαγωγή αναλύσεων. Ο γράφων συμβουλεύει με έμφαση και ενθαρρύνει τους φοιτητές να μην περιοριστούν στα παραδείγματα και τις λυμένες ασκήσεις του συγγράμματος, αλλά να επιδιώξουν με δική τους πρωτοβουλία τη διεξαγωγή περαιτέρω αναλύσεων, με χρήση των συνόλων δεδομένων και του λογισμικού WEKA.

Η οργάνωση του συγγράμματος είναι η ακόλουθη:

Το πρώτο κεφάλαιο αποτελεί μια εισαγωγή στην Επιχειρηματική Ευφυΐα. Δίνονται οι ορισμοί της Επιχειρηματικής Ευφυΐας και των Συστημάτων Επιχειρηματικής Ευφυΐας. Ακολουθεί η παράθεση των επιχειρηματικών και τεχνολογικών παραγόντων οι οποίοι καθόρισαν την άνθηση των νέων αυτών συστημάτων. Παρουσιάζεται η πυραμίδα των συστημάτων Επιχειρηματικής Ευφυΐας και παρατίθενται τα οφέλη τους αλλά και οι περιορισμοί τους. Καταγράφονται και σχολιάζονται συνοπτικά τα κυριότερα πεδία εφαρμογής, όπως η

Διοίκηση Επιχειρησιακής Απόδοσης, η χρηματοοικονομική ανάλυση και διαχείριση κλπ. Τέλος, γίνεται παρουσίαση των σημαντικότερων παρόχων λογισμικού και υπηρεσιών Επιχειρηματικής Ευφυΐας.

Το δεύτερο κεφάλαιο καλύπτει τη θεματική ενότητα των Συστημάτων Υποστήριξης Αποφάσεων (ΣΥΑ). Αρχικά, μελετούνται θεωρητικά θέματα λήψης επιχειρηματικών αποφάσεων. Στη συνέχεια, το κεφάλαιο προχωρά στην καθαυτή παρουσίαση των ΣΥΑ. Παρατίθενται ορισμοί των ΣΥΑ, τα βασικά χαρακτηριστικά τους, καθώς και οι ομοιότητες και διαφορές τους με τα Πληροφοριακά Συστήματα Διοίκησης. Αναλύεται η αρχιτεκτονική των ΣΥΑ και σχολιάζονται τα επιμέρους υποσυστήματα τους. Επίσης, παρουσιάζονται ειδικές υποκατηγορίες των ΣΥΑ, όπως τα Συστήματα Υποστήριξης Ομαδικών Αποφάσεων και τα Συστήματα Υποστήριξης Διοίκησης.

Το τρίτο κεφάλαιο ασχολείται με το αντικείμενο της Μοντελοποίησης Προβλημάτων. Αναλύεται η έννοια του μοντέλου και σχολιάζονται οι βασικές κατηγορίες τους. Αναπτύσσονται μέθοδοι ανάλυσης αποφάσεων, όπως τα Διαγράμματα Επιρροής και οι Πίνακες Αποφάσεων. Παρουσιάζεται η μέθοδος του Γραμμικού προγραμματισμού, η οποία είναι ίσως η πλέον διαδεδομένη μέθοδος λήψης αποφάσεων στη Διοίκηση Επιχειρήσεων. Εξηγούνται οι αναλύσεις τύπου what-if και αναζήτησης στόχου, και παρατίθενται παραδείγματα στα οποία γίνεται χρήση του λογισμικού Excel. Αναφορά γίνεται και στις ευρετικές μεθόδους, με έμφαση στη μέθοδο των Γενετικών Αλγορίθμων. Επίσης, αναπτύσσονται θέματα προσομοίωσης.

Το τέταρτο κεφάλαιο αναφέρεται στις Αποθήκες Δεδομένων (ΑΔ). Παρέχεται ο ορισμός του Inmon και παρουσιάζεται η βασική αρχιτεκτονική των ΑΔ και τα συστατικά τους τμήματα. Γίνεται αναλυτική αναφορά και σύγκριση των συστημάτων OLAP και OLTP. Ακολούθως, παρατίθενται τα σχήματα των ΑΔ, δηλαδή το σχήμα Αστέρα, το σχήμα Χιονονιφάδας και το σχήμα Αστερισμού. Αναλύονται οι κύβοι δεδομένων και οι πράξεις OLAP. Εκτεταμένη αναφορά γίνεται στις εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης δεδομένων. Παρουσιάζεται το σύστημα μεταδεδομένων και σχολιάζονται οι δυνατότητες εφαρμογής των ΑΔ στη σύγχρονη επιχείρηση.

Το πέμπτο κεφάλαιο καλύπτει θέματα οπτικοποίησης και οπτικής ανάλυσης δεδομένων. Οι τεχνικές οπτικοποίησης χρησιμοποιούνται για τη διερευνητική ανάλυση των δεδομένων. Εξηγούνται οι αρχές και η μεθοδολογία της διερευνητικής ανάλυσης δεδομένων με οπτικά μέσα. Στη συνέχεια, παρουσιάζεται μεγάλος αριθμός μεθόδων γραφικής απεικόνισης δεδομένων, όπως τα Διαγράμματα Διασποράς, οι Παράλληλες Συντεταγμένες, οι δενδροχάρτες, τα επαναληπτικά πρότυπα κλπ. Ως μελέτη περίπτωσης παρατίθεται το πρόβλημα της ανίχνευσης απάτης με οπτικά μέσα και παρουσιάζονται σχετικά συστήματα. Το κεφάλαιο ολοκληρώνεται με μια αναφορά στα Ταμπλό (dashboards), τα οποία αποτελούν τον βασικό τρόπο οπτικοποίησης πληροφοριών στα συστήματα Επιχειρηματικής Ευφυΐας.

Το έκτο κεφάλαιο αποτελεί εισαγωγή στην Εξόρυξη Δεδομένων. Παρέχεται ο ορισμός της ΕΔ και σχολιάζονται τα στάδια της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων. Αναλύονται οι διάφορες εργασίες εξόρυξης δεδομένων, όπως η κατηγοριοποίηση, η ανάλυση συστάδων, η ανάλυση κανόνων συσχέτισης κλπ. Το κεφάλαιο ολοκληρώνεται με μια αρκετά διεξοδική παρουσίαση των πεδίων εφαρμογής μεθόδων Εξόρυξης Δεδομένων στη σύγχρονη επιχείρηση.

Το έβδομο κεφάλαιο καλύπτει θέματα προεπεξεργασίας δεδομένων. Αναλύονται τα προβλήματα των πηγαίων δεδομένων και εξηγείται η ανάγκη προεπεξεργασίας τους. Ακολούθως, παρουσιάζονται τα βασικά καθήκοντα προεπεξεργασίας δεδομένων. Σε αυτά περιλαμβάνονται η αντιμετώπιση των χαμένων τιμών, η αντιμετώπιση του θορύβου, η κανονικοποίηση των αριθμητικών τιμών, η διακριτοποίηση των αριθμητικών τιμών και η κατασκευή νέων πεδίων. Ιδιαίτερη μνεία γίνεται στο θέμα της επιλογής σημαντικών χαρακτηριστικών. Για κάθε ένα από αυτά τα θέματα, παρατίθενται και σχολιάζονται οι πιο διαδεδομένες και ευρέως χρησιμοποιούμενες τεχνικές.

Αντικείμενο του ογδόου κεφαλαίου είναι η ανάλυση Κανόνων Συσχέτισης. Αρχικά, παρουσιάζονται και εξηγούνται βασικές έννοιες των κανόνων συσχέτισης, όπως η έννοια του στοιχειοσυνόλου, της υποστήριξης και της εμπιστοσύνης. Ακολούθως, προσδιορίζεται η εργασία εξόρυξης κανόνων συσχέτισης ως μια διαδικασία δύο σταδίων, όπου στο πρώτο στάδιο εντοπίζονται συχνά στοιχειοσύνολα και στο δεύτερο στάδιο παράγονται οι κανόνες. Σχολιάζεται αναλυτικά ο αλγόριθμος εντοπισμού συχνών στοιχειοσυνόλων Apriori. Εξηγείται το στατιστικό μέτρο αξιολόγησης κανόνων Lift. Παρουσιάζονται οι πολυδιάστατοι κανόνες και η εξαγωγή κανόνων από αριθμητικά δεδομένα. Στο κλείσιμο του κεφαλαίου, παρατίθεται μελέτη περίπτωσης, όπου μοντελοποιούνται αποφάσεις εξωτερικών ελεγκτών με χρήση Κανόνων Συσχέτισης

Το ένατο κεφάλαιο καλύπτει εν μέρει τη θεματική ενότητα της Κατηγοριοποίησης. Αρχικά, εξηγούνται οι έννοιες της επιβλεπόμενης μάθησης, της Κατηγοριοποίησης των επαγωγικών αλγορίθμων και των μοντέλων. Παρατίθενται και σχολιάζονται τα τρία στάδια της κατηγοριοποίησης. Εξηγείται το φαινόμενο της υπερπροσαρμογής των μοντέλων και οι συνέπειες του. Επίσης, παρουσιάζονται τρεις πολύ διαδεδομένες μέθοδοι

κατηγοριοποίησης. Οι μέθοδοι αυτές είναι τα Δένδρα Αποφάσεων, τα νευρωνικά δίκτυα τύπου Multilayer Perceptron και οι Μπαϋεσιανοί Κατηγοριοποιητές. Η παρουσίαση των μεθόδων αυτών εμβαθύνει σε σημαντικό βαθμό. Σε μια μελέτη περίπτωσης, εφαρμόζονται οι προαναφερθείσες μέθοδοι κατηγοριοποίησης για τον εντοπισμό παραποιημένων χρηματοοικονομικών καταστάσεων.

Στο δέκατο κεφάλαιο ολοκληρώνεται η κάλυψη της θεματικής ενότητας της Κατηγοριοποίησης. Παρουσιάζονται αναλυτικά οι πολύ διαδεδομένες μέθοδοι των Μηχανών Διανυσμάτων Υποστήριξης και των k-Πλησιέστερων Γειτόνων. Αναφορά γίνεται και σε διάφορες εκδοχές της Παλινδρόμησης, καθώς και στη Λογιστική Παλινδρόμηση. Αναλύονται οι σύνθετοι κατηγοριοποιητές, δηλαδή οι συνδυασμοί κατηγοριοποιητών και οι υβριδικοί κατηγοριοποιητές. Το κεφάλαιο περιλαμβάνει ειδικά θέματα κατηγοριοποίησης, όπως τεχνικές επικύρωσης των μοντέλων, το πρόβλημα της ανισοκατανομής των κλάσεων και το πρόβλημα του διαφορετικού κόστους σφάλματος. Τα θέματα αυτά είναι πολύ συνηθισμένα σε ρεαλιστικά προβλήματα κατηγοριοποίησης. Το κεφάλαιο ολοκληρώνεται με μελέτη περίπτωσης, όπου εφαρμόζονται μέθοδοι κατηγοριοποίησης για την πρόβλεψη του τύπου του εξωτερικού ελεγκτή επιχειρήσεων.

Το ενδέκατο κεφάλαιο αναφέρεται στη θεματική ενότητα της Ανάλυσης Συστάδων. Σχολιάζονται οι έννοιες της μη επιβλεπόμενης μάθησης και της Ανάλυσης Συστάδων. Μελετώνται μέτρα ομοιότητας μεταξύ παρατηρήσεων, όπως η Ευκλείδεια Απόσταση, η απόσταση Manhattan κλπ. Τα μέτρα απόστασης επεκτείνονται, ώστε να καλύπτουν περιπτώσεις δεδομένων με δυαδικά, ονομαστικά και άλλα γνωρίσματα. Στη συνέχεια, παρουσιάζονται οι σημαντικότερες μέθοδοι ανάλυσης συστάδων, όπως οι Ιεραρχικές, οι Διαχωριστικές, οι μέθοδοι που βασίζονται στην πυκνότητα, οι μέθοδοι πλέγματος και οι μέθοδοι που βασίζονται σε μοντέλα. Συγκεκριμένες μέθοδοι, όπως ο αλγόριθμος k-Means και η μέθοδος των Αυτοοργανούμενων Χαρτών, παρουσιάζονται αναλυτικότερα. Στο τέλος του κεφαλαίου, γίνεται σύντομη αναφορά στις εφαρμογές της Ανάλυσης Συστάδων στη σύγχρονη επιχείρηση.

Αντικείμενο του ενδέκατου κεφαλαίου είναι η διαχείριση έργων για την ανάπτυξη Συστημάτων Επιχειρηματικής Ευφυΐας (ΣΕΕ). Αρχικά, παρουσιάζεται το μοντέλο ανάπτυξης ΣΕΕ το οποίο αποτελείται από διακριτά στάδια, διατεταγμένα σε μια κυκλική αλληλουχία, η οποία αποτυπώνει τη διαρκή μετεξέλιξη των ΣΕΕ. Ακολούθως, για κάθε ένα από τα στάδια αυτά, γίνεται παράθεση και σχολιασμός των ειδικών θεμάτων που εμπίπτουν σε αυτό. Μελετώνται θέματα σχετικά με τα ιδιόμορφα υποσυστήματα των ΣΕΕ, όπως είναι οι Αποθήκες Δεδομένων, το υποσύστημα Εξαγωγής, Μετασχηματισμού και Φόρτωσης δεδομένων και το υποσύστημα μεταδεδομένων. Τέλος, εξετάζονται οι κρίσιμοι παράγοντες επιτυχίας Συστημάτων Επιχειρηματικής Ευφυΐας, οι οποίοι διαφέρουν από τους αντίστοιχους άλλων πληροφοριακών συστημάτων.

Στο δέκατο τρίτο κεφάλαιο, γίνεται παρουσίαση του ελεύθερου λογισμικού Μηχανικής Μάθησης και Εξόρυξης Δεδομένων WEKA. Το λογισμικό διαθέτει γραφική διεπαφή, γεγονός που επιτρέπει τη χρήση από τελικούς χρήστες, οι οποίοι στερούνται ικανότητες προγραμματισμού. Το WEKA παρέχει πληθώρα υλοποιημένων αλγορίθμων για προεπεξεργασία δεδομένων, κατηγοριοποίηση, ανάλυση συστάδων, κανόνες συσχέτισης και οπτικοποίηση. Οι φοιτητές μπορούν να χρησιμοποιήσουν αυτό το λογισμικό, για εξάσκηση στις μεθόδους Εξόρυξης Δεδομένων και για την ανάπτυξη δεξιοτήτων εκτέλεσης αναλυτικών εργασιών και ανάπτυξης μοντέλων.

Το παρόν πόνημα αποτελεί καρπό μακροχρόνιας ερευνητικής και διδακτικής ενασχόλησης του συγγραφέα με το αντικείμενο. Παραδίδεται στις τρέχουσες και ερχόμενες γενιές φοιτητών, με την ελπίδα ότι θα αποτελέσει χρήσιμο βοήθημα για την εκπαίδευση τους στο αντικείμενο της Επιχειρηματικής Ευφυΐας και της εφαρμοσμένης Εξόρυξης Δεδομένων.

1 Εισαγωγή στην Επιχειρηματική Ευφυΐα

Σύνοψη

Το παρόν Κεφάλαιο προσφέρει μια εισαγωγή στην Επιχειρηματική Ευφυΐα (ΕΕ). Αρχικά, γίνεται αναφορά στην άνθηση που παρουσιάζει η ΕΕ τα τελευταία χρόνια, και δίνονται ο ορισμός της ΕΕ καθώς και ο ορισμός των Συστημάτων Επιχειρηματικής Ευφυΐας (ΣΕΕ). Στη συνέχεια, γίνεται αναλυτική παρουσίαση των επιχειρηματικών και τεχνολογικών αίτιων, τα οποία καθόρισαν την ανάγκη ύπαρξης, τη δυνατότητα υλοποίησης και την πρόσφατη ραγδαία ανάπτυξη των ΣΕΕ. Παρουσιάζεται η πυραμίδα των ΣΕΕ. και γίνεται συνοπτική αναφορά στα επίπεδα που την απαρτίζουν, από το βασικό επίπεδο των πηγών δεδομένων, μέχρι το τελικό επίπεδο της λήψης αποφάσεων. Ακολούθως, παρατίθενται τα οφέλη που προσφέρει η ΕΕ αλλά και τα σχετικά προβλήματα, οι κίνδυνοι και οι ανασχετικοί παράγοντες.

Η ΕΕ γνωρίζει σήμερα πολλά πεδία εφαρμογής στη σύγχρονη επιχείρηση. Στο παρόν Κεφάλαιο παρουσιάζονται τα κυριότερα πεδία εφαρμογής, όπως η Διοίκηση Επιχειρησιακής Απόδοσης, η χρηματοοικονομική ανάλυση και διαχείριση, οι πωλήσεις, το μάρκετινγκ, η διαχείριση της εφοδιαστικής αλυσίδας κλπ. Τέλος, γίνεται παρουσίαση των σημαντικότερων παρόχων λογισμικού και υπηρεσιών ΕΕ, καθώς και των βασικών προϊόντων τους.

Προαπαιτούμενη γνώση

Η κατανόηση του εισαγωγικού αυτού κεφαλαίου δεν απαιτεί προηγούμενες εξειδικευμένες γνώσεις. Για τους αναγνώστες που επιθυμούν να αναζητήσουν πρόσθετη πληροφόρηση για γενικά θέματα ΕΕ, προτείνονται τα παρακάτω βιβλία και ηλεκτρονικές πηγές:

Βιβλία:

- *Business Intelligence for Dummies του Scheps (2007).*
- *Business Intelligence των Sabherwal and Beccera – Fernandez (2010).*

Ιστοθέσεις:

- www.information-management.com*, ιστοθέση του περιοδικού Information Management | IT Business News ("Information Management | IT Business News," n.d.).*
- www.informationweek.com*, ιστοθέση του περιοδικού Information Week ("Information Week," n.d.).*
- tdwi.org*, ιστοθέση του Data Warehousing Institute ("TDWI | Advancing all things data," n.d.). Η ιστοθέση περιέχει πολλά ηλεκτρονικά έγγραφα σχετικά με τη σημασία και την αξία της Ε.Ε.*
- www.tdan.com*, ιστοθέση της ηλεκτρονικής έκδοσης The Data Administration Newsletter ("TDAN.com," n.d.), που αναφέρεται σε διάφορα θέματα διαχείρισης δεδομένων.*

1.1 Η Επιχειρηματική Ευφυΐα

Αποτελεί κοινό τόπο ότι το επιχειρηματικό περιβάλλον στην αρχή του 21ου αιώνα μπορεί να χαρακτηριστεί πλούσιο, τόσο σε νέες δυνατότητες και ευκαιρίες όσο και σε δυσκολίες που ανέκυψαν από την πρόσφατη οικονομική κρίση. Για την επιτυχή ανταπόκριση των επιχειρήσεων σε αυτές τις νέες προκλήσεις, απαιτείται αναβάθμιση των διοικητικών πρακτικών και βελτίωση των διαδικασιών λήψης αποφάσεων. Προαπαιτούμενο για βελτιωμένες αποφάσεις είναι η βαθιά κατανόηση και γνώση του περιβάλλοντος, αλλά και της ίδιας της επιχείρησης, καθώς και η έγκαιρη και ουσιαστική πληροφόρηση. Έχει ειπωθεί επανειλημμένως ότι η πληροφορία είναι ένα από τα πολυτιμότερα κεφάλαια ενός οργανισμού.

Τα παραπάνω μπορούν να αποτελέσουν μια καταρχήν εξήγηση του γεγονότος ότι η Επιχειρηματική Ευφυΐα βρίσκεται τον τελευταίο καιρό στο επίκεντρο του ενδιαφέροντος του επιχειρηματικού κόσμου. Τα αποτελέσματα της έκθεσης του οίκου Gartner είναι εξόχως αποκαλυπτικά. Ο οίκος Gartner πραγματοποιεί κάθε χρόνο μια έρευνα με στόχο να εντοπίσει τις τεχνολογικές και επιχειρηματικές προτεραιότητες των μεγάλων επιχειρήσεων. Στην έρευνα συμμετέχουν περισσότεροι από 2000 διευθύνοντες σύμβουλοι πολύ μεγάλων επιχειρήσεων, οι οποίοι αντιπροσωπεύουν δεκάδες επιχειρηματικούς κλάδους καθώς και δεκάδες χώρες. Στη σχετική έκθεση των ετών 2012 («Gartner Executive Programs' Worldwide Survey of More Than 2,300 CIOs Shows

Flat IT Budgets in 2012, but IT Organizations Must Deliver on Multiple Priorities,» n.d.), 2013 («Evtm_219_CIOtop10[3].pdf," n.d.) και 2015 ("http://www.gartnerinfo.com/cios9/ CIOLeadershipForum2015Profile.pdf," 2015) η Επιχειρηματική Ευφυΐα βρίσκεται στην κορυφαία θέση του καταλόγου των τεχνολογικών προτεραιοτήτων.

Το ισχυρό ενδιαφέρον του επιχειρηματικού κόσμου για Συστήματα Επιχειρηματικής Ευφυΐας έχει προκαλέσει τη δημιουργία μιας αντίστοιχης, πολύ δυναμικής αγοράς. Σύμφωνα με μια ανάλυση αγοράς της IDC, την οποία δημοσιοποιεί στην ιστοθέση της η εταιρεία συστημάτων επιχειρηματικής ευφυΐας SAS, το συνολικό ύψος εσόδων από πωλήσεις λογισμικού αναλυτικής των επιχειρήσεων ανήλθε το έτος 2013 στο ποσό των 37 δισεκατομμυρίων δολαρίων, παρουσιάζοντας αύξηση σε σχέση με το έτος 2012 της τάξης του 8%. Στην ίδια έκθεση αναφέρεται ότι ο ετήσιος ρυθμός αύξησης μέχρι το έτος 2018 αναμένεται να είναι της τάξης του 9%.

Ο όρος Επιχειρηματική Ευφυΐα (Business Intelligence) δεν είναι πρόσφατος. Πρωτοεμφανίζεται το 1865 στο βιβλίο "Cyclopædia of commercial and business anecdotes" του Devens (1865). Ο Devens χρησιμοποιεί αυτόν τον όρο για να αναφερθεί στον τρόπο με τον οποίο ο τραπεζίτης Sir Henry Furnese αξιοποιούσε πληροφορίες νωρίτερα από τους ανταγωνιστές του, έτσι ώστε να επιτύχει αύξηση των κερδών του. Η επόμενη εμφάνιση του όρου καταγράφεται το 1958 σε τίτλο άρθρου του Luhn (1958) σε περιοδικό της IBM.

Στη σύγχρονη βιβλιογραφία ο αναγνώστης θα συναντήσει διαφοροποιημένους ορισμούς της Επιχειρηματικής Ευφυΐας. Στο παρόν σύγγραμμα, θα ορίσουμε την Επιχειρηματική Ευφυΐα ως ένα σύνολο από μεθόδους ανάλυσης, τεχνολογίες, ικανότητες και στρατηγικές, οι οποίες στόχο έχουν την επεξεργασία των διαθέσιμων δεδομένων και την εξαγωγή χρήσιμης πληροφορίας από αυτά, για την υποστήριξη της διαδικασίας λήψης επιχειρηματικών αποφάσεων. Ένας άλλος συγγενής, αν και όχι ταυτόσημος όρος, ο οποίος γνωρίζει ιδιαίτερη διάδοση τον τελευταίο καιρό είναι «Αναλυτική των Επιχειρήσεων» (Business Analytics). Η Επιχειρηματική Ευφυΐα επιτρέπει σε έναν οργανισμό να μαθαίνει, να αντιλαμβάνεται καταστάσεις και συμβάντα, να σκέφτεται αφαιρετικά, να προβλέπει τάσεις και μελλοντικά συμβάντα, να σχεδιάζει και να καινοτομεί. Η παραγόμενη πληροφορία μετουσιώνεται σε γνώση που αξιοποιείται από τα διοικητικά στελέχη, ώστε να δρομολογήσουν κατάλληλες δράσεις, που θα οδηγήσουν στον καθορισμό και την επίτευξη επιχειρηματικών στόχων, με τρόπο αποτελεσματικό και αποδοτικό.

Τα συστήματα Επιχειρηματικής Ευφυΐας είναι εξειδικευμένα πληροφοριακά συστήματα, τα οποία προσφέρουν ποιοτική πληροφορία. Η πληροφορία βασίζεται σε ποιοτικά και συγκεντρωτικά δεδομένα, τα οποία συνδυάζονται με λογισμικό ικανό να διεξάγει κατάλληλες αναλύσεις. Η βελτίωση της ποιότητας της πληροφορίας οφείλεται στις δυνατότητες αυτών των συστημάτων, τα οποία επιτρέπουν την ταχύτερη πρόσβαση στην πληροφορία, την ευκολότερη υποβολή ερωτημάτων στο σύστημα και τη σύνταξη αναφορών, την προχωρημένη ανάλυση των δεδομένων, καθώς και τη βελτίωση της ποιότητας των δεδομένων. Οι τελικοί αποδέκτες του προϊόντος των συστημάτων ΕΕ, οι οποίοι πολλές φορές αναφέρονται στη βιβλιογραφία ως «εργάτες γνώσης», τροφοδοτούνται έγκαιρα με γνώση που χρησιμοποιούν για τη λήψη αποφάσεων.

Πρόδρομοι των σύγχρονων Συστημάτων Επιχειρηματικής Ευφυΐας μπορούν να θεωρηθούν τα Συστήματα Υποστήριξης Αποφάσεων (ΣΥΑ). Τα ΣΥΑ, τα οποία καθιερώθηκαν ως πεδίο συστηματικής έρευνας τη δεκαετία του 1970, στηρίζονται κυρίως στη χρήση μοντέλων. Κάνοντας χρήση των μοντέλων, ο χρήστης μπορεί να πειραματιστεί με διάφορα σενάρια, όπως π.χ. τι θα συμβεί εάν μεταβληθεί κάποια συνθήκη εισόδου (ανάλυση what-if) ή να καθορίσει το επιθυμητό αποτέλεσμα και να αναζητήσει τις αναγκαίες συνθήκες εισόδου (αναζήτηση στόχου). Οι Αποθήκες Δεδομένων (Data Warehouse) και οι τεχνικές OLAP (OnLine Analytical Processing) αποτέλεσαν τον επόμενο σταθμό στην ιστορία της Επιχειρηματικής Ευφυΐας. Στις Αποθήκες Δεδομένων συγκεντρώνονται δεδομένα που είναι διάσπαρτα σε διάφορες πηγές. Τα δεδομένα αυτά, αφού υποστούν επεξεργασία ώστε να αντιμετωπιστούν διάφορα προβλήματα, αποθηκεύονται σε συγκεντρωτική μορφή (πχ πωλήσεις ανά μήνα ή ανά κατηγορία προϊόντος). Με τις τεχνικές OLAP ο χρήστης μπορεί να προβάλει και να αναλύσει τα δεδομένα σε διάφορα επίπεδα γενίκευσης (π.χ. πωλήσεις ανά μήνα ή ανά τρίμηνο ή ανά έτος). Στη σημερινή εποχή ένας νέος κλάδος της Πληροφορικής, η Εξόρυξη Δεδομένων, έρχεται να δώσει νέα ώθηση στην Επιχειρηματική Ευφυΐα. Η Εξόρυξη Δεδομένων (Data Mining) ή Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases) στοχεύει στην ανακάλυψη γνώσης που είναι κρυμμένη σε μεγάλους όγκους δεδομένων. Οι τεχνικές Εξόρυξης Δεδομένων δεν απαιτούν τον προκαθορισμό μοντέλων. Αντιθέτως, τα μοντέλα προκύπτουν από την επεξεργασία των δεδομένων. Επίσης, τα μοντέλα μπορούν να χρησιμοποιηθούν για τη διατύπωση προβλέψεων.

1.2 Γιατί Επιχειρηματική Ευφυΐα;

Όπως διατυπώθηκε και παραπάνω, η Επιχειρηματική Ευφυΐα βρίσκεται στο επίκεντρο του ενδιαφέροντος των σύγχρονων μεγάλων επιχειρήσεων. Οι κυριότερες αιτίες γι' αυτό το γεγονός είναι οι ακόλουθες:

1.2.1 Λήψη Επιχειρηματικών Αποφάσεων σε συνθήκες αβεβαιότητας

Η λήψη αποφάσεων είναι μια από τις σημαντικότερες ευθύνες της διοίκησης μιας επιχείρησης. Ο ισχυρισμός αυτός, αν και έκδηλα προφανής, στοιχειοθετείται με σαφήνεια στις εργασίες επιστημόνων οι οποίοι ασχολούνται με τη διοίκηση επιχειρήσεων. Ο Fayol (1949) υποστηρίζει ότι η διοίκηση ενός οργανισμού εκτελεί εργασίες πρόβλεψης και κατάστρωσης σχεδίων, οργάνωσης των δομών και διάθεσης υλικών και ανθρωπίνων πόρων, διοίκησης των δραστηριοτήτων και του προσωπικού, συντονισμού, ενοποίησης και εναρμόνισης πρακτικών και τέλος, ελέγχου συμφωνίας με καθορισμένες πρακτικές και πολιτικές. Ο Mintzberg (1990) ασκεί κριτική στον Fayol και ορίζει ότι η διοίκηση επιτελεί τρεις βασικούς ρόλους: διαπροσωπικούς, πληροφοριακούς και ρόλους λήψης αποφάσεων.

Οι αποφάσεις που λαμβάνονται στα πλαίσια της λειτουργίας ενός οργανισμού ποικίλουν ως προς τον βαθμό αβεβαιότητας. Αποφάσεις που σχετίζονται με ζητήματα καθημερινής λειτουργίας είναι συνήθως σχετικά απλές και τυποποιημένες. Θα μπορούσε να πει κανείς ότι είναι περισσότερο διαδικασίες και λιγότερο αποφάσεις. Μια απόφαση για αναπαραγγελία νέων εμπορευμάτων, όταν τα αποθέματα ξεπεράσουν το χαμηλότερο επιτρεπτό όριο, είναι μια απλή απόφαση καθημερινής λειτουργίας. Τέτοιες αποφάσεις μπορούν να τυποποιηθούν και να ληφθούν ακόμα και αυτόματα, με τη χρήση κατάλληλου λογισμικού. Άλλες αποφάσεις όμως, που αφορούν ευρύτερα τμήματα του οργανισμού ή, ακόμα περισσότερο, που αφορούν ζητήματα στρατηγικού προσανατολισμού είναι πολύ πιο περίπλοκες. Για παράδειγμα, η απόφαση μιας επιχείρησης να παράξει ένα πρωτοποριακό προϊόν, το οποίο δημιουργεί μια νέα κατηγορία προϊόντων, είναι ιδιαίτερα απαιτητική. Θα πρέπει να συνεκτιμηθούν οι καταναλωτικές τάσεις, οι προτιμήσεις και ανάγκες των πελατών, ο προσανατολισμός των τεχνολογικών εξελίξεων, η δυναμική που δημιουργεί το νέο προϊόν στην αγορά, οι πιθανές αντιδράσεις των ανταγωνιστών, οι πιθανές αντιδράσεις συνεργατών, οι οποίοι ενδεχομένως να θιγούν από μια τέτοια κίνηση της εταιρείας, τα χαρακτηριστικά που πρέπει να έχει το νέο προϊόν, το κόστος της επένδυσης και τα αναμενόμενα οικονομικά οφέλη, η τιμή του νέου προϊόντος ώστε η πώληση του να είναι εφικτή, καθώς και πολλά άλλα ζητήματα. Η απόφαση της Apple να λανσάρει το iPod είναι μια χαρακτηριστική τέτοια περίπτωση. Προφανώς, αποφάσεις αυτής της εμβέλειας και αυτού του τύπου είναι ιδιαίτερα περίπλοκες, καθώς υπεισέρχεται μεγάλος βαθμός αβεβαιότητας σε σχέση με πολλά ζητήματα.

Εκτός του γεγονότος ότι οι αποφάσεις στρατηγικού προσανατολισμού είναι από τη φύση τους περίπλοκες και απαιτούν τη διαχείριση του ρίσκου ή της αβεβαιότητας, το σύγχρονο επιχειρηματικό περιβάλλον είναι ιδιαίτερα απαιτητικό, με αποτέλεσμα η λήψη αποφάσεων να καθίσταται ακόμα δυσκολότερη. Μερικοί παράγοντες που αυξάνουν τον βαθμό πολυπλοκότητας είναι οι ακόλουθοι:

- Το εξωτερικό περιβάλλον είναι ασταθές και μεταβάλλεται με μεγάλη ταχύτητα.
- Ο ρυθμός λειτουργίας έχει εντατικοποιηθεί, με αποτέλεσμα οι αποφάσεις να λαμβάνονται υπό την πίεση του χρόνου.
- Έχει διαπιστωθεί αύξηση του ανταγωνισμού ποσοτικά αλλά και ποιοτικά.
- Οι επιχειρήσεις γιγαντώνονται και διασπείρονται γεωγραφικά, με αποτέλεσμα να καθίσταται δυσκολότερη η διαχείριση τους.
- Το ανθρώπινο δυναμικό είναι ποιοτικά αναβαθμισμένο και διαθέτει υψηλή εξειδίκευση και αυξημένες δυνατότητες.
- Η απορρύθμιση κανονιστικών διατάξεων επιτρέπει στις επιχειρήσεις μεγαλύτερη ευελιξία κινήσεων, με αποτέλεσμα να αυξάνεται το πλήθος των εναλλακτικών λύσεων.
- Ο ρυθμός παροχής πληροφοριών είναι καταιγιστικός. Η δυνατότητα παροχής πρωτόγνωρα ποιοτικής πληροφόρησης είναι παρούσα.

Τα διοικητικά στελέχη των επιχειρήσεων, κατά τη λήψη αποφάσεων, χρησιμοποιούν τη γνώση τους σχετικά με τον τομέα τους και το αντικείμενο τους, τη διοικητική τους εμπειρία και τα υποκειμενικά στοιχεία του χαρακτήρα τους και τέλος τις διαθέσιμες πληροφορίες. Για τον λόγο αυτό, η παροχή κατάλληλης πληροφόρησης αποτελεί καθοριστικό παράγοντα για τη λήψη επιτυχημένων αποφάσεων. Κατάλληλη πληροφόρηση σημαίνει ότι δίνεται η σωστή πληροφορία στο σωστό άτομο την αναγκαία χρονική στιγμή. Βελτιωμένες αποφάσεις και κατ' επέκταση βελτιωμένο μάνατζμεντ μπορούν να αυξήσουν τις επιδόσεις της επιχείρησης και να

της εξασφαλίσουν το ανταγωνιστικό πλεονέκτημα. Τα συστήματα Επιχειρηματικής Ευφυΐας συμβάλλουν σε αυτήν την κατεύθυνση, προσφέροντας πληροφόρηση και μειώνοντας τον βαθμό αβεβαιότητας κατά τη λήψη αποφάσεων (Ferrari, 2011)

1.2.2 Οι προκλήσεις της παγκοσμιοποίησης

Στην εποχή της παγκοσμιοποίησης το επιχειρηματικό περιβάλλον άλλαξε και εξακολουθεί να αλλάζει με ταχύτατους ρυθμούς. Η παγκοσμιοποιημένη οικονομία προκάλεσε την ανάπτυξη και ολοκλήρωση παγκόσμιων αγορών. Οι επιχειρήσεις πλέον δραστηριοποιούνται και ανταγωνίζονται σε παγκόσμια κλίμακα. Ο περιορισμός των συνοριακών δασμών και η απορρύθμιση των προστατευτικών μέτρων επιτρέπει σε ξένες επιχειρήσεις να εισέλθουν ευκολότερα σε εγχώριες αγορές. Η άρση των εμποδίων και ο περιορισμός του κόστους εισόδου αυξάνει το πλήθος των ανταγωνιστών. Το τελικό αποτέλεσμα είναι η ένταση του ανταγωνισμού, τόσο ποσοτικά όσο και ποιοτικά.

Οι σημερινές επιχειρήσεις είναι διασκορπισμένες σε πολλές χώρες. Το γεγονός αυτό αυξάνει την πολυπλοκότητα τους και καθιστά δυσκολότερη την παρακολούθηση και τη διοίκηση τους. Επίσης, η επιχειρηματική δραστηριοποίηση σε παγκόσμια κλίμακα περιλαμβάνει και την αντιμετώπιση προβλημάτων, που ανακύπτουν από τις διαφορετικές κουλτούρες. Η πρόσληψη του σημαινομένου μιας διαφημιστικής εκστρατείας μπορεί να είναι τελείως διαφορετική σε ανθρώπους διαφορετικών πολιτισμών. Μια εικόνα, η οποία για έναν καταναλωτή δυτικών κοινωνιών είναι ελκυστική, μπορεί να θεωρηθεί κακόγουστη ή και προσβλητική σε μια κοινωνία του ανατολικού κόσμου. Επίσης, το εργατικό δυναμικό πολυεθνικών επιχειρήσεων, το οποίο έχει διαφορετικές θρησκείες και κουλτούρες, μπορεί να αντιδράσει διαφορετικά σε εργασιακές πολιτικές ενθάρρυνσης και παρακίνησης των εργαζομένων.

Τα νέα κανάλια επικοινωνίας και κυρίως το διαδίκτυο επιτρέπουν τη διάχυση της πληροφορίας σε παγκόσμια κλίμακα. Ο καταναλωτής της σημερινής οικονομίας είναι καλύτερα πληροφορημένος, διαθέτει μόρφωση και δεξιότητες χειρισμού νέων τεχνολογιών, έχει υψηλό εισόδημα και για τους λόγους αυτούς έχει και υψηλότερες απαιτήσεις. Η ανταπόκριση στις υψηλές απαιτήσεις των σύγχρονων πελατών αποτελεί νέα πρόκληση για τις επιχειρήσεις.

Μία άλλη σημαντική παράμετρος του σημερινού επιχειρηματικού περιβάλλοντος είναι η ανάδυση των πάλλε ποτέ αναπτυσσόμενων χωρών και η καθιέρωση τους ως πρωταγωνιστικές δυνάμεις, με οικονομικά μεγέθη συγκρίσιμα με αυτά των παραδοσιακά ανεπτυγμένων δυτικών κοινωνιών. Η καταναλωτική άνθηση αυτών των κοινωνιών προσφέρει νέες επιχειρηματικές ευκαιρίες.

Όλοι οι παραπάνω παράγοντες συμβάλλουν στη διαμόρφωση ενός επιχειρηματικού περιβάλλοντος ιδιαίτερα σύνθετου και αβέβαιου. Για την αντιμετώπιση των αυξημένων προκλήσεων της παγκοσμιοποίησης χρειάζεται ιδιαίτερα αποτελεσματική διοίκηση. Η αναβάθμιση των διοικητικών πρακτικών περιλαμβάνει ως βασική συνιστώσα και τη βελτίωση των διαδικασιών λήψης αποφάσεων. Η τροφοδότηση με ποιοτική, δηλαδή ακριβή, σαφή, σχετική με το εξεταζόμενο ζητούμενο και έγκαιρη πληροφορία, επιτρέπει τη λήψη καλύτερων αποφάσεων.

1.2.3 Η οικονομική κρίση και οι νέες κανονιστικές διατάξεις

Οι απαρχές του 21ου αιώνα σημαδεύτηκαν από μια δριμύτατη οικονομική κρίση. Η κρίση πρωτοεμφανίστηκε στην αγορά ακινήτων των ΗΠΑ και στη συνέχεια εξελίχθηκε σε τραπεζική κρίση. Το αποτέλεσμα ήταν η χρεοκοπία εκατοντάδων αμερικανικών τραπεζών και η διάσωση άλλων. Πολύ σύντομα, η κρίση πέρασε τον Ατλαντικό ωκεανό και παρουσιάστηκε και στην Ευρωπαϊκή Ένωση, προκαλώντας προβλήματα στον τραπεζικό τομέα αλλά και στην πιστοληπτική ικανότητα κρατών. Ορισμένα κράτη, μεταξύ των οποίων και η Ελλάδα, οδηγήθηκαν σε προγράμματα δανειοδότησης ελεγχόμενα από θεσμούς, όπως το Διεθνές Νομισματικό Ταμείο και η Ευρωπαϊκή Κεντρική Τράπεζα.

Σε μια προσπάθεια θωράκισης του χρηματοπιστωτικού συστήματος και αντιμετώπισης ατελειών που ανέδειξε η οικονομική κρίση, αρμόδιοι φορείς ενεργοποιήθηκαν για τη θέσπιση ενός νέου κανονιστικού πλαισίου για τη λειτουργία των τραπεζών. Επιδιώκοντας τη μείωση της μόχλευσης, η συνθήκη Βασιλεία ΙΙΙ ορίζει νέους κανόνες που αφορούν στην κεφαλαιακή επάρκεια των τραπεζών, στα τεστ αντοχής και σε κινδύνους σχετικούς με τη ρευστότητα. Σύμφωνα με τα νέες διατάξεις, οι τράπεζες είναι υποχρεωμένες να συντάσσουν και να κοινοποιούν πλήθος αναφορών σχετικά με τα οικονομικά τους στοιχεία. Για την εργασία αυτή απαιτείται η συγκέντρωση, ενοποίηση και επεξεργασία πολλών δεδομένων και η παραγωγή κατάλληλης πληροφορίας. Εξειδικευμένα συστήματα μπορούν να αναλάβουν την αποτελεσματική εκτέλεση αυτών των εργασιών και να διασφαλίσουν την κανονιστική συμμόρφωση (regulatory compliance).

1.2.4 Διαθεσιμότητα δεδομένων

Στη σημερινή εποχή, κάθε επιχείρηση διαθέτει μηχανογραφικό σύστημα, με το οποίο καταγράφει δεδομένα για τις συναλλαγές και τις λοιπές δραστηριότητες της. Τα Συστήματα Σχεδιασμού Επιχειρησιακών Πόρων (Enterprise Resources Planning (ERP)), τα οποία αποτελούν τη βασική πλατφόρμα μηχανοργάνωσης των σημερινών επιχειρήσεων, επιτρέπουν την παρακολούθηση των συναλλαγών σε όλες τις λειτουργικές περιοχές της αλυσίδας αξίας ενός οργανισμού, μέσα από ένα ενιαίο περιβάλλον. Άλλα συστήματα παρακολούθησης συναλλαγών, που γνωρίζουν ιδιαίτερη διάδοση, είναι τα Συστήματα Διαχείρισης Εφοδιαστικής Αλυσίδας (Supply Chain Management (SCM)) και τα Συστήματα Διαχείρισης Σχέσεων Πελατών (Customer Relationship Management (CRM)). Όλα αυτά τα συστήματα καταγράφουν καθημερινά, σε σχεσιακές βάσεις, τεράστιους όγκους δεδομένων, που αφορούν τις δραστηριότητες της επιχείρησης. Η παραγωγή και καταγραφή δεδομένων εντείνεται περαιτέρω, με τη χρήση διαφόρων συσκευών όπως barcode readers, συστήματα ετικετών RFID, συστήματα GPS, κάμερες κλπ.

Οι εταιρικές ιστοθέσεις είναι μια άλλη πηγή παραγωγής και καταγραφής δεδομένων. Οι σύγχρονες επιχειρήσεις επιθυμούν να έχουν παρουσία στον παγκόσμιο ιστό. Οι ιστοθέσεις τους, οι οποίες σε ορισμένες περιπτώσεις είναι κανονικές πύλες (portals), χρησιμοποιούνται καθημερινά από διάφορους χρήστες όπως υπαλλήλους της εταιρείας, προμηθευτές, συνεργάτες και πελάτες. Η χρήση της ιστοθέσης από τους επισκέπτες της παράγει δεδομένα. Τα δεδομένα αυτά, σε αντίθεση με τα δεδομένα των συστημάτων παρακολούθησης συναλλαγών τα οποία είναι δομημένα, είναι κατά κανόνα αδόμητα και μπορούν να αφορούν σχόλια πελατών για τα προϊόντα της επιχείρησης ή το ρεύμα κλικ των επισκεπτών της ιστοθέσης.

Πέρα από τα δεδομένα που παράγονται από τα μηχανογραφικά συστήματα των επιχειρήσεων, είναι διαθέσιμα και πολλά δεδομένα, τα οποία προέρχονται από εξωτερικές πηγές. Τρίτοι φορείς, όπως κρατικές υπηρεσίες, μέσα ενημέρωσης, τράπεζες και άλλες επιχειρήσεις, μπορεί να προσφέρουν σημαντική πληροφόρηση. Επίσης, μια τεράστια και διαρκώς αυξανόμενη δεξαμενή δεδομένων είναι το Web 2.0. Ιστοθέσεις κοινωνικής δικτύωσης, blogs, wikis και γενικώς ιστοθέσεις το περιεχόμενο των οποίων παράγεται από τους χρήστες του δικτύου, επιτρέπουν την ελεύθερη έκφραση των ανθρώπων και την καταγραφή των απόψεων τους. Κατάλληλη επεξεργασία των στοιχείων αυτών μπορεί να αποκαλύψει καταναλωτικές τάσεις και επιχειρηματικές ευκαιρίες.

Συμπερασματικά, η σύγχρονη επιχείρηση έχει στη διάθεση της τεράστιους όγκους εσωτερικών και εξωτερικών δεδομένων. Τα δεδομένα αυτά μπορεί να είναι διάσπαρτα σε διάφορες πηγές και να περιέχουν ελλιπή ή και αντιφατικά στοιχεία. Ταυτόχρονα όμως, περιέχουν και πληροφορία πολύτιμη για την επιχείρηση. Ένας σύγχρονος όρος, που περιγράφει την υπερσυσσώρευση των δεδομένων και αναφέρεται στις τεχνικές επεξεργασίας τους και στη δυνατότητα εύρεσης πληροφορίας σε αυτά, είναι «Big Data». Τα συστήματα Επιχειρηματικής Ευφυΐας στοχεύουν ακριβώς στη συγχώνευση και επεξεργασία, τόσο των εσωτερικών όσο και των εξωτερικών δεδομένων, και στην ανακάλυψη πολύτιμης πληροφορίας που θα χρησιμοποιηθεί για τη λήψη αποφάσεων.

1.2.5 Νέες τεχνολογίες και μέθοδοι ανάλυσης

Η ανάλυση των δεδομένων και η εξαγωγή συμπερασμάτων από αυτά γινόταν παλαιότερα αποκλειστικά με χρήση στατιστικών μεθόδων. Αργότερα, η πολυδιάστατη ανάλυση, με χρήση Αποθηκών Δεδομένων και κύβων, εμπλούτισε το φάσμα των διαθέσιμων τεχνικών. Κοινό χαρακτηριστικό και στις δύο παραπάνω περιπτώσεις είναι ότι ο χρήστης διατυπώνει εκ των προτέρων υποθέσεις και στη συνέχεια ελέγχει την ισχύ τους αναλύοντας τα δεδομένα.

Στη σημερινή εποχή ένας νέος κλάδος της Πληροφορικής, η Εξόρυξη Δεδομένων, προσφέρει πρωτόγνωρες δυνατότητες για την επεξεργασία των δεδομένων και την ανακάλυψη της γνώσης. Κατ' αρχήν, η Εξόρυξη Δεδομένων ασχολείται με την επεξεργασία μεγάλου όγκου δεδομένων, δίνοντας απαντήσεις σε σχετικά προβλήματα. Δεύτερον, ακολουθεί μια ολιστική προσέγγιση και παρέχει μεθοδολογίες για όλα τα στάδια της ανακάλυψης γνώσης, από την αρχική συγκέντρωση και προεπεξεργασία των δεδομένων μέχρι και την οπτικοποίηση των προτύπων και τη διατύπωση των τελικών συμπερασμάτων. Αντιμετωπίζονται προβλήματα όπως οι χαμένες τιμές, ο θόρυβος, ο κατάλληλος μετασχηματισμός των δεδομένων κλπ. Τρίτον, οι μέθοδοι της επεξεργασίας των δεδομένων δεν προέρχονται μόνο από τη Στατιστική. Η Εξόρυξη Δεδομένων κάνει ευρύτατη χρήση μεθόδων οι οποίες προέρχονται από την Τεχνητή Νοημοσύνη, τη Μηχανική Μάθηση και την Αναγνώριση Προτύπων. Έρευνες έχουν αποδείξει ότι οι νέες αυτές μέθοδοι μπορούν να δώσουν καλύτερα αποτελέσματα από τις παραδοσιακές στατιστικές μεθόδους. Επίσης, η ανάλυση Κανόνων Συσχέτισης είναι μια νέα μέθοδος επεξεργασίας, η οποία προέρχεται απ' ευθείας από την Εξόρυξη Δεδομένων. Τέταρτον, πολ-

λές από τις παραπάνω μεθόδους δεν απαιτούν την εκ των προτέρων διατύπωση υποθέσεων. Αντιθέτως, τα μοντέλα προκύπτουν απευθείας από τα δεδομένα, με κατάλληλη επεξεργασία. Τέλος, οι νέες μέθοδοι δίνουν τη δυνατότητα προγνωστικής ανάλυσης, δηλαδή την επεξεργασία ιστορικών στοιχείων και τη διατύπωση προβλέψεων για το μέλλον.

Από τα παραπάνω, καθίσταται σαφές ότι ο σύγχρονος αναλυτής έχει πλέον στη διάθεση του βελτιωμένες μεθόδους για να επεξεργαστεί τους τεράστιους όγκους των αποθηκευμένων δεδομένων και να αντλήσει πληροφόρηση, πολύτιμη για τη λήψη αποφάσεων. Συμπερασματικά, η φύση της διαδικασίας λήψης επιχειρηματικών αποφάσεων, κυρίως σε στρατηγικό επίπεδο, η οποία περιλαμβάνει τη διαχείριση της αβεβαιότητας, σε συνδυασμό με τις νέες προκλήσεις της παγκοσμιοποιημένης οικονομίας και της πρόσφατης οικονομικής κρίσης, έθεσαν επιτακτικά την ανάγκη ποιοτικής και έγκαιρης πληροφόρησης. Ταυτόχρονα, η μαζική εφαρμογή της πληροφορικής πρόσφερε τα αναγκαία δεδομένα, ενώ οι νέες μεθοδολογίες ανάλυσης έδωσαν τη δυνατότητα της επεξεργασίας τους και την εξαγωγή της χρήσιμης πληροφορίας. Οι παραπάνω παράγοντες είναι αυτοί που συνέβαλαν στην άνθιση της Επιχειρηματικής Ευφυΐας.

1.3 Δομικά Επίπεδα Συστημάτων Επιχειρηματικής Ευφυΐας

Τα συστήματα Επιχειρηματικής Ευφυΐας είναι δομημένα σε μια σειρά από επάλληλα επίπεδα, τα οποία συγκροτούν μια πυραμίδα. Στη βάση της πυραμίδας βρίσκονται τα αρχικά ακατέργαστα δεδομένα, ενώ στην κορυφή της βρίσκεται η λήψη των τελικών αποφάσεων. Κάθε μετάβαση από ένα επίπεδο σε κάποιο ανώτερο, αυξάνει τη δυνατότητα υποστήριξης επιχειρηματικών αποφάσεων. Η πυραμίδα Συστημάτων Επιχειρηματικής Ευφυΐας παρουσιάζεται στην Εικόνα 1.1

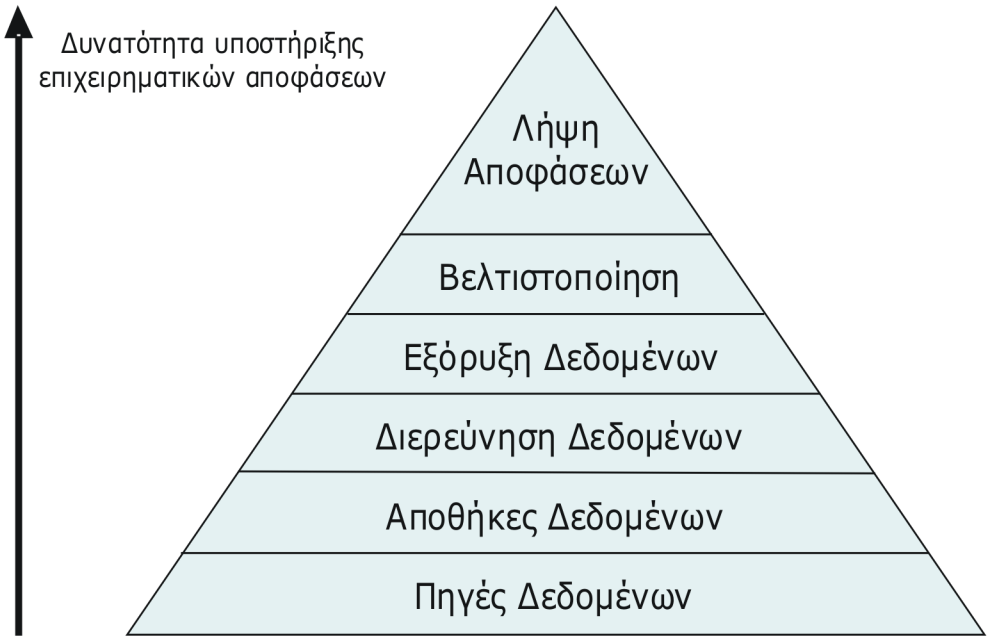
1.3.1 Πηγές Δεδομένων

Στη βάση της πυραμίδας βρίσκονται οι πηγές των αρχικών δεδομένων. Τα δεδομένα αυτά προέρχονται κυρίως από συστήματα παρακολούθησης συναλλαγών, όπως πχ τα συστήματα ERP, και από εταιρικές βάσεις δεδομένων. Άλλες πρόσθετες πηγές δεδομένων είναι οι εταιρικοί δικτυακοί servers, εσωτερικά έγγραφα ή και εξωτερικές πηγές. Τα δεδομένα αυτά μπορεί να είναι σημαντικά για την καθημερινή λειτουργία της επιχείρησης, είναι όμως ακατάλληλα για τη λήψη αποφάσεων. Η πληροφορία ότι το ταμείο No. 3 ενός υποκαταστήματος super market εξέδωσε απόδειξη για την πώληση ενός κουτιού καφέ, μια συγκεκριμένη ημέρα και ώρα, είναι σημαντική για το λογιστήριο και την αποθήκη, είναι όμως αδιάφορη για τη διοίκηση. Αυτό που ενδιαφέρει τη διοίκηση είναι οι συγκεντρωτικές πωλήσεις καφέ, σε μια γεωγραφική περιοχή και σε μια χρονική περίοδο. Τα λειτουργικά δεδομένα είναι υπερβολικά αναλυτικά και για τον λόγο αυτό, ακατάλληλα για επεξεργασία και εξαγωγή συμπερασμάτων. Επίσης, τα δεδομένα αυτά είναι διάσπαρτα σε διάφορες πηγές και πρέπει να ενοποιηθούν. Τέλος, τα δεδομένα μπορεί να έχουν διαφόρων ειδών προβλήματα, τα οποία πρέπει να αντιμετωπιστούν. Αναλυτικότερη παρουσίαση αυτού του θέματος γίνεται στο Κεφάλαιο 7.

1.3.2 Αποθήκες Δεδομένων

Το επόμενο επίπεδο είναι αυτό των Αποθηκών Δεδομένων. Πρόκειται για βάσεις δεδομένων που περιέχουν τα ενοποιημένα, συγκεντρωτικά και καθαρά δεδομένα. Αυτά τα δεδομένα θα χρησιμοποιηθούν για την ανάλυση και την εξαγωγή συμπερασμάτων.

Οι εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης των δεδομένων στις Αποθήκες, γνωστές και ως εργασίες ETL (Extract, Transform, Load), εκτελούνται σε τακτά χρονικά διαστήματα. Στα πλαίσια των εργασιών αυτών, επιλέγονται καταρχήν τα λειτουργικά δεδομένα που είναι σχετικά με την ανάλυση που πρέπει να πραγματοποιηθεί. Οι Αποθήκες Δεδομένων είναι θεματικά προσανατολισμένες, επικεντρώνονται δηλαδή σε θεματικές περιοχές, όπως πχ πελάτες ή προμηθευτές. Για τον λόγο αυτό, πρέπει να περιληφθούν τα σχετικά δεδομένα και να αποκλειστούν τα μη σχετικά. Επίσης τα δεδομένα πρέπει να συνολικοποιηθούν σύμφωνα με θέματα που ενδιαφέρουν τη διοίκηση, όπως πχ πωλήσεις ανά περιοχή ή ανά χρονική περίοδο ή ανά κατηγορία προϊόντος, καθώς επίσης και να οριστεί ο βαθμός λεπτομέρειας ή γενίκευσης, όπως πχ πωλήσεις ανά εβδομάδα ή ανά μήνα ή ανά τρίμηνο. Οι Αποθήκες Δεδομένων εξετάζονται αναλυτικά στο Κεφάλαιο 4.



Εικόνα 1.1 *Η πυραμίδα Συστημάτων Επιχειρηματικής Ευφυΐας*

1.3.3 Διερεύνηση Δεδομένων

Το τρίτο επίπεδο περιλαμβάνει εργασίες αρχικής επεξεργασίας των δεδομένων. Στο στάδιο αυτό ο χρήστης υποβάλλει ερωτήματα (queries) στη βάση δεδομένων, λαμβάνει απαντήσεις και συντάσσει αναφορές. Στις αναφορές μπορεί να περιλαμβάνονται αριθμητικές τιμές αλλά και πίνακες και γραφήματα. Τα γραφήματα μπορούν να αποδώσουν με πιο παραστατικό και ευχάριστο τρόπο την πληροφορία. Γενικώς οι μέθοδοι οπτικοποίησης βοηθούν στην καλύτερη παράθεση και κατανόηση των δεδομένων. Στο στάδιο αυτό μπορεί να γίνει και μια αρχική στατιστική επεξεργασία των δεδομένων. Μπορούν για παράδειγμα να υπολογίζονται μέσοι όροι, τυπικές αποκλίσεις κλπ. Χαρακτηριστικό αυτού του επιπέδου είναι ότι ο χρήστης, σύμφωνα με το σκεπτικό του, αναπτύσσει εκ των προτέρων υποθέσεις και στη συνέχεια χρησιμοποιεί τα εργαλεία ανάλυσης για να επιβεβαιώσει ότι οι υποθέσεις του υποστηρίζονται από τα δεδομένα.

1.3.4 Εξόρυξη Δεδομένων

Στο τέταρτο στάδιο εκτελείται υψηλού επιπέδου ανάλυση των δεδομένων, με τη χρήση των πιο εξελιγμένων τεχνικών. Χρησιμοποιούνται προχωρημένες στατιστικές μέθοδοι, αλλά και μέθοδοι που προέρχονται από την Τεχνητή Νοημοσύνη και τη Μηχανική Μάθηση. Οι μέθοδοι κατηγοριοποίησης (classification) επιτρέπουν την πρόβλεψη της κατηγορίας στην οποία ανήκει ένα αντικείμενο με βάση τα χαρακτηριστικά του. Η πρόβλεψη χρεοκοπίας και η εκτίμηση πιστοληπτικής ικανότητας είναι χαρακτηριστικά παραδείγματα εφαρμογής τεχνικών κατηγοριοποίησης. Μέθοδοι ανάλυσης συστάδων (cluster analysis) επιτρέπουν τον εντοπισμό ομάδων ομοειδών αντικειμένων. Ανάλυση συστάδων μπορεί να εφαρμοστεί σε μελέτες τμηματοποίησης της αγοράς, εύρεσης δηλαδή ομάδων πελατών με ομοειδή χαρακτηριστικά. Οι κανόνες συσχέτισης είναι πολύ χρήσιμοι για την ανάλυση του καταναλωτικού καλαθιού (market basket analysis), την εύρεση δηλαδή προϊόντων που πωλούνται συχνά μαζί. Η πληροφορία αυτή μπορεί να είναι χρήσιμη για τη διαμόρφωση των ραφιών σε super market.

Ένα χαρακτηριστικό που συναντάται συχνά στις μεθόδους αυτού του επιπέδου είναι ότι ο χρήστης δεν χρειάζεται να διατυπώσει δικές του αρχικές υποθέσεις. Οι αλγόριθμοι επεξεργάζονται τα δεδομένα και εξάγουν την πληροφορία απευθείας από αυτά. Συχνά το αποτέλεσμα είναι ένα μοντέλο. Για παράδειγμα ένα δένδρο απόφασης μπορεί να περιγράφει τα χαρακτηριστικά των αγοραστών μιας κατηγορίας προϊόντων, πχ τετρακίνητων αυτοκινήτων. Ο αλγόριθμος θα διαβάσει τα στοιχεία των πωλήσεων, θα εντοπίσει τα κοινά χαρακτηριστικά των καταναλωτών του συγκεκριμένου προϊόντος και θα κατασκευάσει ένα μοντέλο από κανόνες της μορφής εάν-τότε, οι οποίοι θα περιγράφουν ποιοι αγοράζουν το προϊόν και με ποια πιθανότητα. Ο χρήστης δεν χρειάζεται να διατυπώσει καμία αρχική υπόθεση.

1.3.5 Βελτιστοποίηση

Η λήψη αποφάσεων είναι μια διαδικασία επιλογής. Οι αναλύσεις που πραγματοποιήθηκαν στα χαμηλότερα επίπεδα αποφέρουν μια σειρά ενδεχόμενων λύσεων. Ο αποφασίζων καλείται να επιλέξει μια από τις πολλές εναλλακτικές λύσεις. Ως προς το πλήθος των πιθανών λύσεων, τα προβλήματα χωρίζονται σε τρεις κατηγορίες. Τα διχότομα προβλήματα μπορούν να έχουν δύο δυνατές λύσεις, πχ έγκριση του δανείου ή απόρριψη της αίτησης. Τα προβλήματα πολλαπλών λύσεων μπορούν να έχουν έναν περιορισμένο αριθμό ενδεχόμενων λύσεων. Η επιλογή ενός προμηθευτή μέσα από ένα σύνολο υποψήφιων προμηθευτών είναι τέτοιου είδους πρόβλημα. Τέλος, υπάρχουν προβλήματα απεριόριστου αριθμού ενδεχόμενων λύσεων. Αντικείμενο των εργασιών αυτού του επιπέδου είναι ο εντοπισμός της βέλτιστης λύσης. Υπάρχουν διάφορες μέθοδοι για την επιλογή της βέλτιστης απόφασης. Μεταξύ άλλων, επιλέγουμε να αναφέρουμε τον Γραμμικό Προγραμματισμό και τις ευρετικές μεθόδους (heuristics).

1.3.6 Λήψη απόφασης

Στο κορυφαίο επίπεδο της πυραμίδας γίνεται η λήψη της οριστικής απόφασης. Στο σημείο αυτό, είναι σημαντικό να τονιστεί ότι όλες οι μέθοδοι και τα συστήματα που αναφέρονται παραπάνω, έχουν στόχο την υποβοήθηση ενός ανθρώπου στη λήψη της απόφασης και όχι την αυτοματοποιημένη λήψη απόφασης από έναν υπολογιστή. Πρόκειται ουσιαστικά για εργαλεία ανάλυσης δεδομένων και παραγωγής πληροφοριών. Η τελική απόφαση λαμβάνεται από άνθρωπο, ο οποίος φέρει και την ευθύνη για αυτήν την απόφαση. Ο άνθρωπος, όταν λαμβάνει μια απόφαση, διευκολύνεται στην εργασία του εάν χρησιμοποιήσει περίτεχνα εργαλεία, τα οποία θα του προσφέρουν κατάλληλη πληροφόρηση. Την πληροφόρηση αυτή θα τη χρησιμοποιήσει σε συνδυασμό με τη δική του λογική, τη γνώση και τις ικανότητες του. Πέρα όμως από αυτά, ο άνθρωπος διαθέτει και άλλες ικανότητες και ιδιότητες, τις οποίες μπορεί να επιστρατεύσει. Τέτοιες είναι η φαντασία, το ένστικτο, η διαίσθηση καθώς και πλευρές του χαρακτήρα του.

1.4 Οφέλη και Περιορισμοί της Επιχειρηματικής Ευφυΐας

Τα Συστήματα Επιχειρηματικής Ευφυΐας αξιοποιούν τεχνολογίες της Πληροφορικής για να επεξεργαστούν δεδομένα, να παράξουν πληροφορία και να συνδράμουν τη διοίκηση στον έλεγχο και την καλύτερη λειτουργία ενός οργανισμού. Όπως κάθε τεχνολογική λύση, μπορούν να προσφέρουν πολλά οφέλη, ταυτόχρονα όμως υπόκεινται σε περιορισμούς.

1.4.1 Οφέλη της Επιχειρηματικής Ευφυΐας

Τα βασικά οφέλη που προσφέρουν τα συστήματα Επιχειρηματικής Ευφυΐας είναι τα ακόλουθα:

- Καλύτερη κατανόηση πελατών, αγορών, ανταγωνιστών, προμηθειών και πόρων. Η κατάλληλη οργάνωση των δεδομένων και τα εξελιγμένα εργαλεία πληροφορικής δίνουν πρωτόγνωρες δυνατότητες στην εμβάθυνση όλων των παραπάνω ζητημάτων.
- Τροφοδότηση της διοίκησης με τη σωστή πληροφόρηση, την κατάλληλη στιγμή και με τον κατάλληλο τρόπο. Τα συστήματα της Ε.Ε. μπορούν να αναδείξουν την ουσιαστική πληροφορία. Ταυτόχρονο και βασικό μέλημα όμως είναι και η έγκαιρη πληροφόρηση.
- Βελτίωση της ποιότητας των αποφάσεων. Η αναβαθμισμένη και έγκαιρη πληροφόρηση επιτρέπει στη διοίκηση του οργανισμού να λάβει βελτιωμένες αποφάσεις.
- Συμβολή στη διαμόρφωση των στρατηγικών στόχων. Τα συστήματα Ε.Ε. απευθύνονται κυρίως στα υψηλά ή και κορυφαία στελέχη των επιχειρήσεων. Στο επίπεδο αυτό λαμβάνονται οι στρατηγικές αποφάσεις. Η διοίκηση αξιοποιεί τα συστήματα ΕΕ για την άντληση ποιοτικής πληροφόρησης και τον καθορισμό των στρατηγικών στόχων.
- Επίτευξη συγκριτικού πλεονεκτήματος. Η εξασφάλιση συγκριτικού πλεονεκτήματος αποτελεί μόνιμη επιδίωξη κάθε επιχείρησης. Η βελτίωση των αποφάσεων και μέσω αυτού η αύξηση της αποτελεσματικότητας και αποδοτικότητας της διοίκησης, καθώς και ο καθορισμός σωστών στρατηγικών στόχων, μπορούν να αποτελέσουν το συγκριτικό πλεονέκτημα και να οδηγήσουν σε αυξημένη ανταγωνιστικότητα.
- Δυνατότητες αύξησης της κερδοφορίας, μείωσης του κόστους και βελτίωσης της αποδοτικότητας. Η βελτίωση της πληροφόρησης σχετικά με τη διαχείριση της εφοδιαστικής αλυσίδας μπορεί να βο-

ηθήσει στη συμπίεση του κόστους, ενώ η κατανόηση των αγορών μπορεί να αυξήσει τις πωλήσεις και τα κέρδη. Γενικώς, επιτυχημένα συστήματα ΕΕ συμβάλλουν στην αύξηση των επιδόσεων και της κερδοφορίας.

- Αύξηση της πιθανότητας πρόβλεψης συμβάντων και επιχειρηματικών ευκαιριών. Η βαθύτερη κατανόηση της αγοράς επιτρέπει τον εντοπισμό επιχειρηματικών ευκαιριών. Επιπλέον, οι μέθοδοι προγνωστικής ανάλυσης (predictive analytics) επεξεργάζονται ιστορικά δεδομένα και επιτρέπουν τη διατύπωση προβλέψεων.
- Μεγαλύτερη αξιοποίηση των δεδομένων και αύξηση της απόδοσης της επένδυσης σε τεχνολογίες πληροφορικής. Οι σημερινές επιχειρήσεις έχουν επενδύσει εκατομμύρια ευρώ σε πληροφοριακά συστήματα. Τα δεδομένα αυτών των συστημάτων μπορούν να αποδειχθούν πολύτιμη πηγή πρόσθετης, μη συμβατικής πληροφόρησης, εάν αξιοποιηθούν με τη χρήση της Επιχειρηματικής Ευφυΐας. Με τον τρόπο αυτό, οι επενδύσεις πληροφορικής αποδίδουν πρόσθετους καρπούς.

1.4.2 Περιορισμοί της Επιχειρηματικής Ευφυΐας

Η ανάπτυξη συστημάτων Επιχειρηματικής Ευφυΐας έχει να αντιμετωπίσει διάφορους ανασχετικούς παράγοντες, προβλήματα και ενδεχόμενους κινδύνους:

- Κόστος απόκτησης και λειτουργίας Αποθηκών Δεδομένων και συστημάτων ΕΕ. Απαιτούνται επενδύσεις σε υλικό, λογισμικό και τεχνογνωσία. Επίσης οι εργασίες ETL είναι χρονοβόρες, δύσκολες και δαπανηρές. Όλα τα παραπάνω επιφέρουν ένα όχι ευκαταφρόνητο κόστος, το οποίο πρέπει να αναλάβει η επιχείρηση.
- Χαμηλή ποιότητα δεδομένων. Το πρόβλημα αυτό είναι ένα από τα σημαντικότερα στην ανάπτυξη συστημάτων ΕΕ. Τα αρχικά δεδομένα είναι διάσπαρτα, ανομοιογενή, ελλιπή και πιθανώς λανθασμένα ή αντιφατικά. Τροφοδότηση του συστήματος με προβληματικά δεδομένα θα οδηγήσει σε εσφαλμένη πληροφόρηση. Όπως χαρακτηριστικά λέγεται «garbage in, garbage out».
- Ζητήματα συμβατότητας με τα υπάρχοντα συστήματα. Τα συστήματα ΕΕ λειτουργούν επί δεδομένων άλλων συστημάτων. Τα συστήματα αυτά μπορεί να είναι πολλά, διαφορετικά, και πιθανότατα δεν έχει ληφθεί εκ των προτέρων καμία πρόνοια για ενοποίηση των δεδομένων τους. Μπορεί να εμφανιστούν προβλήματα συμβατότητας, τόσο μεταξύ των βασικών συστημάτων όσο και μεταξύ αυτών και του συστήματος ΕΕ.
- Πιθανή ύπαρξη επιφυλάξεων, δυσπιστίας και μη συνεργασίας από την πλευρά των στελεχών. Η ανάπτυξη συστημάτων Ε.Ε. επιφέρει αλλαγές σε λειτουργίες των οργανισμών. Έχει παρατηρηθεί ότι τέτοιες αλλαγές μπορεί να προκαλέσουν τις επιφυλάξεις και τη δυσπιστία των εμπλεκόμενων στελεχών. Είναι πολύ σημαντικό, τα ανώτατα στελέχη της διοίκησης να εφαρμόσουν πολιτικές διαχείρισης της αλλαγής (change management) και να επιληφθούν τέτοιων προβλημάτων.
- Προβλήματα επικοινωνίας και συνεννόησης μεταξύ των στελεχών και των ειδικών πληροφορικής. Τα στελέχη της επιχείρησης και οι ειδικοί της πληροφορικής έχουν ο καθένας τη δική του οπτική γωνία. Τα στελέχη επικεντρώνονται στα επιχειρησιακά ζητήματα, ενώ οι ειδικοί πληροφορικής στα τεχνικά. Αυτό μπορεί να προκαλέσει προβλήματα συνεννόησης. Ειδικά στα συστήματα ΕΕ, όπου τα επιχειρησιακά ζητήματα παίζουν βαρύνοντα ρόλο, το πρόβλημα αυτό μπορεί να ενταθεί.
- Ανάγκη ειδικά εκπαιδευμένου προσωπικού. Πρέπει να προσληφθεί νέο προσωπικό, αλλά κυρίως πρέπει τα στελέχη να μάθουν να χρησιμοποιούν, με τον βέλτιστο τρόπο, τα νέα αυτά συστήματα.
- Κίνδυνος υπερβολικής και άκριτης εμπιστοσύνης στο σύστημα ΕΕ και συνακόλουθης επανάπαυσης. Έχει ήδη τονιστεί ότι ο τελικός υπεύθυνος για τη λήψη των αποφάσεων είναι ο άνθρωπος. Συστήματα ευφυούς ανάλυσης των δεδομένων και κυρίως συστήματα ικανά να διατυπώνουν προβλέψεις, μπορεί μετά από κάποιον χρόνο να εμπνεύσουν υπερβολική εμπιστοσύνη στους χρήστες τους. Τα στελέχη δεν πρέπει να επαναπαύονται στις προβλέψεις του συστήματος, και πρέπει να αντιμετωπίζουν την πληροφόρηση στη βάση της δικής τους υποκειμενικής κρίσης.
- Πολλές περιπτώσεις αποτυχίας σε έργα ΕΕ. Τα έργα Επιχειρηματικής Ευφυΐας έχουν να αντιμετωπίσουν πολλές προκλήσεις. Ως αποτέλεσμα αυτού του γεγονότος καταγράφεται μεγάλο ποσοστό αποτυχίας έργων επιχειρηματικής ευφυΐας. Σύμφωνα με τον Saran (2012), ο οποίος επικαλείται πηγές του οίκου Gartner, λιγότερο από το 30% των έργων ΕΕ επιτυγχάνει τους σκοπούς του.

1.5 Η Επιχειρηματική Ευφυΐα στην Πράξη

Δεδομένου ότι κάθε δραστηριότητα μιας επιχείρησης απαιτεί τη λήψη αποφάσεων, η Επιχειρηματική Ευφυΐα μπορεί να βρει αντίστοιχες δυνατότητες εφαρμογής. Υπό την έννοια αυτή τα πεδία εφαρμογής της Επιχειρηματικής Ευφυΐας στη σύγχρονη επιχείρηση μπορεί να είναι εξαιρετικά ποικίλα και θεωρητικά απεριόριστα. Στο σημείο αυτό θα επιχειρήσουμε μια χοντρική κατηγοριοποίηση και παρουσίαση των συνηθέστερων πεδίων εφαρμογής.

1.5.1 Διοίκηση Επιχειρησιακής Απόδοσης

Σύμφωνα με το γλωσσάριο του οίκου Gartner, η Διοίκηση Επιχειρησιακής Απόδοσης (ΔΕΑ) (Corporate Performance Management (CPM)) ("CPM (corporate performance management) - Gartner IT Glossary," n.d.) είναι ένα σύνολο μεθοδολογιών, μετρικών, διαδικασιών και συστημάτων, τα οποία επιτρέπουν στα διευθυντικά στελέχη ενός οργανισμού να ελέγχουν και να διαχειρίζονται την απόδοση του. Στη σύγχρονη εποχή, η ΔΕΑ υλοποιείται με τη χρήση κατάλληλου λογισμικού. Εφαρμογές κατάλληλες για ΔΕΑ αντιστοιχούν στρατηγική πληροφορία στα επιχειρησιακά σχέδια και παράγουν συγκεντρωτικά αποτελέσματα. Οι εφαρμογές αυτές ολοκληρώνονται με τις διαδικασίες σχεδιασμού και ελέγχου του οργανισμού.

Απαραίτητο στοιχείο για τη ΔΕΑ είναι οι λεγόμενοι Κύριοι Δείκτες Επιδόσεων (ΚΔΕ) (Key Performance Indicators (KPI)). Οι ΚΔΕ είναι καλά καθορισμένοι δείκτες, οι οποίοι αποτυπώνουν την επίδοση του οργανισμού σε σχέση με κάποια δραστηριότητα του. Οι δραστηριότητες αυτές συνηθέστερα αφορούν την εκπλήρωση κάποιου στρατηγικού στόχου ή σχετίζονται με παράγοντες που είναι ζωτικής σημασίας για τον οργανισμό. Οι επιχειρήσεις χρησιμοποιούν τους ΚΔΕ για να ελέγχουν και να μετρούν τον βαθμό επίτευξης στρατηγικών και επιχειρησιακών στόχων.

Ο καθορισμός των κατάλληλων ΚΔΕ δεν είναι μια τετριμμένη εργασία και διαφέρει από επιχείρηση σε επιχείρηση. Οι ΚΔΕ μπορεί να αναφέρονται σε διάφορες δραστηριότητες και λειτουργίες, όπως πχ τις πωλήσεις και τη διαφήμιση, την παραγωγή και τη διοίκηση της εφοδιαστικής αλυσίδας, τα χρηματοοικονομικά και την κερδοφορία, τη διαχείριση ανθρωπίνων πόρων, τη διαχείριση του επιχειρηματικού κινδύνου κλπ. Ένα κρίσιμο ερώτημα είναι το ποιες προϋποθέσεις καθιστούν έναν δείκτη επίδοσης «κύριο». Σε ορισμένες περιπτώσεις η χρήση ΚΔΕ επιβάλλεται από κανονιστικές διατάξεις που διέπουν τη λειτουργία των επιχειρήσεων, όπως ο βρετανικός Companies Act 2006. Εταιρείες συμβούλων και πάροχοι λογισμικού Επιχειρηματικής Ευφυΐας μπορούν να συνδράμουν έναν οργανισμό στη δύσκολη εργασία της επιλογής των κατάλληλων ΚΔΕ. Η Price Waterhouse Coopers (2007) έχει εκδώσει οδηγό για τον καθορισμό ΚΔΕ, στον οποίο παρέχονται οδηγίες για την επιλογή, το περιεχόμενο και τον τρόπο παρουσίασης των ΚΔΕ.

Οι τιμές των ΚΔΕ αντιπαραβάλλονται με προκαθορισμένους στόχους. Τα διευθυντικά στελέχη ορίζουν αρχικά τις τιμές στόχους και στη συνέχεια συγκρίνουν τις τρέχουσες τιμές των ΚΔΕ με τους στόχους. Εάν διαπιστώσουν ότι υπάρχουν υστερήσεις, θα αναζητήσουν τα αίτια και θα προβούν στις αναγκαίες ενέργειες για τη θεραπεία του προβλήματος. Επίσης, μπορεί να αναθεωρήσουν τις τιμές στόχους. Με τον τρόπο αυτό ελέγχουν αλλά και ρυθμίζουν τις επιδόσεις του οργανισμού.

Η Επιχειρηματική Ευφυΐα σχετίζεται άμεσα με τη Διοίκηση Επιχειρησιακής Απόδοσης. Τα συστήματα Επιχειρηματικής Ευφυΐας οφείλουν να συγκεντρώνουν και να προεπεξεργάζονται όλα τα δεδομένα που σχετίζονται με τους ΚΔΕ, να προβαίνουν στον υπολογισμό των τιμών με ταχύτητα και αποτελεσματικότητα και να παρουσιάζουν τα αποτελέσματα με τρόπο κατανοητό. Η παραγόμενη πληροφορία πρέπει να είναι ορθή, έγκαιρη, ουσιαστική και να αποκαλύπτει την πραγματική κατάσταση του υπό διερεύνηση ζητήματος,

1.5.2 Χρηματοοικονομική ανάλυση και διαχείριση

Αντικείμενο είναι ο σχεδιασμός και η παρακολούθηση των χρηματοοικονομικών ροών. Τα στελέχη παρακολουθούν την πορεία των εσόδων και εξόδων της επιχείρησης. Αναλύονται τα εισπρακτέα, τα πληρωτέα και η κατάσταση των αποθεμάτων. Καθίσταται δυνατή η εύκολη σύνταξη χρηματοοικονομικών καταστάσεων με τρέχοντα στοιχεία, ώστε τα στελέχη να εκτιμούν την επίδοση της επιχείρησης. Επίσης, γίνεται σύγκριση με τα μεγέθη του προϋπολογισμού ώστε, αν διαπιστωθούν αποκλίσεις, να ληφθούν οι αναγκαίες μέριμνες. Η διαδικασία ενημέρωσης σε περίπτωση αποκλίσεων μπορεί να είναι και αυτοματοποιημένη.

Αναλυτικότερα, τα συστήματα Επιχειρηματικής Ευφυΐας, για την ανάλυση των χρηματοοικονομικών μεγεθών, παρακολουθούν τα πάγια της επιχείρησης σε όλο τον κύκλο ζωής τους από την απόκτηση μέχρι την απόσβεση. Επίσης, ελέγχουν την κερδοφορία συνολικά, αλλά και ειδικότερα ανά χρονική περίοδο, περιοχή, πελάτες, κατηγορία προϊόντων κλπ. ώστε να εντοπίζονται με αυτόν τον τρόπο τάσεις, δυναμικές και ευκαιρί-

ες. Η παρακολούθηση των εισπρακτέων και πληρωτέων λογαριασμών επιτρέπει την καλύτερη διαχείριση του κεφαλαίου κίνησης και τον έλεγχο των κινδύνων που αφορούν τις απαιτήσεις. Τα τρέχοντα στοιχεία συγκρίνονται με ιστορικά στοιχεία προηγούμενων ετών και με τιμές στόχους, ώστε να παρέχεται πληρέστερη εικόνα για την πορεία της επιχείρησης και τις χρηματοοικονομικές της επιδόσεις.

1.5.3 Πωλήσεις

Τα Συστήματα Επιχειρηματικής Ευφυΐας διευκολύνουν την παρακολούθηση και τον έλεγχο του κρίσιμου τομέα των πωλήσεων, δίνοντας έτσι τη δυνατότητα στις επιχειρήσεις να ανταγωνιστούν αποτελεσματικότερα μέσα στις αγορές. Αναλύονται τα στοιχεία του αγωγού πωλήσεων, από το στάδιο των αρχικών επαφών με τους εν' δυνάμει πελάτες μέχρι την τελική πώληση. Τα στοιχεία αυτά συγκρίνονται με τις τιμές στόχους και εκτιμάται η πορεία των πωλήσεων, ώστε να ληφθούν κατάλληλα μέτρα σε περίπτωση που υπάρχει υστέρηση. Η ανάλυση του αγωγού πωλήσεων μπορεί να αναδείξει και νέες ευκαιρίες. Επίσης η ανάλυση των ιστορικών και άλλων στοιχείων επιτρέπει την ακριβέστερη πρόβλεψη του ύψους των μελλοντικών πωλήσεων. Ένας άλλος σχετικός τομέας είναι αυτός της διαχείρισης του δυναμικού του τμήματος πωλήσεων. Η ανάλυση των στοιχείων μπορεί να γίνει σε διάφορα επίπεδα που να φθάνουν μέχρι τις ατομικές επιδόσεις των πωλητών. Η διοίκηση εντοπίζει τα ισχυρά σημεία αλλά και τις αδυναμίες και στη συνέχεια αξιοποιεί αυτήν την πληροφόρηση και προβαίνει στις αναγκαίες δράσεις, ώστε να επιτευχθεί η διάχυση των βέλτιστων πρακτικών και η αντιμετώπιση προβλημάτων.

1.5.4 Marketing

Η επεξεργασία των στοιχείων που αφορούν τους πελάτες και η άντληση πολύτιμης σχετικής πληροφορίας είναι από τα σημαντικότερα και αποδοτικότερα πεδία εφαρμογής της Επιχειρηματικής Ευφυΐας. Βασικός στόχος είναι η κατανόηση της αγοραστικής συμπεριφοράς των καταναλωτών και η αναγνώριση των αναγκών και των προτιμήσεων τους. Οι πληροφορίες αυτές επιτρέπουν την προώθηση των πωλήσεων και την αξιοποίηση νέων ευκαιριών. Επιπλέον, με τη χρήση των τεχνικών Επιχειρηματικής Ευφυΐας μπορεί να γίνει πολύ επιτυχημένη ανάλυση τμηματοποίησης της αγοράς, εντοπισμός δηλαδή συνόλων πελατών με ομοειδή χαρακτηριστικά και καταναλωτική συμπεριφορά. Αυτή η πληροφορία αξιοποιείται με τη διοργάνωση στοχευμένων διαφημιστικών εκστρατειών. Η αξιολόγηση των αποτελεσμάτων διαφημιστικών εκστρατειών είναι ένας ακόμα τομέας που διευκολύνεται με τη χρήση της Επιχειρηματικής Ευφυΐας. Επιλεγμένες διαφημιστικές δράσεις αποτιμώνται σε σχέση με το κόστος τους και τα οφέλη που απέφεραν, και γίνεται σύγκριση των πραγματικών αποτελεσμάτων με τα προϋπολογισμένα μεγέθη. Με τον τρόπο αυτό επιτυγχάνεται βελτιστοποίηση των διαφημιστικών πρακτικών.

1.5.5 Διαχείριση Εφοδιαστικής Αλυσίδας.

Αντικείμενο είναι η καλύτερη διαχείριση της Εφοδιαστικής Αλυσίδας με την παραγωγή και διάχυση των κατάλληλων πληροφοριών. Γίνεται αποτελεσματικός έλεγχος των επιπέδων των αποθεμάτων, σε συνδυασμό με τις ανάγκες σε υλικά απαραίτητα για την παραγωγή προϊόντων. Εντοπίζονται έγκαιρα και αντιμετωπίζονται ελλείψεις και καθυστερήσεις σε παραγγελίες, ώστε να μην επιβραδύνεται η παραγωγή. Με τον τρόπο αυτό γίνεται καλύτερος έλεγχος της ροής των προϊόντων, αυξάνεται η ικανοποίηση του πελάτη με την έγκαιρη παράδοση και μειώνονται οι ακυρώσεις και οι επιστροφές. Η Επιχειρηματική Ευφυΐα βρίσκει εφαρμογή επίσης στην επιλογή προμηθευτών. Αναλύονται τα ιστορικά στοιχεία των προμηθευτών σχετικά με την ποιότητα των προϊόντων και υπηρεσιών, τους χρόνους παράδοσης, τη συνέπεια, τις τιμολογιακές πολιτικές και τις εκπτώσεις και προσφορές τους κλπ. Επίσης, μπορεί να αξιοποιηθούν και εξωτερικά στοιχεία σχετικά με τους υποψήφιους προμηθευτές που να αφορούν την επιχειρηματική δυναμική τους, τη χρηματοοικονομική τους κατάσταση κλπ.

1.5.6 Διαχείριση Ανθρωπίνων Πόρων

Ζητήματα στελέχωσης της επιχείρησης με ανθρώπινο δυναμικό, αμοιβών και παραγωγικότητας περιλαμβάνονται στα τυπικά αντικείμενα που καλύπτονται από τα συστήματα Επιχειρηματικής Ευφυΐας. Η διοίκηση μπορεί ευκολότερα να διαχειριστεί θέματα μισθοδοσίας όπως αμοιβές, φόρους, ασφαλιστικές εισφορές, υπερωρίες κλπ. Επίσης, επιτυγχάνεται καλύτερος έλεγχος της παραγωγικότητας με υπολογισμό του παραγωγικού

και μη παραγωγικού χρόνου, χρόνους προσέλευσης και αποχώρησης, εντοπισμός των πλέον παραγωγικών εργαζομένων και των ταλέντων, καθώς και ο σχεδιασμός πολιτικών για τη συγκράτηση και εξέλιξη των ταλαντούχων εργαζομένων.

Καθίσταται ευκολότερος ο σχεδιασμός και η σύγκριση διαφορετικών πλάνων, για την κάλυψη των αναγκών σε εργατικό δυναμικό με εναλλακτικούς τρόπους, όπως πρόσληψη μόνιμου ή εποχιακού προσωπικού, πλήρους ή μερικής απασχόλησης, υπερωρίες, εσωτερική κινητικότητα κλπ. Τα προγράμματα διαχείρισης ανθρωπίνων πόρων μπορούν να ποσοτικοποιηθούν και να συγκριθούν ως προς τις οικονομικές και λειτουργικές επιπτώσεις τους. Επιτυγχάνεται η πρόβλεψη των αναγκών σε εργατικό δυναμικό με ανάλυση στοιχείων για συνταξιοδοτήσεις, αποχωρήσεις, επαναπροσλήψεις, απολύσεις κλπ.

1.5.7 Χρηματοπιστωτικός τομέας

Ο τομέας των χρηματοοικονομικών υπηρεσιών, δηλαδή των τραπεζών και των ασφαλειών, βρέθηκε στο επίκεντρο της πρόσφατης οικονομική κρίσης. Προέκυψε λοιπόν η ανάγκη για στενότερη επιτήρηση και έλεγχο των χρηματοπιστωτικών ιδρυμάτων. Οι νέες κανονιστικές διατάξεις που διέπουν τη λειτουργία τους (Βασιλεία ΙΙΙ κλπ.) επιβάλλουν αυστηρούς όρους καθώς και τη δημοσίευση πλήθους αναφορών σχετικά με τα διαθέσιμα κεφάλαια τους, τις συναλλαγές τους, τις εσωτερικές διαδικασίες, τους πελάτες τους κλπ. Στόχος είναι, τόσο η καλύτερη διαχείριση του επιχειρησιακού κινδύνου (operational risk management) όσο και η αντιμετώπιση του οικονομικού εγκλήματος, όπως πχ του «πλυσίματος χρήματος» και της διαφθοράς. Τα πρόστιμα και τα ποσά για αποζημιώσεις πελατών και επενδυτών που μπορεί να προέρθουν από ανεπαρκή διαχείριση του ρίσκου, είναι δυνατόν σήμερα να ανέρχονται στο ύψος δισεκατομμυρίων ευρώ.

Για την εξυπηρέτηση των παραπάνω στόχων και επιδιώξεων χρειάζεται συγκέντρωση επιπλέον δεδομένων, κατάλληλη ενοποίηση τους και ιδιαίτερα αποτελεσματική ανάλυση και αξιοποίηση τους. Τα συστήματα Επιχειρηματικής Ευφυΐας έχουν ακριβώς αυτό το αντικείμενο και είναι τα πλέον κατάλληλα για την ικανοποίηση αυτών των απαιτήσεων. Οι μεθοδολογίες που προσφέρει η Εξόρυξη Δεδομένων είναι ιδιαίτερα ικανές να δίνουν λύσεις σε προβλήματα, όπως η εκτίμηση της πιστοληπτικής ικανότητας των πελατών, η διαχείριση του κινδύνου, η αντιμετώπιση του οικονομικού εγκλήματος και ο εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων. Επιπλέον, η οργανωμένη και συγκεντρωτική διαχείριση των δεδομένων διευκολύνει τη σύνταξη των αναφορών (reports) που απαιτούνται από τη νομοθεσία.

Πρέπει να γίνει κατανοητό ότι η παραπάνω παρουσίαση πεδίων εφαρμογής της Επιχειρηματικής Ευφυΐας είναι ενδεικτική και όχι εξαντλητική. Κατασκευαστές λογισμικού Επιχειρηματικής Ευφυΐας παρέχουν διαφορετικά προϊόντα και προσφέρουν ποικίλες λύσεις για διάφορα πεδία εφαρμογής. Οι ιστοθέσεις κατασκευαστών λογισμικού, όπως η ιστοθέση της Oracle (“Oracle Business Intelligence Applications,” n.d.) και η ιστοθέση της SAP (SAP, 2015), περιέχουν αναλυτικές παρουσιάσεις λογισμικών για εξειδικευμένα πεδία εφαρμογής.

1.6 Πάροχοι λογισμικού και υπηρεσιών Επιχειρηματικής Ευφυΐας

Ως συνέπεια της απαίτησης του επιχειρηματικού κόσμου για λύσεις συστημάτων Επιχειρηματικής Ευφυΐας υψηλού επιπέδου, έχει δημιουργηθεί μια αντίστοιχη μεγάλη αγορά με κύκλο εργασιών της τάξης δισεκατομμυρίων ευρώ. Στην αγορά αυτή δραστηριοποιούνται γνωστές και πολύ μεγάλες εταιρείες πληροφορικής, εταιρείες εξειδικευμένες στο λογισμικό στατιστικής ανάλυσης, εταιρείες που πρωτοστατούσαν στον χώρο των βάσεων δεδομένων και κατασκευαστές συστημάτων ERP. Μεταξύ αυτών, εξέχουσα θέση στην προσφορά συστημάτων Επιχειρηματικής Ευφυΐας κατέχουν οι ακόλουθες:

1.6.1 SAS

Η SAS (Statistical Analysis System) είναι μια εταιρεία, που από την ίδρυση της ασχολήθηκε με το λογισμικό στατιστικής ανάλυσης. Σήμερα αποτελεί έναν από τους σημαντικότερους παρόχους συστημάτων Επιχειρηματικής Ευφυΐας. Προσφέρει λογισμικό αναλυτικής των επιχειρήσεων (Business Intelligence and Analytics) με προχωρημένα εργαλεία οπτικοποίησης, εύκολης ανάλυσης, αυξημένες δυνατότητες χρήσης φορητών συσκευών, καθώς και εργαλεία συνεργασίας. Λογισμικό για τη διαχείριση των πελατών και του μάρκετινγκ (Customer Intelligence) αναλύει καταναλωτικές συμπεριφορές, διευκολύνει την προσωπική στόχευση και επιτρέπει τον σχεδιασμό και αποτίμηση των διαφημιστικών εκστρατειών. Εξειδικευμένο λογισμικό ασφάλειας και αντιμετώπισης απάτης (Fraud and Security Intelligence) ανιχνεύει εκ των προτέρων δόλιες πληρωμές με χρήση κανόνων, μεθόδων εντοπισμού ανωμαλιών και προγνωστικής ανάλυσης, και διασφαλίζει τη συμμόρ-

φωση με κανονιστικές διατάξεις, ελέγχοντας συναλλαγές για παράνομες δραστηριότητες. Λογισμικό διαχείρισης επιδόσεων (Performance Management) επιτρέπει τον συνδυασμένο έλεγχο επίτευξης βραχυπρόθεσμων και στρατηγικών στόχων, διευκολύνει τον εντοπισμό ευκαιριών και κινδύνων και την κατανόηση των πηγών κόστους και παραγόμενης αξίας. Το λογισμικό για τη διαχείριση του ρίσκου (Risk Management) ασχολείται με θέματα ιδίων κεφαλαίων, με διαχείριση του πιστωτικού κινδύνου και με δοκιμές αντοχής πιστωτικών ιδρυμάτων. Τέλος, παρέχεται εξειδικευμένο λογισμικό για τη διαχείριση της εφοδιαστικής αλυσίδας (Supply Chain Intelligence). Οι επιχειρηματικές λύσεις που προσφέρει η SAS αντιμετωπίζουν ζητήματα όπως η διαχείριση των δεδομένων (Data Management), η ανάλυση δεδομένων μεγάλου όγκου (Big Data) και η λειτουργία σε περιβάλλον υπολογιστικού νέφους (SAS Cloud Analytics). Συνολικά η εταιρεία διαθέτει περισσότερα από 200 προϊόντα. Ιδιαίτερη μνεία γίνεται στο SAS Enterprise Miner, λογισμικό εξόρυξης δεδομένων για επιχειρήσεις με αυξημένες δυνατότητες περιγραφικής και προγνωστικής μοντελοποίησης.

1.6.2 IBM

Η IBM, εταιρεία σταθμός στην ιστορία της πληροφορικής, έχει αναπτύξει πολύπλευρη δραστηριότητα στον τομέα του υλικού και του λογισμικού, και έχει εισάγει ριζοσπαστικά καινοτόμα προϊόντα, μεταξύ των οποίων και το περιβόητο IBM Personal Computer, το οποίο αποτέλεσε πρότυπο για τους μελλοντικούς προσωπικούς υπολογιστές (PCs). Η IBM διαθέτει μακροχρόνια εμπειρία στον τομέα της τεχνητής νοημοσύνης και έχει να επιδείξει διάφορα πρωτοποριακά σχετικά προϊόντα όπως ο υπολογιστής Deep Blue, ο οποίος νίκησε τον παγκόσμιο πρωταθλητή σκακιού Kasparov και το σύστημα Watson, το οποίο το 2011 αντιμετώπισε στο τηλεοπτικό κουίζ Jeopardy προηγούμενους νικητές. Επιπλέον, πρόσφατα, με μια σειρά εξαγορών, η IBM απόκτησε διάφορες εταιρείες το αντικείμενο των οποίων άπτεται των συστημάτων Επιχειρηματικής Ευφυΐας. Τέτοιες περιπτώσεις είναι η εταιρεία συστημάτων Επιχειρηματικής Ευφυΐας και διαχείρισης επίδοσης Cognos, η εταιρεία στατιστικού λογισμικού SPSS, η εταιρεία αποθηκών δεδομένων Netezza, καθώς και πολλές άλλες.

Σήμερα η IBM θεωρείται ένας από τους μεγαλύτερους παρόχους συστημάτων Επιχειρηματικής Ευφυΐας και προσφέρει έναν μακρύ κατάλογο σχετικών προϊόντων και λύσεων. Το λογισμικό IBM SPSS χρησιμοποιείται για διαχείριση δεδομένων, στατιστική ανάλυση, εξόρυξη δεδομένων και κειμένου, βελτιστοποίηση αποφάσεων και συνεργασία. Το IBM Cognos προσφέρει dashboards, scorecards, what-if σενάρια, εργαλεία για σχεδιασμό, προϋπολογισμό και πρόβλεψη, διαχείριση επίδοσης, προχωρημένα εργαλεία οπτικοποίησης, αυτοματοποιημένα εργαλεία για σύνταξη χρηματοοικονομικών αναφορών και πολλά άλλα. Το νέο σύστημα Watson Analytics προσφέρει εξελιγμένη ανάλυση των επιχειρηματικών δεδομένων για τον έλεγχο υποθέσεων και απάντηση ερωτημάτων, καθώς επίσης και βελτιωμένα εργαλεία οπτικοποίησης. Το λογισμικό OpenPages έχει αντικείμενο τη διαχείριση του ρίσκου, τη συμμόρφωση με τις νέες κανονιστικές διατάξεις, την αυτοματοποίηση των διαδικασιών χρηματοοικονομικών ελέγχων και τη διευκόλυνση των διαδικασιών εσωτερικού ελέγχου. Το λογισμικό IBM Algorithmics απευθύνεται σε χρηματοοικονομικούς οργανισμούς και προσφέρει λύσεις διαχείρισης ρίσκου για πιστώσεις και ρευστότητα, διαχείρισης κεφαλαίου και υποθηκών, διαχείρισης χαρτοφυλακίου και επενδυτικών αποφάσεων. Η IBM συμμετέχει πρωταγωνιστικά στη διαμόρφωση των νέων τάσεων. Αξιοποιώντας τις τεχνολογίες κινητής υπολογιστικής, προσφέρει μέσω κινητών συσκευών πληροφόρηση σε οποιοδήποτε σημείο. Προϊόντα προσφέρονται υπό το σχήμα «Λογισμικό ως υπηρεσία» (Software As A Service) σε περιβάλλον υπολογιστικού νέφους (Cloud computing).

1.6.3 ORACLE

Η Oracle, πασίγνωστη για την ηγετική της παρουσία στον χώρο των βάσεων δεδομένων, δραστηριοποιείται σήμερα και στον χώρο του υλικού υπολογιστών, κυρίως μετά την εξαγορά της Sun Microsystems, αλλά και στον χώρο του λογισμικού επιχειρησιακών συστημάτων, προσφέροντας λύσεις σχεδιασμού επιχειρησιακών πόρων (ERP), διαχείρισης εφοδιαστικής αλυσίδας (SCM) και διαχείρισης σχέσεων πελατών (CRM). Επίσης, θεωρείται ένας από τους κορυφαίους σύγχρονους παρόχους συστημάτων Επιχειρηματική Ευφυΐας και κάτοχος του μεγαλύτερου τμήματος της σχετικής αγοράς.

Η πλατφόρμα Enterprise Business Intelligence περιλαμβάνει εξελιγμένα εργαλεία ανάλυσης, δημιουργίας αναφορών, υποβολής ερωτημάτων, dashboards και scorecards, πράξεων OLAP, ειδοποίησης σε πραγματικό χρόνο κλπ. Το λογισμικό Oracle Essbase είναι ένας ισχυρός server πολυδιάστατης ανάλυσης και πράξεων OLAP, που επιτρέπει τη γρήγορη ανάπτυξη σύνθετων επιχειρηματικών μοντέλων και τη διεξαγωγή αναλύσεων what-if. Η πλατφόρμα Oracle Advanced Analytics συνδυάζει τη βάση δεδομένων της Oracle με δύο ισχυρότατα εργαλεία ανάλυσης, το Oracle Data Mining για εξόρυξη δεδομένων και προγνωστικές αναλύσεις,

καθώς επίσης και με την ελεύθερη γλώσσα προγραμματισμού R, η οποία χρησιμοποιείται για στατιστικές αναλύσεις και εξόρυξη δεδομένων. Το σύστημα Oracle Exalytics συνίσταται σε μια ολοκληρωμένη λύση, που συνδυάζει υψηλότατης ποιότητας υλικό υπολογιστών (hardware), κορυφαίο λογισμικό Επιχειρηματικής Ευφυΐας και τεχνολογία βάσεων δεδομένων in-memory, συστήματα βάσεων δεδομένων δηλαδή, που λειτουργούν πρωτίστως στην κύρια μνήμη του υπολογιστή, εξασφαλίζοντας πολύ μεγαλύτερη ταχύτητα. Ως προς τις επιχειρηματικές λύσεις που παρέχει η Oracle, αυτές καλύπτουν όλα τα πεδία εφαρμογής που αναφέρονται στο υποκεφάλαιο 'Η Επιχειρηματική Ευφυΐα στην Πράξη', δηλαδή χρηματοοικονομική διοίκηση, πωλήσεις, μάρκετινγκ, διαχείριση εφοδιαστικής αλυσίδας, διαχείριση ανθρωπίνων πόρων, χρηματοπιστωτικός τομέας, καθώς και πολλές επιπλέον, όπως διαχείριση ρίσκου και κανονιστική συμμόρφωση, διαχείριση χαρτοφυλακίου, διαχείριση κοινωνικών σχέσεων κλπ. Ως κυρίαρχη δύναμη στον χώρο των βάσεων δεδομένων, η Oracle διαθέτει εξαιρετική τεχνογνωσία σε ζητήματα διαχείρισης δεδομένων, τεχνογνωσία την οποία αξιοποιεί και στον νέο χώρο του Big Data. Μια σειρά από εργαλεία και εφαρμογές δίνουν προωθημένες λύσεις σε ζητήματα Big Data. Επίσης, η Oracle τα τελευταία χρόνια έχει εξαγοράσει πολλές εταιρείες που ασχολούνταν με το υπολογιστικό νέφος, εξασφαλίζοντας έτσι σημαντική παρουσία και σε αυτόν τον χώρο.

1.6.4 SAP

Η SAP είναι μια ευρωπαϊκή εταιρεία που κυριαρχεί στον χώρο των συστημάτων Σχεδιασμού Επιχειρησιακών Πόρων (Enterprise Resources Planning), και είναι ένας από τους μεγαλύτερους παραγωγούς λογισμικού παγκοσμίως. Το 2007 η SAP εξαγόρασε την Business Objects, μια γαλλική εταιρεία εξειδικευμένη στα συστήματα Επιχειρηματικής Ευφυΐας, εντείνοντας την παρουσία της σε αυτόν τον χώρο, και σήμερα θεωρείται μια από τις πρωταγωνίστριες δυνάμεις.

Υπό τον τίτλο SAP Business Objects, η εταιρεία προσφέρει μια σειρά από σουίτες εφαρμογών Επιχειρηματικής Ευφυΐας. Το SAP Business Objects BI platform περιλαμβάνει εργαλεία για πρόσβαση σε δεδομένα διαφόρων κατασκευαστών (IBM, Oracle, Teradata κλπ.), εργαλεία για την αποτελεσματική σύνταξη αναφορών με δυνατότητες επεξεργασίας Big Data και ενσωμάτωσης αναφορών σε εφαρμογές, εργαλεία για τη δημιουργία ισχυρών διαδραστικών dashboards, λογισμικό για την αποτελεσματική και γρήγορη απάντηση επιχειρηματικών ερωτήσεων καθώς και λύσεις κινητής υπολογιστικής που διανέμουν πληροφόρηση σε φορητές συσκευές. Η έκδοση Analytics Edition συνδυάζει την ολοκλήρωση και διαχείριση δεδομένων με εξελιγμένο λογισμικό Επιχειρηματικής Ευφυΐας. Κάνοντας χρήση προχωρημένων αναλυτικών μεθόδων επιτρέπει την αναγνώριση τάσεων και εξαιρέσεων, την αξιοποίηση επιχειρηματικών ευκαιριών και την έγκαιρη αντιμετώπιση κινδύνων. Η έκδοση OLAP edition προσφέρει εργαλεία πολυδιάστατης ανάλυσης. Το λογισμικό SAP Crystal Reports έχει αντικείμενο τη δημιουργία καλαίσθητων αναφορών με δυνατότητα επεξεργασίας δεδομένων από διάφορες πηγές, ενώ το SAP Lumira περιλαμβάνει εξελιγμένα εργαλεία οπτικοποίησης. Τα συστήματα Επιχειρηματικής Ευφυΐας της SAP δίνουν δυνατότητες προγνωστικής ανάλυσης και προσφέρουν λύσεις για τη διαχείριση και έλεγχο της επίδοσης της επιχείρησης, καθώς και για τον έλεγχο του ρίσκου και την κανονιστική συμμόρφωση.

1.6.5 Microsoft

Η Microsoft, ο μεγαλύτερος κατασκευαστής λογισμικού παγκοσμίως ως προς τα έσοδα, είναι ευρύτερα γνωστή κυρίως για το λειτουργικό σύστημα Windows και τη σουίτα εφαρμογών αυτοματισμού γραφείου MS Office. Επίσης, η παιχνιδομηχανή Xbox και τα tablets Microsoft Surface είναι πολύ γνωστά προϊόντα hardware. Στον μακρύ κατάλογο προϊόντων λογισμικού της εταιρείας περιλαμβάνονται και εφαρμογές για επιχειρήσεις, όπως συστήματα ERP και λογισμικό Επιχειρηματικής Ευφυΐας. Δύο προϊόντα της, η βάση δεδομένων SQL Server και το Microsoft Office, ειδικότερα η εφαρμογή φύλλων εργασίας Excel και το πρόγραμμα δημιουργίας παρουσιάσεων Power Point, έπαιξαν σημαντικό ρόλο στην καθιέρωση της ως ένας από τους βασικούς παρόχους λογισμικού Επιχειρηματικής Ευφυΐας.

Η βάση δεδομένων SQL Server και ειδικότερα η έκδοση Business Intelligence, προσφέρει ένα περιβάλλον Επιχειρηματικής Ευφυΐας που επιτρέπει την ταχεία και διαδραστική διερεύνηση και οπτικοποίηση των δεδομένων, τη συγχώνευση δομημένων και αδόμητων δεδομένων και την ταχεία ανάλυση τους με τη χρήση της εγκατεστημένης στη μνήμη αναλυτικής μηχανής (analytics engine). Ο SQL Server Analysis Services δίνει τη δυνατότητα δημιουργίας πολυδιάστατων μοντέλων, και περιλαμβάνει εργαλεία οπτικοποίησης και σύνταξης αναφορών. Επίσης περιλαμβάνονται εργαλεία εξόρυξης δεδομένων για τη διεξαγωγή προγνωστικών αναλύσεων. Τα εργαλεία αυτά είναι διαθέσιμα ως add-ins του Excel αλλά και μέσω του SQL Server Development

Tools για πιο περίτεχνες αναλύσεις. Η πλατφόρμα ανάπτυξης εφαρμογών Microsoft Azure προσφέρει λογισμικό μηχανικής μάθησης για την εξόρυξη δεδομένων και τη διατύπωση προβλέψεων, συνδυασμένο με μια φιλική προς τον χρήστη διεπαφή. Το Azure υποστηρίζει και τη γλώσσα R.

Μεγάλη βαρύτητα δίνει η Microsoft στο υπολογιστικό νέφος και το Big Data. Όλες οι ιστοσελίδες της εταιρείας που αναφέρονται στα συστήματα Επιχειρηματικής Ευφυΐας, τονίζουν με έμφαση τις δυνατότητες αξιοποίησης του νέφους και της λειτουργίας του λογισμικού στα πλαίσια του. Το Microsoft Data Warehouse επιτρέπει τη διαχείριση εξωτερικών δεδομένων μεγάλου όγκου. Τα δομημένα επιχειρηματικά δεδομένα μπορούν εύκολα να συνδυαστούν με αδόμητα δεδομένα από το Hadoop, ώστε να αποτελέσουν μια ολοκληρωμένη βάση πληροφόρησης. Το νέο Office 365, λογισμικό βασισμένο στο νέφος, περιλαμβάνει το Power BI, ένα εύχρηστο περιβάλλον κατάλληλο για εργασίες Επιχειρηματικής Ευφυΐας, προσαρμόσιμες στις μεταβαλλόμενες ανάγκες του χρήστη. Η Microsoft αξιοποιεί και τη βαθιά τεχνογνωσία της στον αυτοματισμό γραφείου. Το Share Point προσφέρει ένα ελκυστικό περιβάλλον για τη δημιουργία και διανομή αναφορών και dashboards. Το Excel, το οποίο στο παρελθόν χρησιμοποιήθηκε κατά κόρον από επιχειρηματικά στελέχη για τη διεξαγωγή αναλύσεων, ενισχύεται με δυνατότητες εξόρυξης δεδομένων. Το ευρύτατα διαδεδομένο Microsoft Office αποτελεί χρήσιμη πλατφόρμα για σύνταξη αναφορών. Ακόμα και τρίτοι κατασκευαστές συστημάτων Επιχειρηματικής Ευφυΐας, όπως η Oracle και η SAP, τονίζουν τη δυνατότητα του λογισμικού τους να συνδεθεί με τα προγράμματα του Office και να ενσωματώσει λειτουργικότητες και αποτελέσματα σε φύλλα εργασίας του Excel, σε παρουσιάσεις του Power Point και σε έγγραφα του Word.

1.6.6 Qlik

Η Qlik είναι μια εταιρεία παραγωγής λογισμικού εξειδικευμένη στα συστήματα Επιχειρηματικής Ευφυΐας. Ιδρύθηκε το 1993 στη Σουηδία και γνώρισε ταχύτατη ανάπτυξη. Σήμερα είναι μια διεθνής εταιρεία με δεκάδες χιλιάδες πελάτες σε περισσότερες από 100 χώρες. Τα βασικά προγράμματα της εταιρείας είναι το QlikView και το QlikSense. Το QlikView είναι μια πλατφόρμα για την ανάπτυξη εφαρμογών Επιχειρηματικής Ευφυΐας. Το λογισμικό διαθέτει μια σειρά από ιδιότητες που το καθιστούν αποτελεσματικό και ελκυστικό. Προβλέπεται διαχείριση των δεδομένων μέσα στη μνήμη ώστε να αυξάνεται η ταχύτητα επεξεργασίας. Υπάρχει δυνατότητα χρήσης του μέσα από internet browsers με τη χρήση κατάλληλων plug-ins. Επίσης αξιοποιείται η κινητή υπολογιστική και η εφαρμογή είναι προσβάσιμη μέσα από κινητές συσκευές όπως tablets και smartphones. Με το QlikView Desktop ο χρήστης μπορεί να αποκτά πρόσβαση σε δεδομένα, να εκτελεί αναλύσεις και να σχεδιάζει αναφορές και dashboards. Το QlikView Workbench είναι ένα plug in για Microsoft Visual Studio, που επιτρέπει την εύκολη ανάπτυξη εφαρμογών για την επέκταση των λειτουργιών του QlikView. Το πρόγραμμα μπορεί να έχει πρόσβαση σε μεγάλους όγκους δεδομένων μέσα από πηγές συμβατές με πρότυπα όπως το ODBC και το XML. Επίσης το πρόγραμμα μπορεί να συνδεθεί με λογισμικά άλλων κατασκευαστών όπως το SAP ERP, το Salesforce και το Informatica.

Το QlikSense είναι μια εφαρμογή οπτικοποίησης δεδομένων και δημιουργίας αναφορών. Ο χρήστης μπορεί με διαδραστικό και εύκολο τρόπο να διερευνά τα δεδομένα, να υποβάλλει ερωτήσεις και να κατασκευάζει dashboards. Το λογισμικό είναι ικανό να συνδυάζει δεδομένα από πολλαπλές πηγές. Επίσης, είναι προσβάσιμο από φορητές συσκευές και προσαρμόζεται αυτόματα σε αυτές. Έχουν προβλεφθεί ιδιαίτερες λειτουργικότητες που διευκολύνουν τη συνεργασία και τη διανομή των αναλύσεων και των πληροφοριών σε ομάδες. Έμφαση έχει δοθεί στην ευχρηστία και την προσαρμοστικότητα του λογισμικού, ώστε κάθε χρήστης να μπορεί να το χειριστεί σύμφωνα με τις επιθυμίες και τις ανάγκες του.

Βιβλιογραφία/Αναφορές

SAP. (2015). *Business Intelligence Tools | BI & Analytics | SAP.* Retrieved 25 May, 2015, from http://go.sap.com/solution/platform-technology/business-intelligence.html

CIOLeadershipForum2015Profile.pdf. (2015). Retrieved 25 May, 2015, from http://www.gartnerinfo.com/cios9/CIOLeadershipForum2015Profile.pdf

CPM (Corporate Performance Management) – Gartner IT Glossary. (n.d.). Retrieved 22 May, 2015, from http://www.gartner.com/it-glossary/cpm-corporate-performance-management

Devens, M. (1865). *Cyclopædia of commercial and business anecdotes*. New York, NY: D. Appleton and Company.

Evtm_219_CIOtop10[3].pdf. (n.d.). Retrieved 25 May, 2015, from http://www.gartnerinfo.com/sym23/evtm_219_CIOtop10%5B3%5D.pdf

Ferrari, A. (2011). Business Intelligence Systems, Uncertainty in Decision-Making and Effectiveness of Organizational Coordination. In A. Carugati & C. Rossignioli (Eds.), *Emerging Themes in Information Systems and Organization Studies* (pp. 155-167). Berlin: Springer – Verlag.

Gartner Executive Programs' Worldwide Survey on More Than 2300 CIOs Shows flat IT Budgets in 2012, but IT Organizations Must Deliver on Multiple Priorities. (n.d.). Retrieved 25 May, 2015, from http://www.gartner.com/newsroom/id/1897514

Information Management | IT Business News. (n.d.). Retrieved 27 December, 2014, from http://www.information-management.com/

Information Week. (n.d.). Retrieved 27 December, 2014, from http://www.informationweek.com/software.asp

Luhn, H. P. (1958). A Business Intelligence System. *IBM Journal of Research and Development*, *2*(4), 314-319.

Mintzberg, H. (1990). *Mintzberg on Management: Inside our Strange World of Organizations*. New York, NY: Free Press.

Oracle Business Intelligence Applications. (n.d.). Retrieved 22 May, 2015, from http://www.oracle.com/technetwork/middleware/bi-applications/overview/index.html

Price Waterhouse Coopers. (2007). *Guide to Performance Indicators*.

Saran, C. (2012). Almost a Third of BI Projects Fail to Deliver on Business Objectives. *Computer Weekly*. Retrieved from http://www.computerweekly.com/news/2240113585/Almost-a-third-of-BI-projects-fail-to-deliver-on-business-objectives

Sabherwal, R., & Beccera – Fernandez, I. (2010). *Business Intelligence*. Hoboken, NJ: John Wiley and Sons Inc.

Scheps, S. (2007). *Business Intelligence for Dummies*. Hoboken, NJ: Willey Publishing Inc.

TDWI | Advancing all things data. (n.d.). Retrieved 27 December, 2014, from http://tdwi.org/Home.aspx

TDAN.com. Retrieved 27 December, 2014, from http://www.tdan.com

Fayol, H. (1949). *General and Industrial Management*. London, UK: Pitman.

2 Συστήματα Υποστήριξης Αποφάσεων

Σύνοψη

Τα Συστήματα Υποστήριξης Αποφάσεων (ΣΥΑ) πρωτοπαρουσιάστηκαν τη δεκαετία του 70 και χρησιμοποιήθηκαν σε μεγάλο βαθμό από τις επιχειρήσεις για την υποβοήθηση της διαδικασίας λήψης αποφάσεων. Το παρόν Κεφάλαιο καλύπτει αυτήν τη θεματική ενότητα. Αρχικά, μελετώνται γενικότερα θέματα που σχετίζονται με τη λήψη αποφάσεων. Εξηγείται η έννοια των Λογικών Αποφάσεων και αναφέρονται οι τύποι λογικής που σχετίζονται με αυτήν. Ακολούθως, παρουσιάζεται το μοντέλο του Simon, σύμφωνα με το οποίο η διαδικασία λήψης αποφάσεων χωρίζεται σε τέσσερα στάδια, το στάδιο της πληροφόρησης, του σχεδιασμού, της επιλογής και της υλοποίησης. Επιπλέον διευκρινίζεται το ειδικό αντικείμενο του κάθε σταδίου. Στη συνέχεια, γίνεται αναφορά στα διάφορα είδη αποφάσεων. Οι αποφάσεις χωρίζονται ως προς τη δομή του προβλήματος σε δομημένες, ημιδομημένες και αδόμητες. Επίσης, ως προς το διοικητικό επίπεδο στο οποίο λαμβάνονται, χωρίζονται σε λειτουργικές, τακτικές και στρατηγικές. Για κάθε μια από αυτές τις κατηγορίες παρέχονται ορισμοί και επεξηγήσεις και αναλύεται το πώς μειώνεται ο βαθμός δόμησης, με τη μετάβαση από το λειτουργικό προς το στρατηγικό επίπεδο. Σύμφωνα με τον Mintzberg, η λήψη αποφάσεων είναι ένα από τα βασικά καθήκοντα των διοικητικών στελεχών. Προκειμένου να λάβουν αποφάσεις, τα στελέχη αναζητούν σχετική πληροφόρηση και κάνουν χρήση πληροφοριακών συστημάτων. Τα συστήματα τα οποία χρησιμοποιούν τα στελέχη, καθώς και ο τρόπος που τα χρησιμοποιούν αναλύονται στα πλαίσια αυτού του κεφαλαίου.
Στη συνέχεια, το κεφάλαιο προχωρά με την καθεαυτό παρουσίαση των ΣΥΑ. Παρατίθενται διάφοροι ορισμοί που έχουν κατά καιρούς προταθεί για τα ΣΥΑ. Τα ΣΥΑ, ως συστήματα ειδικού σκοπού, διαθέτουν ειδικά χαρακτηριστικά. Τα χαρακτηριστικά αυτά των ΣΥΑ παρατίθενται και σχολιάζονται αναλυτικά. Τα Πληροφοριακά Συστήματα Διοίκησης είναι συγγενή συστήματα με τα ΣΥΑ. Οι δύο κατηγορίες συστημάτων μοιράζονται ομοιότητες και διαφορές. Καταγράφονται οι ομοιότητες και οι διαφορές των δύο συστημάτων. Ακολούθως, παρουσιάζεται η δομή των ΣΥΑ με τα επιμέρους υποσυστήματα τους. Ιδιαίτερη αναφορά γίνεται στο Σύστημα Διαχείρισης Βάσης Μοντέλων. Τέλος, γίνεται αρκετά αναλυτική παρουσίαση ειδικών κατηγοριών των ΣΥΑ, όπως είναι τα Συστήματα Υποστήριξης Ομαδικών Αποφάσεων, τα Συστήματα Υποστήριξης Διοίκησης και τα Ευφυή Συστήματα Υποστήριξης Αποφάσεων.

Προαπαιτούμενη γνώση

Τα Συστήματα Υποστήριξης Αποφάσεων είναι ένα ευρύ αντικείμενο με μακρά ιστορία στον χώρο της Πληροφορικής. Στη διάρκεια της πορείας τους, γνώρισαν διάφορα στάδια εξέλιξης και υπάρχει πλούσια βιβλιογραφία που ανταποκρίνεται στα στάδια αυτά και στην αντίστοιχη προβληματική της εποχής. Επίσης, τα ΣΥΑ χρησιμοποιούν έννοιες και σχετίζονται σε μεγαλύτερο ή μικρότερο βαθμό με άλλους επιστημονικούς κλάδους, όπως η λήψη αποφάσεων, οι γνωστικές επιστήμες, οι επιστήμες οργάνωσης, η επιστήμη συστημάτων και η τεχνητή νοημοσύνη. Ο αναγνώστης μπορεί να αναζητήσει γνώσεις υποδομής στην εκτεταμένη βιβλιογραφία που αναφέρεται σε αυτά τα ζητήματα.
Για μια ανασκόπηση της ιστορίας των ΣΥΑ, με καταγραφή των ορόσημων της και ειδική αναφορά στους ερευνητές οι οποίοι έπαιξαν καθοριστικό ρόλο στην εξέλιξη των ΣΥΑ, καθώς και στις απόψεις τους και την προβληματική τους για τις μελλοντικές εξελίξεις, ο αναγνώστης μπορεί να ανατρέξει στη εργασία των Power, Burstein and Sharda (2011). Το βιβλίο του Mintzberg (1975) υπήρξε σταθμός και καθόρισε το μέλλον της επιστήμης της Διοίκησης Επιχειρήσεων, όπως και το βιβλίο του Simon (1977) καθόρισε το μέλλον της λήψης αποφάσεων. Στα δύο αυτά βιβλία ο αναγνώστης θα βρει λεπτομέρειες για βασικές έννοιες που σχετίζονται με τα ΣΥΑ. Οι έννοιες αυτές αναφέρονται στα αρχικά κεφάλαια πολλών συγγραμμάτων για τα ΣΥΑ. Σημαντικές βασικές έννοιες για τη λήψη αποφάσεων και πληροφορίες για τις πρώτες γενιές των ΣΥΑ υπάρχουν στο βιβλίο των Gorry and Scott Morton (1971). Για θέματα λήψης ομαδικών αποφάσεων, ο αναγνώστης μπορεί να ανατρέξει στο βιβλίο του Black (1948). Στους αναγνώστες που ενδιαφέρονται να αναζητήσουν πρόσθετη και αναλυτικότερη πληροφόρηση για σύγχρονα θέματα σχετικά με τα Συστήματα Υποστήριξης Αποφάσεων, υποδεικνύουμε ένα πρόσφατο σύγγραμμα, το βιβλίο των Sharda, Delen and Turban (2015).

2.1 Λήψη Αποφάσεων

2.1.1 Λογικές Αποφάσεις

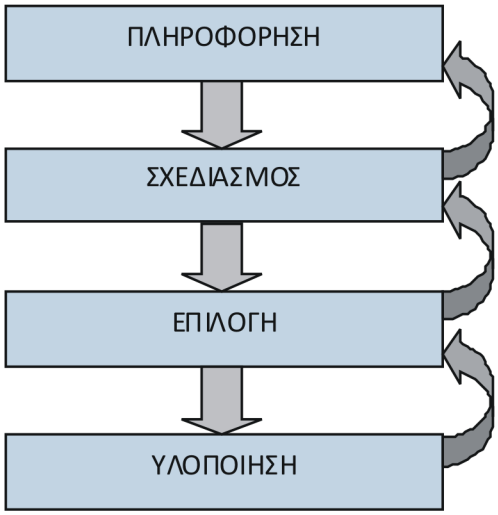
Η λήψη αποφάσεων είναι μια καθημερινή ανθρώπινη δραστηριότητα. Σε επίπεδο προσωπικών αποφάσεων, σημαντικό ρόλο παίζουν διάφοροι εξωλογικοί παράγοντες όπως συναισθηματικοί, ψυχολογικοί κλπ. Σε επίπεδο όμως αποφάσεων που αφορούν οργανισμούς, όσο κι αν άλλοι παράγοντες δεν μπορούν να αποκλειστούν τελείως, καθοριστικό ρόλο παίζει η ανθρώπινη λογική. Οι αποφάσεις αυτές χαρακτηρίζονται Λογικές Αποφάσεις (Rational Decisions). Για τη λήψη Λογικών Αποφάσεων προσμετρούνται τα οφέλη που θα προκύψουν από την απόφαση, όπως επίσης και οι απαραίτητοι πόροι που πρέπει να διατεθούν, εφαρμόζονται κριτήρια αξιολόγησης της επιτυχίας και απαιτούνται πληροφορίες. Στην περίπτωση των επιχειρηματικών αποφάσεων, στο επίκεντρο τοποθετούνται τα οικονομικά ζητήματα. Ωστόσο, για τη λήψη επιχειρηματικών αποφάσεων τα κριτήρια δεν είναι μόνον οικονομικά. Η Sauter (1997) ορίζει έξι τύπους λογικής που σχετίζονται με τη λογική λήψη αποφάσεων:

- Οικονομική. Αναφέρεται στο οικονομικό όφελος που θα αποφέρει η απόφαση και στο κόστος για την υλοποίηση της. Προφανώς το ζητούμενο είναι η μεγιστοποίηση του οφέλους και η ελαχιστοποίηση του κόστους.
- Τεχνική. Μελετά την αποτελεσματικότητα των προτεινόμενων λύσεων. Λύσεις οι οποίες δεν θα επιφέρουν την επίτευξη των στόχων αποκλείονται. Τεχνικά θέματα που σχετίζονται με τις εναλλακτικές λύσεις μελετώνται στα πλαίσια αυτής της λογικής.
- Νομική. Ασχολείται με το κατά πόσο οι λύσεις δεν παραβιάζουν νομικές διατάξεις και κατ' επέκταση η εφαρμογή τους θα επισύρει κυρώσεις. Για παράδειγμα, η εγκατάσταση και λειτουργία ενός εργοστασίου σε μια χώρα πρέπει να είναι σύμφωνη με την περιβαλλοντολογική της νομοθεσία.
- Κοινωνική. Αφορά την ηθική διάσταση των πιθανών λύσεων, όπως αυτή γίνεται κατανοητή με βάση τα τρέχοντα κοινωνικά θέσφατα. Ενέργειες που θίγουν το κοινό περί δικαίου αίσθημα ή προσβάλλουν πάγιες ηθικές αξίες, θα καταστήσουν τον οργανισμό στόχο αρνητικής κριτικής ή και κακόβουλων ενεργειών.
- Διαδικαστική. Μελετά εάν οι προτεινόμενες λύσεις είναι σε συμφωνία με τις διαδικασίες και τις υποδομές του οργανισμού ή του περιβάλλοντος. Σε μια εταιρεία θαλάσσιων μεταφορών, λύσεις που απαιτούν εκτεταμένη χρήση αεροπορικών μέσων είναι πιθανότατα μη εφαρμόσιμες.
- Πολιτική. Οι αποφάσεις εφαρμόζονται σε ένα πραγματικό κόσμο που αποτελείται από αλληλοσχετιζόμενες οντότητες (ανθρώπους, οργανισμούς, κράτη κλπ.). Οι ενέργειες έχουν επιπτώσεις στις οντότητες ή στις σχέσεις τους. Η πολιτική λογική εξετάζει αυτές τις επιπτώσεις. Οι πιθανές αντιδράσεις συνεργατών ή ανταγωνιστών της επιχείρησης σε μια ενέργεια της είναι ένα παράδειγμα ζητήματος πολιτικής.

Όπως φαίνεται από τα παραπάνω, η λήψη μιας απόφασης είναι μια σύνθετη διαδικασία, η οποία πρέπει να συνεκτιμήσει πολλούς και διαφορετικούς παράγοντες. Στα πλαίσια μιας ιδανικής διαδικασίας θα έπρεπε να συγκεντρωθούν πληροφορίες για όλους αυτούς τους παράγοντες, να γίνει κατανοητή η βαρύτητα και η επιρροή του κάθε παράγοντα, να γίνει εξαντλητική απαρίθμηση και σχολαστική μελέτη όλων των πιθανών λύσεων και να εκτιμηθούν τα κέρδη και το κόστος για κάθε μια από αυτές. Μια τέτοια ιδανική διαδικασία θα απέδιδε τη βέλτιστη λύση. Στο παρελθόν, η προβληματική περί λήψης αποφάσεων περιστρέφονταν γύρω από την εύρεση της βέλτιστης λύσης. Ωστόσο, με πρακτικούς όρους μια τόσο εξαντλητική διαδικασία θα ήταν ανέφικτη. Στον πραγματικό κόσμο υφίστανται περιορισμοί. Όπως κατέδειξε ο Simon (1957), οι ανθρώπινες αποφάσεις είναι μερικώς λογικές εξαιτίας υπαρκτών περιορισμών. Ένα πρώτο είδος περιορισμού είναι η έκταση της πληροφόρησης που είναι διαθέσιμη. Κατά κανόνα η πληροφόρηση δεν είναι πλήρης, αλλά μερική και συνήθως η αναζήτηση κάθε δυνατής πληροφορίας για ένα ζήτημα θα ήταν αδύνατη. Επιπλέον, οι μέθοδοι επεξεργασίας των δεδομένων είναι ατελείς. Ένα δεύτερο είδος περιορισμού αφορά τις γνωστικές και αντιληπτικές ικανότητες των ανθρώπων. Οι άνθρωποι δεν είναι τέλεια όντα και οι αντιληπτικές τους ικανότητες έχουν όρια. Αργότερα, άλλοι ερευνητές επεσήμαναν ότι οι αποφάσεις πρέπει να ληφθούν εντός συγκεκριμένων χρονικών ορίων, οπότε υφίστανται και χρονικοί περιορισμοί. Γι' αυτούς τους λόγους, τα στελέχη που παίρνουν αποφάσεις δεν αναζητούν την καλύτερη δυνατή λύση αλλά αναζητούν μια «αρκετά καλή» λύση. Μια λογική αυτού του τύπου χαρακτηρίστηκε από τον Simon ως «Οριοθετημένη Λογική» (Bounded Rationality).

2.1.2 Φάσεις στη Λήψη Αποφάσεων

Η διαδικασία λήψης αποφάσεων έχει καταστεί αντικείμενο μελέτης από πλήθος ερευνητών και έχει εξεταστεί από διάφορες οπτικές γωνίες. Η σημαντικότερη ίσως συνδρομή στη μελέτη αυτή αποδίδεται στον ψυχολόγο, οικονομολόγο και πολιτικό επιστήμονα Herbert Simon, καθώς επηρέασε καθοριστικά τις πρακτικές διοίκησης των επιχειρήσεων. Ο Simon θεώρησε τη λήψη αποφάσεων ως μια διαδικασία επιλογής μεταξύ εναλλακτικών λύσεων, η οποία υπόκειται σε γνωστικούς, πληροφοριακούς και άλλους περιορισμούς. Επίσης, όρισε τη λήψη αποφάσεων ως μια συστηματική διαδικασία που αποτελείται από τρία στάδια (Simon, 1977). Αργότερα προστέθηκε ένα επιπλέον στάδιο. Τα τέσσερα στάδια του μοντέλου του Simon παρουσιάζονται στο Σχήμα 2.1.



Σχήμα 2.1 *Στάδια λήψης αποφάσεων κατά Simon*

Σύμφωνα με το μοντέλο του Simon, σε κάθε ένα από τα τέσσερα στάδια αντιστοιχούν ορισμένες εργασίες. Επίσης, υπάρχει ανάδραση μεταξύ των σταδίων και ευρήματα ενός σταδίου μπορεί να ανατροφοδοτήσουν ένα προηγούμενο στάδιο. Οι επιμέρους εργασίες που λαμβάνουν χώρα σε κάθε στάδιο έχουν ως εξής:

Πληροφόρηση. Βασικό καθήκον είναι ο καθορισμός του προβλήματος. Σε πρώτο στάδιο γίνεται συλλογή όλων των πληροφοριών που απαιτούνται. Η πληροφόρηση αφορά αρχικά τα συμπτώματα του προβλήματος. Ακολούθως, διερευνάται το κατά πόσον τα συμπτώματα είναι εκφάνσεις ενός άλλου, βαθύτερου προβλήματος. Αν για παράδειγμα διαπιστωθεί καθυστέρηση στη διανομή των προϊόντων, πρέπει να ελεγχθεί αν αυτό οφείλεται σε υποστελέχωση σε προσωπικό, ελλείψεις στα μέσα μεταφοράς, κακή οργάνωση της αποθήκης ή υπεραισιόδοξες εκτιμήσεις κατά τον σχεδιασμό του πλάνου διανομής.

Σχεδιασμός. Στο στάδιο του σχεδιασμού ορίζονται μια σειρά από εναλλακτικές λύσεις, ενέργειες δηλαδή που θα επιφέρουν τη λύση του προβλήματος. Επίσης, ορίζονται τα κριτήρια με βάση τα οποία θα γίνει η αξιολόγηση των λύσεων. Ιδιαίτερα χρήσιμος σε αυτό το στάδιο μπορεί να είναι ο σχεδιασμός ενός μοντέλου, που θα καταγράφει και θα αναπαριστά τους σημαντικούς παράγοντες του προβλήματος, καθώς και τις σχέσεις αλληλεξάρτησης τους.

Επιλογή. Στο στάδιο αυτό γίνεται συστηματική μελέτη των εναλλακτικών λύσεων που προτάθηκαν. Οι λύσεις αποτιμώνται στη βάση των κριτηρίων που ορίστηκαν στο προηγούμενο στάδιο. Εκτιμάται η αποτελεσματικότητα και αποδοτικότητα της κάθε λύσης, καθώς και το κόστος της. Το μοντέλο μπορεί να χρησιμοποιηθεί για να διερευνηθούν τα αποτελέσματα των διάφορων λύσεων. Με την ολοκλήρωση του σταδίου έχει επιλεγεί η πλέον συμφέρουσα λύση.

Υλοποίηση. Στο τελευταίο στάδιο γίνεται η εφαρμογή της απόφασης. Γίνεται κατανομή των αρμοδιοτήτων και των πόρων και υλοποιείται το σχέδιο εφαρμογής. Σε ακόλουθο χρόνο γίνεται η αξιολόγηση των αποτελεσμάτων και επιπτώσεων.

2.1.3 Είδη Αποφάσεων

Η συστηματική μελέτη της λήψης λογικών αποφάσεων έχει καταδείξει ότι η λήψη αποφάσεων είναι μια δομημένη διαδικασία, που συνεκτιμά πολλούς και διαφορετικούς παράγοντες, που υφίσταται περιορισμούς και που αρθρώνεται σε διακριτά στάδια. Πέρα όμως από τα κοινά χαρακτηριστικά τους, οι αποφάσεις διαφέρουν μεταξύ τους σε σημαντικό βαθμό και με ποικίλους τρόπους. Ο επιμερισμός των αποφάσεων σε κατηγορίες συμβάλλει στη βαθύτερη κατανόηση του προβλήματος λήψης αποφάσεων και στην καλύτερη αντιμετώπιση του. Οι αποφάσεις μπορεί να κατηγοριοποιηθούν σύμφωνα με δύο κριτήρια, τη δομή του προβλήματος και το διοικητικό επίπεδο λήψης της απόφασης.

Ως προς τη δομή του προβλήματος οι αποφάσεις χωρίζονται σε τρεις κατηγορίες δηλαδή σε δομημένες, αδόμητες και ημιδομημένες (Gorry & Scott Morton, 1971).

Δομημένες Αποφάσεις. Πρόκειται για συνήθεις αποφάσεις ρουτίνας που επαναλαμβάνονται. Το πρόβλημα που αφορούν είναι απολύτως κατανοητό και οι σχετικές λύσεις προκαθορισμένες. Οι αποφάσεις λαμβάνονται με τυποποιημένες διαδικασίες και οι λύσεις μπορεί να προέρχονται από κάποιο μαθηματικό μοντέλο ή κάποιον αλγόριθμο της διοικητικής επιστήμης ή της επιχειρησιακής έρευνας. Θα μπορούσαν να θεωρηθούν περισσότερο ως δομημένες διαδικασίες. Το γεγονός ότι η διαδικασία λήψης απόφασης είναι απολύτως σαφής και καθορισμένη, επιτρέπει την αυτοματοποιημένη λήψη της απόφασης από κάποιο κατάλληλο λογισμικό. Η εκτέλεση μια παραγγελίας όταν ικανοποιούνται οι προϋποθέσεις είναι ένα παράδειγμα δομημένης απόφασης.

Αδόμητες Αποφάσεις. Βρίσκονται στον αντίποδα των δομημένων αποφάσεων. Αφορούν καταστάσεις και προβλήματα που δεν είναι επαναλαμβανόμενα, αλλά είναι πρωτότυπα και σημαντικά. Το πρόβλημα δεν μπορεί να περιγραφεί με απόλυτη ακρίβεια και υπάρχει σημαντικός βαθμός αβεβαιότητας. Οι πιθανές λύσεις δεν είναι προκαθορισμένες και χαρακτηρίζονται και αυτές από σημαντική αβεβαιότητα. Επίσης η διαδικασία λήψης της απόφασης δεν μπορεί να καθοριστεί ως ακολουθία συγκεκριμένων βημάτων. Για τη λύση του προβλήματος πρέπει να χρησιμοποιηθεί σε μεγάλο βαθμό η ανθρώπινη κρίση και διαίσθηση. Παράδειγμα αδόμητης απόφασης είναι η ανάπτυξη μιας νέας τεχνολογίας από την επιχείρηση.

Ημιδομημένες Αποφάσεις. Τοποθετούνται μεταξύ δομημένων και αδόμητων αποφάσεων. Πλευρές του προβλήματος χαρακτηρίζονται από κάποιον βαθμό αβεβαιότητας. Για τη λύση του προβλήματος εφαρμόζονται τυποποιημένες διαδικασίες, σε συνδυασμό με την ανθρώπινη κρίση. Η εκτίμηση της πιστοληπτικής ικανότητας μπορεί να θεωρηθεί παράδειγμα ημιδομημένης απόφασης. Οι ημιδομημένες αποφάσεις είναι ο συνηθέστερος τύπος αποφάσεων.

Πρέπει να επισημανθεί ότι ο παραπάνω διαχωρισμός δεν είναι απόλυτος και άκαμπτος. Ένα συγκεκριμένο στέλεχος πιθανώς να θεωρήσει ότι ένα πρόβλημα μπορεί να αποτυπωθεί σε ένα μαθηματικό μοντέλο. Στην περίπτωση αυτή το πρόβλημα αντιμετωπίζεται ως δομημένο. Αντιθέτως, ένα άλλο στέλεχος πιθανώς να θεωρήσει ότι το συγκεκριμένο πρόβλημα είναι περισσότερο περίπλοκο και δεν μπορεί να μοντελοποιηθεί. Το πρόβλημα τότε αντιμετωπίζεται ως αδόμητο ή ημιδομημένο. Μια άλλη παράμετρος είναι η δυνατότητα πρόσβασης στην πληροφορία και οι διαθέσιμες μέθοδοι επεξεργασίας της. Πληροφοριακά συστήματα, που παρέχουν πρόσβαση σε δεδομένα υψηλής ποιότητας και προσφέρουν δυνατότητες προωθημένης επεξεργασίας τους, μειώνουν τον βαθμό αβεβαιότητας και διευκολύνουν την τυποποίηση διαδικασιών. Οργανισμοί που διαθέτουν τέτοια συστήματα βρίσκονται σε πλεονεκτική θέση έναντι ανταγωνιστών τους που τα στερούνται.

Το δεύτερο κριτήριο κατηγοριοποίησης των αποφάσεων αφορά το διοικητικό επίπεδο λήψης τους. Σύμφωνα με αυτό το κριτήριο οι αποφάσεις χωρίζονται σε λειτουργικές, τακτικές και στρατηγικές.

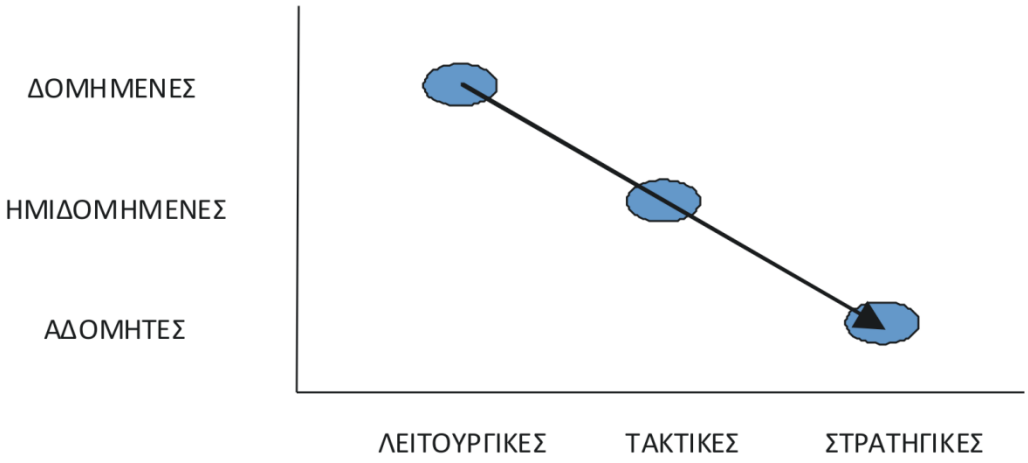
Λειτουργικές Αποφάσεις. Αφορούν ζητήματα άμεσης λειτουργίας και συγκεκριμένων εργασιών. Έχουν βραχυπρόθεσμο έως άμεσο χρονικό ορίζοντα και λαμβάνονται από χαμηλόβαθμα στελέχη που είναι επιφορτισμένα με τη λειτουργία ενός υποτμήματος ή με την εκτέλεση μιας εργασίας.

Τακτικές Αποφάσεις. Αφορούν τακτικές για την υλοποίηση των στρατηγικών στόχων. Μπορούν να σχετίζονται με την αποτελεσματικότητα χρήσης πόρων ή την αποδοτικότητα λειτουργικών μονάδων. Συνήθως επηρεάζουν ένα τμήμα του οργανισμού (πχ το τμήμα πωλήσεων) και έχουν βραχυπρόθεσμο ή μεσοπρόθεσμο ορίζοντα. Τακτικές αποφάσεις λαμβάνονται από τα μεσαία στελέχη (πχ διευθυντής εργοστασίου).

Στρατηγικές Αποφάσεις. Αφορούν τον καθορισμό των στόχων, των πόρων και της πολιτικής της επιχείρησης καθώς και τον έλεγχο για την εκπλήρωση των στόχων. Έχουν από μεσοπρόθεσμο έως μακροπρόθεσμο χρονικό ορίζοντα. Η σημασία τους είναι βαρύνουσα και μπορούν να επηρεάσουν ολόκληρο τον οργανισμό ή ένα σημαντικό τμήμα του. Στρατηγικές αποφάσεις λαμβάνονται από τα κορυφαία διοικητικά στελέχη.

Ο βαθμός δόμησης ενός προβλήματος και το επίπεδο λήψης της απόφασης σχετίζονται μεταξύ τους, αν και όχι με απόλυτο τρόπο. Συνήθως ο βαθμός δόμησης των αποφάσεων μειώνεται καθώς μεταβαίνουμε από το λειτουργικό στο στρατηγικό επίπεδο. Στο λειτουργικό επίπεδο, οι καλά καθορισμένες διαδικασίες καθημερινής λειτουργίας και η ακριβής πληροφόρηση επιτρέπουν τη λήψη δομημένων αποφάσεων. Αντιθέτως, στο

στρατηγικό επίπεδο, η περιπλοκότητα των συνθηκών και το πλήθος των πιθανών λύσεων επιβάλλει τη λήψη ως επί το πλείστον αδόμητων αποφάσεων. Οι στρατηγικές αποφάσεις λαμβάνονται κυρίως σε συνθήκες ρίσκου ή και αβεβαιότητας. Συνθήκες ρίσκου υπάρχουν όταν οι συνθήκες και τα αποτελέσματα είναι πιθανολογικά ενδεχόμενα Μια σειρά ενεργειών μπορεί να επιφέρει διάφορα αποτελέσματα και η πιθανότητα εμφάνισης ενός αποτελέσματος μπορεί να υπολογιστεί. Αβεβαιότητα υπάρχει όταν οι συνθήκες είναι απρόβλεπτες και όταν μια σειρά ενεργειών μπορεί να επιφέρει διαφορετικά αποτελέσματα με άγνωστη πιθανότητα εμφάνισης. Η σχέση επιπέδου λήψης απόφασης και δόμησης των προβλημάτων αναπαρίσταται γραφικά στο Σχήμα 2.2. Θα έπρεπε ωστόσο να τονιστεί ότι ο συσχετισμός δόμησης του προβλήματος και επιπέδου λήψης απόφασης δεν είναι απόλυτος. Σε όλα τα επίπεδα υπάρχουν και δομημένες και αδόμητες αποφάσεις. Η τεχνολογία της πληροφορικής εδώ και πολλά χρόνια έχει προσφέρει σημαντικά εργαλεία για τη λήψη δομημένων και ημιδομημένων αποφάσεων. Η πρόκληση της σημερινής εποχής είναι η χρήση της πληροφορικής για τη λήψη αδόμητων αποφάσεων.



Σχήμα 2.2 *Επίπεδο λήψης αποφάσεων και βαθμός δόμησης.*

2.1.4 Διοικητικά Στελέχη και Λήψη Αποφάσεων

Σε έναν οργανισμό οι αποφάσεις λαμβάνονται από τα διοικητικά του στελέχη. Η επιστήμη της διοίκησης επιχειρήσεων έχει ορίσει με σαφήνεια τα διοικητικά καθήκοντα και έχει αναδείξει τη λήψη αποφάσεων ως μια από τις βασικές συνισταμένες της άσκησης διοίκησης. Σύμφωνα με τον Mintzberg (1990) η διοίκηση ενός οργανισμού επιτελεί δέκα βασικούς ρόλους, που εντάσσονται σε τρεις κατηγορίες, δηλαδή την κατηγορία των διαπροσωπικών ρόλων, των ρόλων πληροφόρησης και την κατηγορία των ρόλων απόφασης:

- **Διαπροσωπικοί ρόλοι**
  - Εκπροσώπηση του οργανισμού και εκτέλεση συμβολικών καθηκόντων νομικής ή οικονομικής φύσης.
  - Ηγεσία του οργανισμού που παρακινεί, καθοδηγεί και διοικεί του υφισταμένους.
  - Σύνδεση του οργανισμού με το εξωτερικό περιβάλλον.
- **Πληροφοριακοί ρόλοι**
  - Αναζήτηση εσωτερικής και εξωτερικής πληροφόρησης σχετικά με τον οργανισμό.
  - Διάχυση της πληροφορίας στον οργανισμό.
  - Παροχή πληροφόρησης σε τρίτους σχετικά με τον οργανισμό.
- **Ρόλοι λήψης απόφασης**

- Άσκηση επιχειρηματικής δραστηριότητας και προώθηση της αλλαγής και της καινοτομίας.
- Αντιμετώπιση των κλυδωνισμών και των δυσκολιών.
- Διαχείριση και διάθεση των πόρων του οργανισμού όπως τον χρόνο, τα κεφάλαια, τον εξοπλισμό και το ανθρώπινο δυναμικό.
- Διεξαγωγή διαπραγματεύσεων για τον οργανισμό.

Ο Mintzberg (1975) μελέτησε τα χαρακτηριστικά και τον τρόπο εργασίας των στελεχών που λαμβάνουν τις αποφάσεις. Σύμφωνα με την Sauter (1997), που συνοψίζει τα ευρήματα του Mintzberg, τα στελέχη επιθυμούν να λειτουργούν με τον δικό τους προσωπικό τρόπο, ο οποίος κατά τη γνώμη τους έχει αποδειχθεί αποτελεσματικός. Προτιμούν έναν άμεσο και ανεπίσημο τρόπο πρόσβασης στην πληροφορία από έναν τυπικό τρόπο, όπου θα ζητούσαν από κάποιον υφιστάμενο τους να συντάξει μια επίσημη μελέτη. Ο τρόπος σκέψης τους δεν είναι γραμμικός και συχνά η αναζήτηση τους εκτρέπεται σε νέα θέματα τα οποία ανακύπτουν στην πορεία. Είναι σημαντικό γι' αυτούς να γνωρίζουν την πηγή των πληροφοριών. Θα αποδεχθούν ευκολότερα μια πληροφορία αν γνωρίζουν ότι η πηγή της είναι αξιόπιστη. Για την εκτέλεση της εργασίας τους χρειάζονται πρόσθετες πηγές, που να τους βοηθούν στην κατανόηση της πληροφορίας. Τέλος, εκτιμούν τη συμμετοχή και επιθυμούν την εμπλοκή πολλών μερών στη διαδικασία λήψης αποφάσεων. Τα παραπάνω χαρακτηριστικά μπορούν να αποτελέσουν χρήσιμες υποδείξεις για τον σχεδιασμό ενός Συστήματος Υποστήριξης Αποφάσεων.

2.1.5 Λήψη αποφάσεων και πληροφοριακά συστήματα.

Τα παλαιότερα χρόνια, η λήψη αποφάσεων θεωρούνταν περισσότερο ως μια τέχνη, ως ένα σύνολο προσωπικών ικανοτήτων, που αναπτύχθηκαν μέσω της εμπειρίας με την πάροδο του χρόνου. Στη σημερινή εποχή, η προσέγγιση αυτή δεν επαρκεί. Ο όγκος της παρεχόμενης πληροφόρησης είναι τόσο μεγάλος, που η τήρηση του χωρίς τη χρήση εξειδικευμένων εργαλείων βρίσκεται έξω από τις ανθρώπινες δυνατότητες. Το ίδιο συμβαίνει και με την ανάγκη επεξεργασίας όλων αυτών των δεδομένων. Οι σύγχρονοι μάνατζερς, πέρα από τις ιδιαίτερες προσωπικές τους ικανότητες, πρέπει να είναι συστηματικοί στην εργασία τους και να αξιοποιούν τα νέα εργαλεία που τους προσφέρονται. Η τεχνολογία της πληροφορικής έχει αλλάξει το τοπίο και στο πεδίο λήψης αποφάσεων.

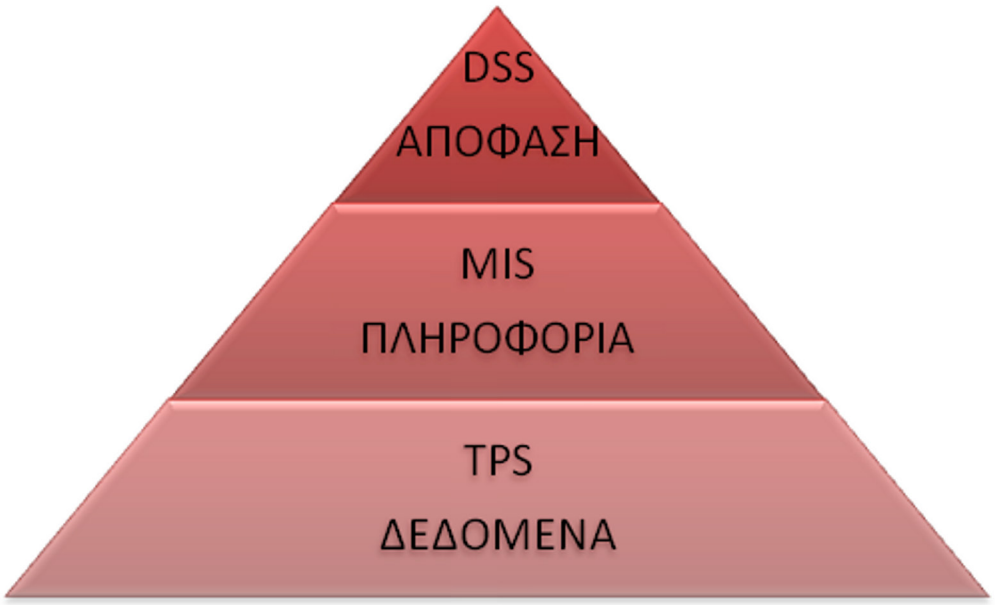
Στη σημερινή εποχή, η εφαρμογή της πληροφορικής στις επιχειρήσεις είναι πλήρης. Μια σειρά από πληροφοριακά συστήματα είναι εγκατεστημένα και λειτουργούν παρέχοντας δυνατότητες τήρησης και επεξεργασίας δεδομένων καθώς και επικοινωνίας. Τέτοια συστήματα είναι τα παρακάτω:

- Πληροφοριακά Συστήματα Αυτοματισμού Γραφείου (Office Automation Systems).
- Πληροφοριακά Συστήματα Παρακολούθησης Συναλλαγών (Transaction Processing Systems).
- Πληροφοριακά Συστήματα Διοίκησης (Management Information Systems).
- Συστήματα Υποστήριξης Αποφάσεων (Decision Support Systems).
- Συστήματα Υποστήριξης Ανώτατων Στελεχών (Executive Information Systems).
- Συστήματα Διαχείρισης Γνώσης (Knowledge Management Systems).

Τα πληροφοριακά συστήματα παρακολούθησης συναλλαγών είναι επιφορτισμένα με την παρακολούθηση των συναλλαγών που πραγματοποιούνται στα πλαίσια της λειτουργίας της επιχείρησης. Στην κατηγορία αυτή ανήκουν κυρίως τα συστήματα Σχεδιασμού Επιχειρησιακών Πόρων (Enterprise Resources Planning – ERP), αλλά και άλλα συστήματα, όπως Διαχείρισης Πελατειακών Σχέσεων (Customer Relationship Management (CRM)) και Διαχείρισης Εφοδιαστικής Αλυσίδας (Supply Chain Management (SCM)). Τα Πληροφοριακά Συστήματα Διοίκησης προσφέρουν πληροφόρηση για τον οργανισμό, τα Συστήματα Υποστήριξης Αποφάσεων, όπως και τα Συστήματα Υποστήριξης Στελεχών, διευκολύνουν τη λήψη αποφάσεων και τα Συστήματα Διαχείρισης Γνώσης επιτρέπουν την τήρηση, οργάνωση και επικοινωνία της συλλογικής γνώσης ενός οργανισμού. Όλα αυτά τα συστήματα, σε μικρότερο ή μεγαλύτερο βαθμό και με διαφορετικό τρόπο, μπορούν να συμβάλλουν στη διαδικασία λήψης αποφάσεων. Ωστόσο, ορισμένα από αυτά είναι ειδικά σχεδιασμένα γι' αυτόν τον σκοπό.

Τα Συστήματα Παρακολούθησης Συναλλαγών, τα οποία έχουν καταγεγραμμένες όλες τις συναλλαγές της επιχείρησης, αποτελούν την κύρια πηγή δεδομένων. Όποιος αναζητά πληροφόρηση με τη μέγιστη δυνατή λεπτομέρεια θα πρέπει να ανατρέξει σε αυτά. Στη σημερινή εποχή όμως, υπάρχουν πρόσθετες πηγές βασικών δεδομένων, και μάλιστα η σημασία τους αυξάνεται με γρήγορο ρυθμό. Τέτοιες πηγές σχετίζονται με το διαδίκτυο και είναι οι διαδικτυακοί σέρβερς της επιχείρησης αλλά και εξωτερικές πηγές τρίτων παρόχων, καθώς και το ραγδαία αναπτυσσόμενο Web 2.0. Τα Πληροφοριακά Συστήματα Διοίκησης είναι ικανά να αντλούν πληροφό-

ρηση κυρίως από τα Συστήματα Παρακολούθησης Συναλλαγών, και να την παρουσιάζουν στα στελέχη ώστε να τους διευκολύνουν στη λήψη αποφάσεων. Σε υψηλότερη βαθμίδα χρησιμότητας για τη λήψη αποφάσεων βρίσκονται τα εξειδικευμένα συστήματα Λήψης Αποφάσεων. Η ιεράρχηση των τριών αυτών συστημάτων παρουσιάζεται διαγραμματικά στο Σχήμα 2.3



Σχήμα 2.3 *Πυραμίδα Πληροφοριακών Συστημάτων*

Τα διοικητικά στελέχη κατά την άσκηση των καθηκόντων τους αξιοποιούν τις δυνατότητες των πληροφοριακών συστημάτων. Ο ακριβής καθορισμός της χρήσης των πληροφοριακών συστημάτων εξαρτάται από το θεωρητικό πλαίσιο που χρησιμοποιεί κανείς. Από την άποψη του ρόλου των στελεχών κατά Mintzberg, τα στελέχη χρησιμοποιούν ηλεκτρονικά συστήματα επικοινωνίας για τη διασύνδεση του οργανισμού με το εξωτερικό περιβάλλον, Πληροφοριακά Συστήματα Διοίκησης για την αναζήτηση πληροφορίας σχετικά με τον οργανισμό και τη διάχυση της σε αυτόν, και Συστήματα Υποστήριξης Αποφάσεων για την κατανομή και διάθεση πόρων του οργανισμού. Από την άποψη του μοντέλου λήψης αποφάσεων κατά Simon, τα στελέχη χρησιμοποιούν Πληροφοριακά Συστήματα Διοίκησης κατά το στάδιο της Πληροφόρησης για να αντλήσουν πληροφορία σχετικά με το πρόβλημα και Συστήματα Υποστήριξης Αποφάσεων κατά το στάδιο του Σχεδιασμού και της Επιλογής, για να πειραματιστούν με διάφορες λύσεις και να επιλέξουν κάποια από αυτές.

Οι Turban, Aronson and Liang (2005) συνοψίζουν τις δυνατότητες που προσφέρουν τα Συστήματα Υποστήριξης Αποφάσεων στη διαδικασία λήψης αποφάσεων ως εξής:

- Ταχείς υπολογισμοί. Μπορούν να εκτελεστούν περίπλοκοι υπολογισμοί με μεγάλη ταχύτητα και χαμηλό κόστος.
- Βελτιωμένη επικοινωνία. Ομάδες στελεχών που αποφασίζουν συλλογικά έχουν αυξημένες δυνατότητες επικοινωνίας.
- Αυξημένη παραγωγικότητα. Τα εξελιγμένα εργαλεία βελτιώνουν την παραγωγικότητα των αναλυτών. Επίσης διευκολύνεται η συνεργασία ατόμων που βρίσκονται σε διαφορετικές γεωγραφικές περιοχές.
- Τεχνική υποστήριξη. Η χρήση υπολογιστών επιτρέπει την αποθήκευση, επεξεργασία και μετάδοση δεδομένων με μεγάλη ταχύτητα και με οικονομικό τρόπο.
- Πρόσβαση σε Αποθήκες Δεδομένων.
- Ποιοτική Υποστήριξη. Επιτυγχάνεται με την πρόσβαση σε περισσότερα δεδομένα, δοκιμή περισσότερων εναλλακτικών, χρήση προσομοίωσης και τεχνητής νοημοσύνης κλπ.
- Ανταγωνιστικό πλεονέκτημα. Με τη βελτίωση των αποφάσεων επιτυγχάνεται βελτίωση της ποιότητας, των χρονοδιαγραμμάτων, της υποστήριξης πελατών κλπ.
- Υπέρβαση των ανθρώπινων αντιληπτικών ορίων.

2.2 Συστήματα Υποστήριξης Αποφάσεων

2.2.1 Ορισμός.

Τα διοικητικά στελέχη των επιχειρήσεων χρησιμοποιούν εξειδικευμένα συστήματα, τα Συστήματα Υποστήριξης Αποφάσεων (ΣΥΑ), για να υποβοηθηθούν στη διαδικασία λήψης αποφάσεων. Κατά καιρούς έχουν δοθεί διάφοροι ορισμοί για τα ΣΥΑ. Ο Scott Morton, πρωτοπόρος των ΣΥΑ, σε συνεργασία με τον Gorry (Gorry & Scott Morton, 1971) ορίζουν τα ΣΥΑ ως αλληλεπιδραστικά συστήματα, βασισμένα στους υπολογιστές, που βοηθούν τους αποφασίζοντες να χρησιμοποιούν δεδομένα και μοντέλα για να επιλύουν ημιδομημένα προβλήματα. Αργότερα, οι Keen and Scott-Morton (1978), με έναν πιο περιγραφικό ορισμό, αναφέρουν ότι τα ΣΥΑ συνδυάζουν τους διανοητικούς πόρους ατόμων με τις δυνατότητες των υπολογιστών, για να βελτιώσουν την ποιότητα των αποφάσεων, και ορίζουν ότι πρόκειται για συστήματα, που βασίζονται στους υπολογιστές και υποστηρίζουν διοικητικά στελέχη, τα οποία λαμβάνουν αποφάσεις για ημιδομημένα προβλήματα. Ωστόσο, οι Turban et al. (2005) επισημαίνουν ότι τέτοιοι ορισμοί, όπως και αντίστοιχοι που αναφέρονται στα Πληροφοριακά Συστήματα Διοίκησης, σημαίνουν διαφορετικά πράγματα σε διαφορετικούς ανθρώπους, και ότι δεν υπάρχει ένας γενικά αποδεκτός ορισμός για τα ΣΥΑ. Τονίζουν επίσης ότι ο όρος ΣΥΑ είναι ένας όρος-ομπρέλα, που καλύπτει κάθε σύστημα που βασίζεται σε υπολογιστές και το οποίο υποστηρίζει τη λήψη αποφάσεων σε έναν οργανισμό. Με βάση αυτόν τον ορισμό, μια Αποθήκη Δεδομένων, την οποία συμβουλεύονται στελέχη για να πάρουν αποφάσεις, μπορεί να θεωρηθεί ΣΥΑ. Επίσης, λογισμικό τεχνητής νοημοσύνης, που χρησιμοποιείται από ορκωτούς ελεγκτές για να αναπτύξουν προσδοκίες σχετικά με τις τιμές ορισμένων λογαριασμών, μπορεί να θεωρηθεί και αυτό ΣΥΑ.

Άλλοι ορισμοί των ΣΥΑ επικεντρώνουν σε άλλα ζητήματα, όπως πχ τα συστατικά τους μέρη ή τη διαδικασία ανάπτυξης τους. Οι Bonczek, Holsapple and Whinston (1980) ορίζουν τα ΣΥΑ ως συστήματα βασισμένα σε υπολογιστές, τα οποία αποτελούνται από τρία συστατικά μέρη που αλληλεπιδρούν. Τα μέρη αυτά είναι ένα σύστημα γλώσσας, δηλαδή ένα σύστημα επικοινωνίας μεταξύ του χρήστη και των άλλων μερών του ΣΥΑ, ένα σύστημα γνώσης, δηλαδή ένα σύστημα που περιέχει πληροφορίες σχετικά με το πρόβλημα, οι οποίες μπορεί να έχουν τη μορφή δεδομένων ή διαδικασιών, και τέλος, ένα σύστημα επεξεργασίας προβλημάτων, που διαθέτει διάφορες ικανότητες χειρισμού προβλημάτων, το οποίο θα χρησιμοποιηθεί για τη λήψη αποφάσεων. Ο Keen (1980) αναφέρεται στα ΣΥΑ ως καταστάσεις, όπου ένα τελικό σύστημα μπορεί να αναπτυχθεί μέσα από μια προσαρμοστική διαδικασία μάθησης και εξέλιξης. Οι χρήστες του ΣΥΑ, ο κατασκευαστής του ΣΥΑ και το ίδιο το ΣΥΑ αλληλεπιδρούν μεταξύ τους και συμβάλλουν από κοινού στην εξέλιξη του συστήματος. Το πλήθος των διαφορετικών ορισμών επιβεβαιώνει τη ρήση του Turban περί έλλειψης ενός γενικώς αποδεκτού ορισμού για τα ΣΥΑ. Σημαντικό για τον αναγνώστη είναι να κατανοήσει ότι τα ΣΥΑ εμφανίστηκαν τη δεκαετία του ‘70 ως συστήματα που χρησιμοποιούσαν δεδομένα και μαθηματικά μοντέλα, και που στόχο είχαν την υποβοήθηση ανθρώπων στη λήψη αποφάσεων. Με την πάροδο του χρόνου, αναπτύχθηκαν άλλα συστήματα με ιδιαίτερα χαρακτηριστικά, όπως οι Αποθήκες Δεδομένων και νέοι κλάδοι της Πληροφορικής, όπως η Εξόρυξη Δεδομένων, οι οποίοι δεν αυτοπροσδιορίζονται ως ΣΥΑ, μπορούν όμως να χρησιμοποιηθούν για την υποστήριξη της λήψης αποφάσεων. Το παρόν σύγγραμμα αφιερώνει ειδικά Κεφάλαια για τις Αποθήκες Δεδομένων και για τις μεθοδολογίες Εξόρυξης Δεδομένων. Στο τρέχον Κεφάλαιο, δίνεται έμφαση στα ΣΥΑ που χρησιμοποιούν μαθηματικά μοντέλα.

2.2.2 Ειδικά χαρακτηριστικά και χρησιμότητα των ΣΥΑ

Τα ΣΥΑ είναι πληροφοριακά συστήματα ειδικά σχεδιασμένα, ώστε να παρέχουν υποστήριξη σε ανθρώπους που λαμβάνουν αποφάσεις. Για να μπορέσουν να πετύχουν αυτόν τον στόχο, τα ΣΥΑ πρέπει να διαθέτουν ειδικά χαρακτηριστικά και να περιλαμβάνουν λειτουργίες, οι οποίες τα καθιστούν χρήσιμα με συγκεκριμένους τρόπους.

Πρόσβαση σε δεδομένα. Ο χρήστης του συστήματος θα πρέπει να μπορεί να αντλεί πληροφόρηση από δεδομένα τρέχοντα αλλά και παλαιότερα. Τα δεδομένα αυτά μπορεί να προέρχονται από πολλές και διαφορετικές πηγές. Τα δεδομένα, εκτός από πληροφόρηση, αναγκαία για την κατανόηση της πραγματικής τωρινής κατάστασης, περιέχουν πρότυπα που περιγράφουν κανόνες λειτουργίας και άλλες χρήσιμες πληροφορίες. Το αρχείο με τα δεδομένα για τα δάνεια, που έχει εκχωρήσει μια τράπεζα, δίνει πληροφορίες για το ύψος των δανείων, τις εγγυήσεις που έχει λάβει και τις σχετικές επισφάλειες, αλλά ταυτόχρονα περιέχει πρότυπα σχετικά με το υπό ποιές προϋποθέσεις εγκρίνονται ή απορρίπτονται οι αιτήσεις.

Ημιδομημένα και αδόμητα προβλήματα. Η χρήση εργαλείων πληροφορικής για λύση δομημένων προβλημάτων είναι σχετικά απλή. Τα ΣΥΑ δεν περιορίζονται σε αυτό, αλλά μπορούν να χρησιμοποιηθούν για τη λύση ημιδομημένων ή και αδόμητων προβλημάτων.

Χρήση από στελέχη διαφορετικών διοικητικών επιπέδων. Επιθυμητή ιδιότητα των ΣΥΑ είναι να μπορούν να χρησιμοποιηθούν από στελέχη που βρίσκονται σε διάφορα διοικητικά επίπεδα του οργανισμού. Τα ανώτερα διοικητικά στελέχη μπορούν να παρακολουθούν τη διαδικασία λήψης αποφάσεων σε κατώτερα διοικητικά επίπεδα. Με τον τρόπο αυτό εξασφαλίζεται συνοχή στις αποφάσεις.

Χρήση από ομάδες και από άτομα. Τα ΣΥΑ υποστηρίζουν ατομικές αποφάσεις παρέχοντας διάφορα εργαλεία. Οι περισσότερες αποφάσεις όμως λαμβάνονται από ομάδες. Τα ΣΥΑ επιτρέπουν τη συνεργασία πολλών ατόμων για τη λήψη αποφάσεων.

Ενσωμάτωση στη διαδικασία λήψης αποφάσεων. Όπως ήδη αναφέρθηκε, η λήψη αποφάσεων είναι μια διαδικασία που χωρίζεται σε στάδια. Τα ΣΥΑ μπορούν να χρησιμοποιηθούν και παρέχουν υποστήριξη σε όλα τα στάδια, δηλαδή της πληροφόρησης, του σχεδιασμού, της επιλογής και της υλοποίησης. Επίσης, στόχος των ΣΥΑ είναι να ενταχθούν οργανικά στη διαδικασία λήψης αποφάσεων και να αποτελέσουν αναπόσπαστο τμήμα της.

Ευελιξία και προσαρμοστικότητα. Τα στελέχη, και ειδικά τα ανώτερα, έχουν τον δικό τους προσωπικό τρόπο λειτουργίας. Τα ΣΥΑ πρέπει να επιτρέπουν στον χρήστη να τα προσαρμόζει στον δικό του τρόπο εργασίας. Επίσης, τα ΣΥΑ πρέπει να είναι ικανά να ανταποκρίνονται στις μεταβαλλόμενες συνθήκες του πραγματικού κόσμου. Ο χρήστης θα πρέπει να μπορεί να μεταβάλλει, να διαγράφει και να προσθέτει μοντέλα και λειτουργικότητες, έτσι ώστε να προσαρμόζει το σύστημα σε νέες απαιτήσεις με εύκολο και γρήγορο τρόπο.

Διαδραστικότητα. Σημαντικό χαρακτηριστικό των ΣΥΑ είναι η διαδραστικότητα. Ο χρήστης πλοηγείται στο σύστημα, μπορεί να υποβάλλει ερωτήσεις, να επικεντρώνει σε δεδομένα, να προβάλλει τα δεδομένα σε διαφορετικό επίπεδο λεπτομέρειας, να εκτελεί διάφορες αναλύσεις, όπως αναλύσεις what-if και αναζήτησης στόχου, να χρησιμοποιεί διάφορα μοντέλα ή και να τα συνδυάζει και να επιλέγει μεταξύ διαφορετικών μεθόδων ανάλυσης. Γενικώς τα ΣΥΑ δεν παρέχουν άκαμπτη πληροφόρηση αλλά επιτρέπουν σε μεγάλο βαθμό την αλληλεπίδραση με τον χρήστη.

Μοντελοποίηση. Βασικό χαρακτηριστικό είναι η αναπαράσταση περιπτώσεων λήψης αποφάσεων με τη χρήση μοντέλων, τα οποία προέρχονται από την επιχειρησιακή έρευνα και τη στατιστική. Η ύπαρξη των μοντέλων είναι αυτή που διαφοροποιεί ένα ΣΥΑ από άλλα πληροφοριακά συστήματα. Τα ΣΥΑ διαθέτουν μια συλλογή από μοντέλα. Επιπλέον, επιτρέπουν στον χρήστη να κατασκευάσει πρόσθετα μοντέλα ή να συνδυάσει επιμέρους μοντέλα για την κατασκευή ενός πιο σύνθετου μοντέλου. Ο χρήστης κάνει εκτεταμένη χρήση μοντέλων και πειραματίζεται με διάφορα σενάρια.

Αυτοματοποίηση αποφάσεων. Η μοντελοποίηση των προβλημάτων μπορεί να καταστήσει εφικτή την αυτοματοποίηση ορισμένων αποφάσεων. Συγκεκριμένες περιπτώσεις τυποποιούνται και μεταφράζονται σε συγκεκριμένες αποφάσεις, έτσι ώστε να ανταποκρίνονται σε κανόνες του οργανισμού.

Διέγερση δημιουργικότητας. Τα ΣΥΑ πρέπει να είναι σχεδιασμένα έτσι ώστε να διεγείρουν την περιέργεια του χρήστη, να τον ενθαρρύνουν, αλλά και να τον διευκολύνουν να αναζητήσει πρόσθετη πληροφόρηση, να εξετάσει διαφορετικά σενάρια, να χρησιμοποιήσει διαφορετικά εργαλεία και μεθόδους κλπ.

Αύξηση της αποτελεσματικότητας. Ο βασικότερος στόχος των ΣΥΑ είναι να βοηθήσουν ανθρώπους να λάβουν πιο αποτελεσματικές αποφάσεις. Παρέχοντας πληροφόρηση και μέσα ανάλυσης, τροφοδοτούν πολύπλευρα τη διαδικασία λήψης απόφασης. Έχοντας στη διάθεση του όγκους πληροφορίας και ποικιλία μέσων ανάπτυξης, ο χρήστης ενθαρρύνεται να αναζητήσει καλύτερες λύσεις. Σε πολλές περιπτώσεις, αυτό μπορεί να μεταφράζεται σε περισσότερη ενασχόληση, δηλαδή σε επένδυση χρόνου. Το κόστος αυτό θεωρείται αποδεκτό. Στόχος του ΣΥΑ είναι οι «καλύτερες αποφάσεις» και όχι οι «ταχύτερες αποφάσεις».

Φιλικότητα διεπαφής. Οι χρήστες των ΣΥΑ δεν είναι ειδικοί πληροφορικής. Είναι όμως ειδικοί στα επιχειρησιακά ζητήματα. Η διεπαφή του συστήματος πρέπει να μπορεί να «μιλά στη γλώσσα τους», να είναι δηλαδή εύχρηστη και κυρίως, να επικεντρώνει στα επιχειρησιακά ζητήματα. Επίσης, τα μέσα παρουσίασης της πληροφορίας πρέπει να είναι κατανοητά και συνοπτικά. Η χρήση φυσικής γλώσσας και γραφικών βοηθούν σε μεγάλο βαθμό σε αυτήν την κατεύθυνση.

2.2.3 Σύγκριση Πληροφοριακών Συστημάτων Διοίκησης και Συστημάτων Υποστήριξης Αποφάσεων

Τα διοικητικά στελέχη των επιχειρήσεων, κατά την εκτέλεση των καθηκόντων τους, χρησιμοποιούν διάφορα πληροφοριακά συστήματα. Όμως δύο κατηγορίες συστημάτων απευθύνονται κυρίως σε αυτούς, τα Πληροφο-

ριακά Συστήματα Διοίκησης (ΠΣΔ) και τα Συστήματα Υποστήριξης Αποφάσεων (ΣΥΑ). Και οι δύο κατηγορίες συστημάτων μπορούν να βοηθήσουν στη λήψη αποφάσεων. Έχουμε ήδη αναφέρει διάφορους ορισμούς για τα ΣΥΑ. Όσον αφορά τα ΠΣΔ, έχουν προταθεί επίσης διάφοροι ορισμοί. Ένας από τους πλέον επιτυχημένους είναι αυτός των Laudon and Laudon (1998), οι οποίοι ορίζουν τα ΠΣΔ ως συστήματα υποστήριξης της διοίκησης, τα οποία παρέχουν συνήθεις περιληπτικές εκθέσεις σχετικά με την επίδοση της επιχείρησης, και που χρησιμοποιούνται για την παρακολούθηση και τον έλεγχο της επιχείρησης και την πρόβλεψη της μελλοντικής επίδοσης. Οι Watson, Carroll and Mann (1987) περιγράφουν τα ΠΣΔ ως μια μέθοδο παροχής προηγούμενων, τωρινών και προβλεπόμενων πληροφοριών, σχετικών με εσωτερικές λειτουργίες και εξωτερική ευφυΐα. Τα ΠΣΔ υποστηρίζουν τον σχεδιασμό, έλεγχο και λειτουργία ενός οργανισμού, παρέχοντας ενιαία πληροφόρηση την κατάλληλη χρονική στιγμή, ώστε να βοηθηθούν τα στελέχη που λαμβάνουν αποφάσεις. Ένα ΠΣΔ παρέχει πληροφορίες σχετικά με το τι έχει συμβεί στο παρελθόν, τι συμβαίνει στο παρόν και τι είναι πιθανόν να συμβεί στο μέλλον. Η πληροφόρηση δίνεται με τη μορφή αναφορών (reports), και χρησιμοποιείται από τα στελέχη για να πάρουν αποφάσεις και να επιλύσουν προβλήματα του οργανισμού (McLeod, 1990). Σε γενικές γραμμές τα ΠΣΔ χρησιμοποιούνται από τους μάνατζερ για να παρακολουθούν και να ελέγχουν τις δραστηριότητες και τις επιδόσεις της επιχείρησης τους. Ο μάνατζερ μπορεί να υποβάλει ερωτήματα, να αντλήσει πληροφόρηση και να συντάξει αναφορές.

Παρά τις όποιες ομοιότητες τους, οι δύο κατηγορίες συστημάτων έχουν και πολλές διαφορές. Στον Πίνακα 2.1 γίνεται μια συγκριτική παράθεση των χαρακτηριστικών και ιδιοτήτων των ΠΣΔ και των ΣΥΑ

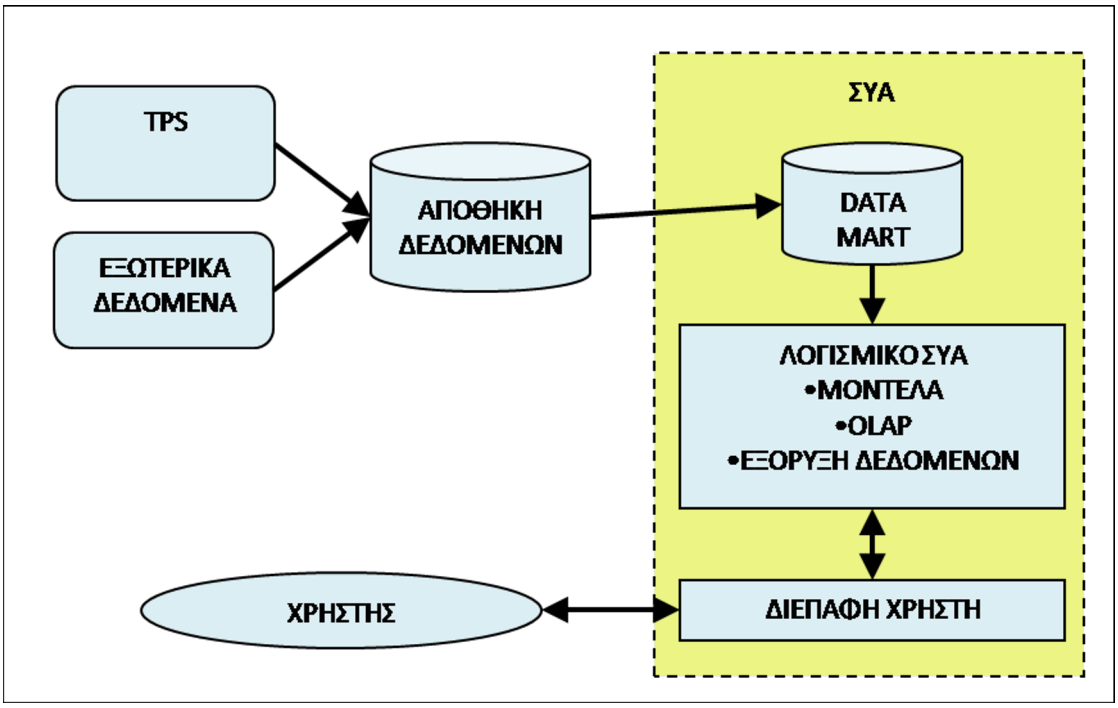
ΠΣΔ	**ΣΥΑ**
Κύριος στόχος είναι η παροχή πληροφόρησης για τον οργανισμό.	Κύριος στόχος είναι η ενεργητική υποστήριξη στη λήψη αποφάσεων.
Χρησιμοποιούνται κυρίως για τη λύση δομημένων προβλημάτων και για εργασίες ρουτίνας.	Χρησιμοποιούνται κυρίως για τη λύση ημιδομημένων ή και αδόμητων προβλημάτων.
Χρησιμοποιούνται κυρίως για λειτουργικές και τακτικές αποφάσεις.	Χρησιμοποιούνται κυρίως για στρατηγικές και τακτικές αποφάσεις.
Προσφέρουν πληροφόρηση σχετικά με τον οργανισμό.	Προσφέρουν μεθόδους μοντελοποίησης και ανάλυσης προβλημάτων.
Γίνεται κυρίως αποθήκευση, εξαγωγή, χειρισμός και παρουσίαση των δεδομένων.	Γίνεται κυρίως επεξεργασία των δεδομένων με χρήση μοντέλων και αναλυτικών μεθόδων.
Μηδενική ή σχετικά απλή χρήση αναλυτικών μεθόδων.	Εκτεταμένη χρήση εξελιγμένων αναλυτικών μεθόδων.
Χρησιμοποιούνται κυρίως στο πρώτο στάδιο της διαδικασίας λήψης αποφάσεων για την παροχή πληροφόρησης.	Χρησιμοποιούνται σε όλα τα στάδια της διαδικασίας λήψης αποφάσεων.
Συμβάλλουν στη βελτίωση της αποδοτικότητας των λειτουργιών του οργανισμού.	Συμβάλλουν στη βελτίωση της αποτελεσματικότητας των αποφάσεων.
Χρησιμοποιούν κυρίως εσωτερικά δεδομένα που προέρχονται από συστήματα επεξεργασίας συναλλαγών.	Χρησιμοποιούν και εξωτερικά δεδομένα.
Μικρότερη ευελιξία στην παραγωγή πληροφορίας.	Μεγαλύτερη ευελιξία στην παραγωγή αποτελεσμάτων ανάλυσης
Σε μεγάλο ποσοστό παράγουν περιοδικές αναφορές.	Δεν παράγουν περιοδικές αναφορές. Τα αποτελέσματα αφορούν συγκεκριμένα σενάρια και μπορεί να παραχθούν μια μόνο φορά.
Οι αναφορές είναι προκαθορισμένες και γενικές.	Τα αποτελέσματα είναι εξειδικευμένα και αφορούν συγκεκριμένο πρόβλημα.
Δεν επιτρέπουν την πρόβλεψη των αποτελεσμάτων μιας απόφασης ούτε τη σύγκριση μεταξύ εναλλακτικών λύσεων.	Επιτρέπουν την πρόβλεψη των αποτελεσμάτων μιας απόφασης ή/και τη σύγκριση μεταξύ εναλλακτικών λύσεων.

Πίνακας 2.1 *Σύγκριση ΠΣΔ - ΣΥΑ*

2.2.4 Δομή ΣΥΑ

Ένα Σύστημα Υποστήριξης Αποφάσεων μπορεί να αναλυθεί σε ένα σύνολο επιμέρους υποσυστημάτων. Τα κύρια υποσυστήματα που απαρτίζουν ένα ΣΥΑ είναι:

- **Το υποσύστημα δεδομένων.** Τα ΣΥΑ πραγματοποιούν αναλύσεις επί δεδομένων. Απαιτείται λοιπόν μια συλλογή δεδομένων και ένα σύστημα πρόσβασης σε αυτά. Πρωταρχικά, τα δεδομένα προέρχονται από τα συστήματα παρακολούθησης συναλλαγών του οργανισμού. Ωστόσο, τα ΣΥΑ μπορεί να χρησιμοποιούν και εξωτερικά δεδομένα, όπως μακροοικονομικούς δείκτες ή και δεδομένα από τους εταιρικούς διαδικτυακούς σέρβερ, ακόμα και από το Web 2.0. Τυπικά τα δεδομένα από διαφορετικές πηγές συγχωνεύονται και ομογενοποιούνται σε μια Αποθήκη Δεδομένων, που αποτελεί και τον κύριο μηχανισμό πρόσβασης σε αυτά. Για τις ειδικές ανάγκες του ΣΥΑ μπορεί να χρησιμοποιηθεί ένα υποσύνολο των δεδομένων, το οποίο αποθηκεύεται σε ένα Πρατήριο Δεδομένων (Data Mart). Το ΣΥΑ αντλεί τα απαραίτητα δεδομένα από το Πρατήριο Δεδομένων και πραγματοποιεί αναλύσεις. Η αρχιτεκτονική αυτή ωστόσο δεν είναι απόλυτη. Το ΣΥΑ μπορεί να αντλεί τα δεδομένα του απευθείας από την Αποθήκη Δεδομένων ή ακόμα και από τα συστήματα παρακολούθησης συναλλαγών.
- **Το υποσύστημα Λογισμικού Αναλύσεων.** Για τη διεξαγωγή των αναλύσεων απαιτείται εξειδικευμένο λογισμικό. Παραδοσιακά, τα ΣΥΑ χρησιμοποιούν μοντέλα, που προέρχονται από τη Στατιστική και την Επιχειρησιακή Έρευνα. Σύγχρονα ΣΥΑ χρησιμοποιούν και λογισμικό που επιτρέπει την εκτέλεση πράξεων OLAP ή και τη χρήση τεχνικών Εξόρυξης Δεδομένων και Τεχνητής Νοημοσύνης..
- **Το υποσύστημα διεπαφής χρήστη.** Τα ΣΥΑ χαρακτηρίζονται από υψηλή διαδραστικότητα. Οι χρήστες επιλέγουν δεδομένα και αναλυτικές μεθοδολογίες, πραγματοποιούν αναλύσεις και λαμβάνουν αντίστοιχες απαντήσεις. Το σύστημα διεπαφής περιλαμβάνει τη «γλώσσα» με την οποία ο χρήστης δίνει εντολές στο ΣΥΑ (action language) και τα μέσα με τα οποία το ΣΥΑ παρουσιάζει τα αποτελέσματα στον χρήστη (presentation language). Πολλοί χρήστες αντιλαμβάνονται το σύστημα διεπαφής ως το ίδιο το σύστημα. Για τον λόγο αυτό, ο σχεδιασμός και τα χαρακτηριστικά του έχουν πολύ μεγάλη σημασία. Η διεπαφή πρέπει να χαρακτηρίζεται από απλότητα και ευχρηστία, να είναι σχεδιασμένη έτσι ώστε να επικεντρώνει σε επιχειρηματικά και όχι τεχνικά ή διαδικαστικά ζητήματα, και να διεγείρει τη δημιουργικότητα του χρήστη. Επίσης, πρέπει να είναι ευέλικτη και να επιτρέπει στον χρήστη να προσαρμόσει το σύστημα στον δικό του τρόπο εργασίας, να επιλέγει δεδομένα, να εφαρμόζει διάφορες αναλυτικές μεθόδους, να συγκρίνει σενάρια, να προσθέτει, να αφαιρεί και να δημιουργεί νέα μοντέλα. Η χρήση γραφικών μέσων βοηθά πολύ στην καλύτερη παρουσίαση και κατανόηση των αποτελεσμάτων. Επίσης, ενδείκνυται η χρήση του γνώριμου περιβάλλοντος ενός Web browser.



Σχήμα 2.4 *Η δομή ενός ΣΥΑ*

2.2.5 Σύστημα Διαχείρισης Βάσης Μοντέλων

Η ύπαρξη και χρήση μοντέλων για τη διεξαγωγή αναλύσεων είναι ένα βασικό χαρακτηριστικό των ΣΥΑ. Τυπικά, ένα ΣΥΑ διαθέτει πλήθος μοντέλων, που προέρχονται από τη Στατιστική, την Επιχειρησιακή Έρευνα και τη Χρηματοοικονομική Ανάλυση. Η διαχείριση των μοντέλων γίνεται μέσω ενός εξειδικευμένου συστήματος, του Συστήματος Διαχείρισης Βάσης Μοντέλων (ΣΔΒΜ), το οποίο διευκολύνει την πρόσβαση στα μοντέλα, καθώς και τη χρήση τους. Ειδικότερα, το ΣΔΒΜ προσφέρει τις παρακάτω λειτουργίες:

- Αποθηκεύει και διατηρεί μια συλλογή (ή βάση) από μοντέλα. Σύμφωνα με τους Turban et al. (2005), τα μοντέλα χωρίζονται στις παρακάτω κατηγορίες:
  - Στρατηγικά. Χρησιμοποιούνται από κορυφαία στελέχη για ζητήματα στρατηγικού σχεδιασμού, όπως τον καθορισμό των στρατηγικών στόχων και του στρατηγικού προσανατολισμού.
  - Τακτικά. Χρησιμοποιούνται από μεσαία στελέχη για την κατανομή και τον έλεγχο των πόρων του οργανισμού.
  - Λειτουργικά. Χρησιμοποιούνται από κατώτερα στελέχη και σχετίζονται με τη μοντελοποίηση καθημερινών λειτουργιών.
  - Αναλυτικά. Χρησιμοποιούνται για την ανάλυση των δεδομένων εφαρμόζοντας μεθόδους από τη Στατιστική και την Εξόρυξη Δεδομένων.
- Διατηρεί έναν κατάλογο μοντέλων, ο οποίος μπορεί να περιλαμβάνει τους ορισμούς των μοντέλων, λεπτομέρειες για τη δομή και τους αλγορίθμους τους, οδηγίες για τη χρήση τους κλπ. Η τεκμηρίωση των μοντέλων διευκολύνει τη χρήση τους (Smith, Gunther, Rao & Ratlife, 2001).
- Επιτρέπει την τροποποίηση των υπαρχόντων μοντέλων και τη δημιουργία νέων με χρήση κάποιας γλώσσας προγραμματισμού.
- Επιτρέπει την κατασκευή νέων μοντέλων από άλλα βασικά μοντέλα. Βασικά μοντέλα και εξειδικευμένες ρουτίνες που εκτελούν μια εργασία, χρησιμοποιούνται ως δομικοί λίθοι, οι οποίοι συνδυάζονται για να συγκροτήσουν ένα νέο σύνθετο μοντέλο.
- Για την υποστήριξη της κατασκευής των νέων μοντέλων διαθέτει ένα περιβάλλον ανάπτυξης μοντέλων και μια γλώσσα ορισμού μοντέλων (Model Definition Language (MDL)), έτσι ώστε τα μοντέλα να αποθηκεύονται με κατάλληλο τρόπο στη βάση μοντέλων.
- Επιτρέπει στον χρήστη να αναζητήσει, να επιλέξει και να ανασύρει το κατάλληλο μοντέλο.
- Προσφέρει μια διεπαφή εισόδου, που επιτρέπει στον χρήστη να τροφοδοτήσει με δεδομένα εισόδου τα μοντέλα.
- Διευκολύνει τον χειρισμό των μοντέλων, μετατρέποντας τα δεδομένα του χρήστη σε μορφή κατάλληλη για τα μοντέλα, και απλοποιώντας τις εντολές χρήσης των μοντέλων.
- Προσφέρει ένα περιβάλλον για την εκτέλεση και την παρακολούθηση των λειτουργιών των μοντέλων.
- Παρουσιάζει τα αποτελέσματα σε μορφή εύκολα κατανοητή από τον χρήστη. Οθόνες αποτελεσμάτων με στατιστικές λεπτομέρειες απαιτούν αυξημένες γνώσεις στατιστικής για την κατανόηση τους. Αντιθέτως, η χρήση φυσικής γλώσσας, που αποκρύπτει μαθηματικές λεπτομέρειες και εξηγεί τα αποτελέσματα με απλό τρόπο, διευκολύνει την κατανόηση των αποτελεσμάτων και τη χρήση του συστήματος. Επίσης, το σύστημα μπορεί να υποδεικνύει στον χρήστη κατευθύνσεις για περαιτέρω ανάλυση, και γενικώς να τον ενθαρρύνει να πειραματιστεί με το μοντέλο.

2.2.6 Συστήματα Υποστήριξης Ομαδικών Αποφάσεων

Η λήψη αποφάσεων δεν είναι πάντα μια ατομική εργασία. Σε πολλές περιπτώσεις η ευθύνη για τη λήψη αποφάσεων ανήκει σε ένα συλλογικό όργανο. Το όργανο αυτό απαρτίζεται από μέλη που συμμετέχουν, επικοινωνούν, συνεργάζονται και τελικά αποφασίζουν, πιθανότατα μέσω ψηφοφορίας. Εκτός από τα μέλη που συμμετέχουν θεσμικά, μπορεί να κληθούν στην ομάδα και πρόσθετα εξωτερικά μέλη, όπως ειδικοί σε κάποιο ζήτημα, αναλυτές, συντονιστές, εκπρόσωποι συνεργαζόμενων οργανισμών κλπ. οι οποίοι κατά κανόνα παίζουν συμβουλευτικό ρόλο. Η λήψη αποφάσεων από ομάδα ατόμων έχει τις δικές της ιδιαιτερότητες και παρουσιάζει πλεονεκτήματα και μειονεκτήματα. Βασικό πλεονέκτημα είναι ότι κάθε ένας από τους συμμετέχοντες συνεισφέρει τη δική του γνώση, εμπειρία και προβληματική, εμπλουτίζοντας με τον τρόπο αυτόν την όλη διαδικασία. Θετικά αποτελέσματα μπορεί να είναι η καλύτερη κατανόηση του προβλήματος, η παραγωγή περισσότερων ιδεών και ο ευκολότερος εντοπισμός λαθών. Βεβαίως, η συλλογική λήψη αποφάσεων έχει δεχθεί και έντονη κριτική. Χαρακτηριστική είναι η ρήση του γνωστού Έλληνα μηχανικού Αλέξανδρου Ισσιγόνη,

σχεδιαστή του πολύ επιτυχημένου αυτοκινήτου Mini Cooper, ο οποίος είπε ότι «η καμήλα είναι ένα άλογο που σχεδιάστηκε από επιτροπή». Η συνύπαρξη διαφορετικών και ίσως αντικρουόμενων απόψεων, χρονικές καθυστερήσεις και παρελκυστικές τακτικές, ο ελλιπής συντονισμός και άλλα προβλήματα μπορεί να οδηγήσουν σε κακές αποφάσεις.

Για την υποστήριξη της λήψης ομαδικών αποφάσεων με χρήση τεχνολογιών της πληροφορικής, έχουν προταθεί τα λεγόμενα Συστήματα Υποστήριξης Ομαδικών Αποφάσεων (ΣΥΟΑ) (Group Decision Support Systems (GDSS)). Τα ΣΥΟΑ υποστηρίζουν τη λύση ημιδομημένων ή και αδόμητων προβλημάτων από ομάδες χρηστών. Παρέχουν όλες τις δυνατότητες των ΣΥΑ όπως έχουν αναφερθεί ανωτέρω. Οι χρήστες έχουν πρόσβαση σε δεδομένα, χειρίζονται μοντέλα και πραγματοποιούν αναλύσεις. Επιπλέον όμως, τα ΣΥΟΑ προσφέρουν πρόσθετες δυνατότητες επικοινωνίας και συνεργατικής εργασίας. Τέτοιες δυνατότητες προσφέρονται από μια κατηγορία εξειδικευμένου λογισμικού, των Συστημάτων Υποστήριξης Ομάδων (ΣΥΟ) (Group Support Systems (GSS)) ή Λογισμικού Ομάδων (ΛΟ) (Groupware).

Το Λογισμικό Ομάδων είναι εξειδικευμένο λογισμικό για τη διευκόλυνση της επικοινωνίας και της συνεργασίας ομάδων εργασίας. Οι Huber, Valacich and Jessup (1993) ορίζουν τα ΣΥΟ ως τεχνολογίες υπολογιστών, που χρησιμοποιούνται για να βοηθήσουν συλλογικές προσπάθειες, οι οποίες σκοπεύουν στον εντοπισμό και αντιμετώπιση προβλημάτων, ευκαιριών και ζητημάτων. Στόχος των ΣΥΟ είναι να ενισχύσουν τα πλεονεκτήματα που προφέρει η συνεργατική εργασία και ταυτόχρονα να περιορίσουν τα μειονεκτήματα. Δύο βασικές λειτουργίες των ΣΥΟ είναι η επικοινωνία και η συνεργασία.

Οι σύγχρονες τεχνολογίες επικοινωνίας, οι οποίες χρησιμοποιούν δίκτυα υπολογιστών, προσφέρουν πρωτόγνωρες δυνατότητες και μέσα επικοινωνίας, που πριν μερικές δεκαετίες θα ήταν αδιανόητες. Η ευρύτατη εξάπλωση του Διαδικτύου εξασφαλίζει μια παγκόσμια πλατφόρμα επικοινωνίας. Με τη βοήθεια των νέων τεχνολογιών αναπτύχθηκε ένα πλήθος νέων μέσων, που προσφέρουν εναλλακτικές λύσεις επικοινωνίας με τρόπο φθηνό, γρήγορο και αξιόπιστο. Οι χρήστες μπορούν να έχουν πρόσβαση σε πληροφορίες που βρίσκονται αποθηκευμένες σε οποιοδήποτε μέρος του κόσμου, να ανταλλάσσουν αρχεία, να συζητούν σε πραγματικό χρόνο, να επικοινωνούν ασύγχρονα με τη βοήθεια του ηλεκτρονικού ταχυδρομείου και να συμμετέχουν σε newsgroups. Η έλευση του Web 2.0 πολλαπλασίασε τους τρόπους επικοινωνίας προσθέτοντας τα blogs, τα wikis και τα μέσα κοινωνικής δικτύωσης. Οι εικονικοί χώροι και τα άβαταρ, που ήδη εφαρμόζονται, δίνουν μια γεύση από τις δυνατότητες του κοντινού μέλλοντος. Προς το παρόν, το κορυφαίο μέσο για την επικοινωνία ομάδων λήψης αποφάσεων είναι η τηλεδιάσκεψη. Οι χρήστες μπορούν να έχουν εικόνα και να συζητούν χρησιμοποιώντας δικτυωμένους υπολογιστές και κατάλληλο λογισμικό. Η τηλεδιάσκεψη και τα άλλα σύγχρονα μέσα επικοινωνίας προσφέρουν μια σειρά από πλεονεκτήματα. Ο Davis (1999) αναφέρει τα ακόλουθα:

- αύξηση της παραγωγικότητας των εργαζομένων,
- συμμετοχή περισσότερων ατόμων στη λήψη αποφάσεων,
- υπέρβαση των γεωγραφικών ορίων,
- δημιουργία μιας ενιαίας επιχειρησιακής κουλτούρας,
- βελτίωση της ποιότητας ζωής των εργαζομένων.

Πέραν της επικοινωνίας, τα ΣΥΟ δημιουργούν και ένα περιβάλλον συνεργασίας. Η συνεργασία είναι ένα θέμα πολύ πιο σύνθετο από την απλή επικοινωνία. Τα άτομα πρέπει να εργάζονται από κοινού σε ένα αντικείμενο, πχ να συντάσσουν από κοινού ένα έγγραφο ή να εργάζονται σε συμπληρωματικά αντικείμενα. Πρέπει επίσης να μοιράζονται γνώσεις και να αποφασίζουν μέσω ψηφοφορίας. Επιπλέον, χρειάζεται ένα σύστημα διαχείρισης της ροής της εργασίας.

Δύο σημαντικές παράμετροι στην ομαδική εργασία είναι ο χρόνος και ο χώρος στον οποίο λαμβάνει χώρα η συνεργασία. Υπάρχουν τέσσερεις δυνατοί συνδυασμοί:

- Ίδιος χρόνος και ίδιος χώρος. Πρόκειται για τον παραδοσιακό τρόπο συνεργασίας, όπου τα μέλη της ομάδας συγκεντρώνονται στον ίδιο χώρο και εργάζονται από κοινού.
- Ίδιος χρόνος αλλά διαφορετικός χώρος. Μοντέρνος και διαδεδομένος τρόπος συνεργασίας. Εφαρμόζεται συνήθως όταν τα άτομα βρίσκονται σε απομακρυσμένες γεωγραφικά περιοχές. Τα μέλη χρησιμοποιούν τηλεδιάσκεψη και άλλα μέσα για τη συνεργασία τους.
- Διαφορετικός χρόνος αλλά ίδιος χώρος. Σπάνιος τρόπος συνεργασίας. Τα μέλη εργάζονται στον ίδιο χώρο σε διαφορετικά χρονικά διαστήματα. Απαιτείται σύστημα διαχείρισης της ροής της εργασίας.
- Διαφορετικός χρόνος και διαφορετικός χώρος. Ιδιαίτερα διαδεδομένη τακτική που εφαρμόζεται όταν τα μέλη έχουν περιορισμένη διαθεσιμότητα χρόνου και βεβαρημένο πρόγραμμα ή όταν βρίσκονται σε γεωγραφικές περιοχές με μεγάλη διαφορά ώρας.

Στο Σχήμα 2.5 περιγράφονται οι δυνατοί συνδυασμοί τόπου και χρόνου για ομαδική εργασία.

Σχήμα 2.5 *Συνδυασμός τόπου και χρόνου για ομαδική εργασία*

Τα ΣΥΟ προσφέρουν διάφορα μέσα συνεργατικής εργασίας. Σε αυτά περιλαμβάνονται η πρόσβαση σε κοινές βάσεις δεδομένων, διαχείριση εγγράφων, ανταλλαγή ηλεκτρονικών μηνυμάτων, τηλεδιάσκεψη, διαχείριση της ροής εργασίας, μέσα ψηφοφορίας, εργαλεία χρονοπρογραμματισμού, μέσα επίλυσης διχογνωμιών, διαδικασίες κοινής ανακάλυψης νέων ιδεών κλπ.

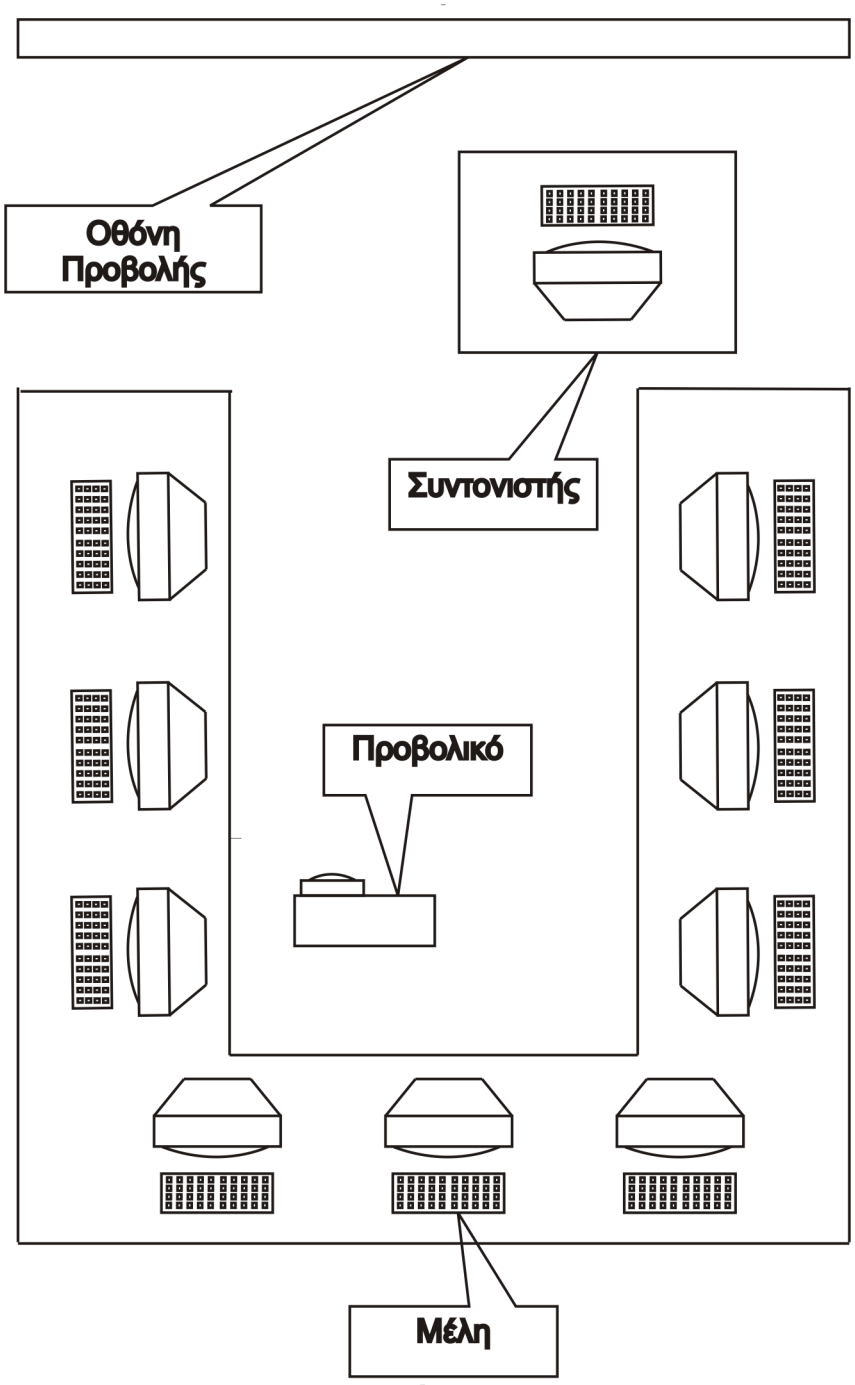
Τα Συστήματα Υποστήριξης Ομαδικών Αποφάσεων ολοκληρώνουν τις δυνατότητες των Συστημάτων Υποστήριξης Αποφάσεων με τις δυνατότητες των Συστημάτων Υποστήριξης Ομάδων. Στόχος τους είναι να διευκολύνουν τη συνεργατική διεξαγωγή αναλύσεων και την παραγωγή εναλλακτικών προτάσεων με τρόπο που να ενισχύει τη δυναμική της ομάδας. Στις επιμέρους δυνατότητες των δύο κατηγοριών συστημάτων προστίθενται η κοινή διαχείριση και χρήση των μοντέλων και η κοινή διεξαγωγή αναλύσεων. Ως προς τη διεξαγωγή αναλύσεων, υπάρχουν διάφορα εναλλακτικά σχήματα. Σύμφωνα με την απλούστερη εκδοχή, τα μέλη της ομάδας εργάζονται μόνοι τους και χρησιμοποιούν ένα κοινό ΣΥΑ για να πραγματοποιούν τις αναλύσεις τους. Κατά το συνεργατικό στάδιο θα προσκομίσουν τα αποτελέσματα και τις απόψεις τους, συνεισφέροντας με αυτόν τον τρόπο στην κοινή αναζήτηση. Μία άλλη εκδοχή προβλέπει τη συμμετοχή ενός εξειδικευμένου αναλυτή. Ο αναλυτής συναντά τα μέλη ατομικά ή συλλογικά, συζητά μαζί τους, πραγματοποιεί τις αναλύσεις του και στη συνέχεια συμμετέχει στη διάσκεψη. Εναλλακτικά, ο αναλυτής μπορεί να εργάζεται κατά τη διάρκεια της διάσκεψης και να πραγματοποιεί αναλύσεις σε συνεργασία με τα μέλη της ομάδας. Σύμφωνα με την τρίτη εκδοχή, τα μέλη της ομάδας χρησιμοποιούν το σύστημα για τη διεξαγωγή αναλύσεων και εργάζονται συλλογικά ή οργανωμένοι σε ομάδες. Η συνεργασία μπορεί να είναι σύγχρονη ή ασύγχρονη και να διεξάγεται με φυσική παρουσία ή με τηλεδιάσκεψη.

Συνεδριάσεις που σχετίζονται με θέματα στρατηγικού προσανατολισμού ή παρόμοιας εμβέλειας ζητήματα, συχνά πραγματοποιούνται με φυσική παρουσία σε ειδικά διαμορφωμένους χώρους και με τη βοήθεια ενός ΣΥΟΑ. Οι χώροι αυτοί ονομάζονται Δωμάτια Αποφάσεων (Decision Rooms) ή Δωμάτια Ηλεκτρονικών Συνεδριάσεων (Electronic Meeting Rooms). Πρόκειται για αίθουσες ειδικά διαμορφωμένες και εξοπλισμένες με κατάλληλο υλικό και λογισμικό υπολογιστών. Οι θέσεις των μελών της ομάδας εργασίας είναι διατεταγμένες σε ημικυκλικά τόξα, έτσι ώστε να επιτρέπουν την οπτική επαφή μεταξύ των μελών. Στο κέντρο του ημικύκλιου είναι τοποθετημένος ο συντονιστής – μεσολαβητής. Το έργο του συντονιστή είναι σημαντικό για την επιτυχία της συνεδρίασης (Miranda & Bostrom, 1999). Τα μέλη της ομάδας διαθέτουν ατομικούς δικτυωμένους υπολογιστές, ενώ ένα προβολικό μηχάνημα είναι συνδεδεμένο σε ένα σέρβερ και προβάλλει εικόνα σε μια κεντρική οθόνη ορατή σε όλα τα μέλη. Τα μέλη μπορούν αν εργάζονται ατομικά ή σε ομάδες, να πειραματίζονται με μοντέλα και να διεξάγουν αναλύσεις. Μοντέλα και αναλύσεις μπορεί να παρουσιάζονται και στην ολομέλεια μέσω του προβολικού μηχανήματος, επιτρέποντας τη συνεργασία και με αυτόν τον τρόπο. Συνεδριάσεις σε δωμάτια αποφάσεων είναι του τύπου ίδιος χώρος – ίδιος χρόνος. Στο Σχήμα 2.6 παρουσιάζεται διαγραμματικά ένα Δωμάτιο Αποφάσεων.

Για συνεδριάσεις στις οποίες τα μέλη βρίσκονται σε διαφορετικό χώρο, χρησιμοποιούνται συστήματα δικτυωμένων υπολογιστών, εξοπλισμένων με κατάλληλο λογισμικό. Κατά κανόνα, τα συστήματα αυτά προσφέρουν δυνατότητες τηλεδιάσκεψης για σύγχρονη εργασία. Εναλλακτικά, η συνεργασία μπορεί να είναι ασύγχρονη και τα μέλη να συμμετέχουν διαφορετικές χρονικές στιγμές. Στην περίπτωση αυτή ορίζονται προθεσμίες για την εκτέλεση των εργασιών. Αναλύσεις, τροποποιήσεις μοντέλων, σχόλια κλπ. καταγράφονται

και διαβιβάζονται στα μέλη της ομάδας όταν συνδεθούν με το σύστημα. Μια άλλη σημαντική λειτουργία των ΣΥΟΑ είναι ότι μπορούν να καταγράφουν τη συνεδρίαση. Σε κατοπινό χρόνο, τα μέλη ή άλλοι ενδιαφερόμενοι, μπορούν να ανατρέξουν στα γεγονότα της συνεδρίασης, να ελέγξουν τις πληροφορίες που χρησιμοποιήθηκαν, να θυμηθούν τις προτάσεις που κατατέθηκαν και γενικώς να κατανοήσουν τον τρόπο με τον οποίο ελήφθη η τελική απόφαση.

Ένα επιπλέον σημαντικό πλεονέκτημα των ΣΥΟΑ είναι ότι μπορούν να εξασφαλίσουν την ανωνυμία. Τα μέλη της συνεδρίασης μπορούν να καταθέσουν τις απόψεις τους και να ψηφίσουν χωρίς να αποκαλύψουν την ταυτότητα τους. Η ανωνυμία έχει θετικές επιπτώσεις με διάφορους τρόπους. Προσωπικές συμπάθειες ή αντιπάθειες, φιλίες ή και σχέσεις εξάρτησης είναι πιθανό να επηρεάσουν την ψήφο των μελών. Επίσης, η προσωπικότητα ή η διοικητική θέση ενός ατόμου, μπορεί να επηρεάσει περισσότερο από την αξία των επιχειρημάτων του. Οι ανώνυμες προτάσεις τυγχάνουν αντικειμενικότερης κριτικής. Με τον τρόπο αυτό, βελτιώνεται η διαδικασία λήψης αποφάσεων και μπορούν να επιτευχθούν καλύτερες αποφάσεις.



Σχήμα 2.6 *Δωμάτιο Αποφάσεων*

2.2.7 Συστήματα Υποστήριξης Διοίκησης

Τα Συστήματα Υποστήριξης Διοίκησης (ΣΥΔ) (Executive Support Systems (ESS)) είναι μια ειδική κατηγορία ΣΥΑ. Θα μπορούσε να πει κανείς ότι είναι η κορυφαία εκδοχή των ΣΥΑ, με την έννοια ότι απευθύνονται στην κορυφή της διοικητικής πυραμίδας των οργανισμών. Για την ακρίβεια, πρόκειται για Συστήματα Υποστήριξης Αποφάσεων ειδικά σχεδιασμένα για τις ανάγκες των πιο υψηλόβαθμων διοικητικών στελεχών. Τα ανώτατα διοικητικά στελέχη χαράσσουν τη στρατηγική του οργανισμού και παρακολουθούν την επίδοση του. Στα πλαίσια των καθηκόντων τους παρακολουθούν τις δραστηριότητες των ανταγωνιστών, τη μεταβολή των συνθηκών της αγοράς, προσπαθούν να εντοπίσουν προβλήματα και ευκαιρίες και να προβλέψουν τάσεις. Για τον έλεγχο της επίδοσης, χρησιμοποιούν τους λεγόμενους Κύριους Δείκτες Επίδοσης, ειδικούς αριθμοδείκτες που αναφέρονται σε ζητήματα κρίσιμης σημασίας για τον οργανισμό, όπως η κερδοφορία, οικονομικά στοιχεία, ανθρώπινους και τεχνολογικούς πόρους, στοιχεία πωλήσεων κλπ. Τα ΣΥΔ προσφέρουν στα στελέχη τα μέσα για την αποτελεσματικότερη και ταχύτερη εκτέλεση των καθηκόντων τους. Τους παρέχουν την απαραίτητη πληροφόρηση, καθώς επίσης και εργαλεία διεξαγωγής αναλύσεων, επιτρέποντας τους να κατανοήσουν καλύτερα τις συνθήκες, να ελέγξουν τον οργανισμό και να βελτιώσουν τις αποφάσεις τους. Με τον τρόπο αυτό, η ποιοτική πληροφόρηση μετατρέπεται σε στρατηγικό πόρο. Σύμφωνα με τη Sauter (1997), τα ΣΥΔ επιτυγχάνουν τα εξής:

- Λειτουργούν ως εργαλεία στρατηγικού σχεδιασμού.
- Βελτιώνουν την ποιότητα των αποφάσεων των κορυφαίων στελεχών.
- Μειώνουν τον χρόνο που χρειάζεται για τον εντοπισμό προβλημάτων και ευκαιριών.
- Βελτιώνουν την ποιότητα των σχεδιασμών στα κορυφαία επίπεδα των οργανισμών.
- Παρέχουν μηχανισμούς που βελτιώνουν τον έλεγχο των οργανισμών.
- Παρέχουν καλύτερη πρόσβαση σε δεδομένα και μοντέλα.

Τα ΣΥΔ παρέχουν υποστήριξη στο στρατηγικό επίπεδο μιας επιχείρησης. Διαθέτουν στοιχεία σχεδίασης παρόμοια με τα ΣΥΑ και προσφέρουν πληροφόρηση και αναλυτικές δυνατότητες με τη χρήση μοντέλων. Στόχος τους είναι η υποστήριξη αδόμητων και ημιδομημένων αποφάσεων. Σε στρατηγικό επίπεδο, οι ανάγκες για πληροφόρηση είναι πολύ υψηλές. Τα ανώτατα στελέχη πρέπει να έχουν στη διάθεση τους την ευρύτερη δυνατή βάση πληροφοριών. Οι πληροφορίες που συγκεντρώνει το σύστημα αφορούν το εσωτερικό του οργανισμού αλλά και το εξωτερικό του περιβάλλον. Τα εσωτερικά δεδομένα προέρχονται κυρίως από τα συστήματα παρακολούθησης συναλλαγών και τους εταιρικούς διαδικτυακούς σέρβερ, ενώ τα εξωτερικά δεδομένα είναι ειδήσεις, ανεξάρτητες βάσεις δεδομένων, οικονομικές πληροφορίες, ενώ τελευταία, σημαντικές γίνονται οι πληροφορίες από το Web 2.0. Όλα τα σημαντικά γεγονότα και οι τάσεις, εσωτερικές και εξωτερικές, πρέπει να είναι καταγεγραμμένες και διαθέσιμες στον χρήστη. Επίσης, τα δεδομένα που τηρεί το σύστημα είναι και ποσοτικά και ποιοτικά.

Τα ΣΥΔ τηρούν και ιστορική και τρέχουσα πληροφορία. Τα στελέχη ανατρέχουν σε στοιχεία του παρελθόντος για να εκτελέσουν αναλύσεις και να εξάγουν συμπεράσματα, πρέπει όμως να έχουν και σαφή εικόνα της τωρινής κατάστασης του οργανισμού. Για τον λόγο αυτό, η τήρηση τρέχουσας πληροφορίας έχει βαρύνουσα σημασία. Ιδανική συνθήκη για το σύστημα είναι η απόλυτη πληρότητα πληροφόρησης. Το σύστημα πρέπει να είναι ικανό να παρέχει πληροφόρηση για κάθε σημαντικό ζήτημα που αφορά τον οργανισμό. Επίσης, για κάθε ζήτημα, οποιαδήποτε πληροφορία είναι σημαντική πρέπει να είναι καταγεγραμμένη και διαθέσιμη. Πέραν της πληρότητας, σημαντικό στοιχείο είναι η ακρίβεια. Εσφαλμένη ή αποκλίνουσα πληροφόρηση μπορεί να οδηγήσει σε εσφαλμένες αποφάσεις. Όταν οι αποφάσεις αυτές αφορούν ζητήματα στρατηγικού προσανατολισμού, οι επιπτώσεις μπορεί να είναι ολέθριες. Τα ανώτατα στελέχη των επιχειρήσεων λειτουργούν υπό την πίεση του χρόνου. Κατά κανόνα, ο διαθέσιμος χρόνος τους είναι περιορισμένος και για τον λόγο αυτό, το σύστημα πρέπει να λειτουργεί με ταχύτητα. Ο χρόνος αναμονής για μια αναζήτηση δεν πρέπει να ξεπερνά τα λίγα δευτερόλεπτα. Επίσης, η πληροφορία πρέπει να παραδίδεται στον παραλήπτη της έγκαιρα. Τα ΣΥΔ αντλούν πληροφορίες από Αποθήκες Δεδομένων, οι οποίες με τη σειρά τους φορτώνουν δεδομένα από άλλα συστήματα. Η όλη διαδικασία εξαγωγής, μεταφοράς και φόρτωσης των δεδομένων πρέπει να είναι ταχεία ώστε το στέλεχος να έχει διαθέσιμη την «εικόνα της τελευταίας στιγμής». Καθυστερήσεις στην ενημέρωση έχουν κόστος. Μία ευκαιρία που εντοπίστηκε καθυστερημένα μπορεί κάλλιστα να είναι μια χαμένη ευκαιρία, εάν κάποιος ανταγωνιστής έχει προλάβει να την εκμεταλλευτεί. Επίσης, προβλήματα που δεν εντοπίζονται έγκαιρα διογκώνονται ή προκαλούν αλυσιδωτές αντιδράσεις σε άλλες διαδικασίες του οργανισμού.

Η πληροφόρηση που παρέχει το ΣΥΔ είναι συνοπτική και αποκαλυπτική για την πραγματική κατάσταση του οργανισμού. Παρουσιάζονται συμπυκνωμένες πληροφορίες που σχετίζονται με ζητήματα τα οποία είναι

κρίσιμα για τον οργανισμό. Επιπλέον όμως, τα συστήματα αυτά είναι πολύ διαδραστικά και επιτρέπουν την εργασία σε διάφορα επίπεδα γενίκευσης. Ο χρήστης αρχικά τροφοδοτείται με τη γενική εικόνα, μπορεί όμως, εάν το επιθυμήσει, να εμβαθύνει σε λεπτομέρειες. Αν για παράδειγμα ο χρήστης διαπιστώσει από μια συνοπτική έκθεση πωλήσεων ότι υπάρχει κάμψη των πωλήσεων σε μια περιοχή τον τελευταίο μήνα, μπορεί να δει τα αναλυτικά στοιχεία πωλήσεων για την περιοχή και τον μήνα αυτόν. Εξειδικεύοντας περαιτέρω την ανάλυση του, μπορεί να διαπιστώσει αν η πτωτική τάση αφορά το σύνολο των πωλήσεων στην περιοχή ή αν αφορά μια συγκεκριμένη κατηγορία προϊόντων, πχ τρόφιμα. Αφού αποκτήσει ακριβή εικόνα του προβλήματος, θα αναζητήσει την αιτία του, η οποία στο παραπάνω παράδειγμα μπορεί να είναι η δραστηριοποίηση στην περιοχή που παρουσιάστηκε το πρόβλημα ενός ανταγωνιστή εξειδικευμένου στον κλάδο των τροφίμων. Με τον τρόπο αυτό, ο χρήστης κινείται από το γενικό στο ειδικό, αντλώντας την απαραίτητη πληροφόρηση.

Τα ΣΥΔ διαθέτουν αναλυτικές δυνατότητες. Μια σειρά από μοντέλα επιτρέπουν στον χρήστη τη διεξαγωγή αναλύσεων και την εξαγωγή συμπερασμάτων. Χαρακτηριστικό των μοντέλων που περιλαμβάνει ένα ΣΥΔ είναι ότι αναφέρονται σε στρατηγικά θέματα και σχετίζονται με τους κύριους δείκτες επίδοσης του οργανισμού. Με τη βοήθεια των μοντέλων, ο χρήστης πραγματοποιεί αναλύσεις what-if, αναζήτησης στόχου κλπ. ή εκτελεί εργασίες στρατηγικού σχεδιασμού. Επίσης, μπορεί να εντοπίσει τάσεις και παρεκκλίσεις. Σημαντικό είναι να παρέχεται η δυνατότητα συγκρίσεων. Ο χρήστης συγκρίνει αποτελέσματα γεωγραφικών περιοχών, χρονικών περιόδων κλπ. ώστε να ανακαλύψει ιδιομορφίες και επιμέρους επιδόσεις. Πολύ συνηθισμένη είναι η σύγκριση μεταξύ πραγματικών επιδόσεων και καθορισμένων στόχων. Με τον τρόπο αυτό, ελέγχεται το κατά πόσον οι επιδόσεις του οργανισμού είναι ευθυγραμμισμένες με τους προγραμματισμένους στόχους. Εάν διαπιστωθούν αποκλίσεις, αναζητούνται τα αίτια, αντιμετωπίζονται τα προβλήματα που προκάλεσαν τις αποκλίσεις ή αναθεωρούνται οι στόχοι σε περίπτωση που διαπιστωθεί ότι ήταν υπεραισιόδοξοι.

Ένα από τα σημαντικότερα αντικείμενα εργασίας των ανώτατων διοικητικών στελεχών είναι η διατύπωση προβλέψεων. Σε μεγάλο βαθμό η λήψη αποφάσεων και η κατάστρωση σχεδίων απαιτεί τη διατύπωση προβλέψεων σχετικά με μελλοντικά συμβάντα ή μεγέθη. Για παράδειγμα, η κατάρτιση σχεδίων για προμήθεια υλικών, πρόσληψη προσωπικού, επενδύσεις σε μηχανολογικό εξοπλισμό κλπ. σχετίζεται άμεσα με προβλέψεις για τη μελλοντική ζήτηση. Η πρόβλεψη βοηθά τα στελέχη να μειώσουν την αβεβαιότητα σχετικά με μελλοντικά συμβάντα και να καταρτίσουν πιο επιτυχημένα σχέδια. Για τον λόγο αυτό, τα ΣΥΔ πρέπει να περιλαμβάνουν πολλά μοντέλα ικανά να διατυπώνουν ακριβείς προβλέψεις. Τα μοντέλα πρόβλεψης έχουν βαρύνουσα σημασία στα ΣΥΔ. Ιδιαίτερης αξίας είναι επίσης ο εντοπισμός εξαιρέσεων. Δείκτες με υπερβολικά μεγάλες ή μικρές τιμές είναι πιθανό να σηματοδοτούν επιχειρηματικές ευκαιρίες ή προβλήματα στις διαδικασίες και υποδομές του οργανισμού. Είναι τόσο μεγάλη η σημασία των εξαιρέσεων, ώστε στην επιστήμη της Διοίκησης Επιχειρήσεων υπάρχει καθιερωμένος όρος «Διοίκηση μέσω Εξαιρέσεων» (Management by Exception). Ο όρος αυτός περιγράφει ένα στυλ διοίκησης, το οποίο επικεντρώνει στις εξαιρέσεις. Τα ΣΥΔ πρέπει να περιλαμβάνουν αποτελεσματικά μοντέλα για τον εντοπισμό και την ανάλυση των εξαιρέσεων.

Στα ΣΥΔ υπάρχουν ειδικές απαιτήσεις και από το σύστημα διεπαφής χρήστη. Βασικό ζητούμενο είναι η ευχρηστία. Οι Watson and Satzinger (1994) αναφέρουν ότι το σύστημα πρέπει να υπερβαίνει τα όρια της απλής φιλικότητας προς τον χρήστη και να γίνεται δελεαστικό και «διαισθητικό» (intuitive). Συνήθως, τα στελέχη έχουν έναν προσωπικό τρόπο εργασίας και επιθυμούν να είναι η διεπαφή προσαρμοσμένη σε αυτόν. Επιπλέον, οι ανάγκες τους αλλάζουν ταχύτατα, οπότε ο χρήστης πρέπει να μπορεί να τροποποιεί τη διεπαφή σύμφωνα με τις νέες ανάγκες. Μεγάλες πρέπει να είναι και οι δυνατότητες διαδραστικής χρήσης. Παράθυρα που μοιάζουν με άκαμπτες ηλεκτρονικές αναφορές δεν έχουν μεγάλη αξία. Ο χρήστης πλοηγείται στο σύστημα με τρόπο εύκολο, γρήγορο και αποτελεσματικό, αναζητώντας πρόσθετη πληροφόρηση, συγκρίνοντας στοιχεία και εκτελώντας αναλύσεις. Η πληροφορία πρέπει να παρουσιάζεται με τρόπο απολύτως κατανοητό και σαφή, και η πρόσβαση σε αυτήν πρέπει να επιτυγχάνεται με λίγες απλές κινήσεις. Τα μέσα παρουσίασης της πληροφορίας ποικίλουν και γίνεται εκτεταμένη χρήση γραφικών.

2.2.8 Ευφυή Συστήματα Υποστήριξης Αποφάσεων

Τα Ευφυή Συστήματα Υποστήριξης Αποφάσεων (ΕΣΥΑ) (Intelligent Decision Support Systems (IDSS)) αποτελούν εξελιγμένη εκδοχή των ΣΥΑ. Διαφοροποιούνται από τα υπόλοιπα ΣΥΑ από το γεγονός ότι κάνουν χρήση μεθοδολογιών, οι οποίες προέρχονται από την Τεχνητή Νοημοσύνη και τη Μηχανική Μάθηση. Ενδεικτικά και όχι περιοριστικά, τέτοιες μεθοδολογίες είναι τα έμπειρα συστήματα, η περιπτωσιολογική συλλογιστική, η ασαφής λογική, τα Νευρωνικά Δίκτυα, οι γενετικοί αλγόριθμοι, οι ευφυείς πράκτορες κλπ. Ο εμπλουτισμός των αναλυτικών τεχνικών των ΣΥΑ με μεθοδολογίες Τεχνητής Νοημοσύνης αποφέρει νέες δυνατότητες για την εξαγωγή συμπερασμάτων και τη λήψη αποφάσεων, και αυξάνει την ακρίβεια, την αξιοπιστία και τη

χρηστικότητα του συστήματος. Τα ΕΣΥΑ αξιοποιούν προηγούμενη γνώση για να εξάγουν συμπεράσματα για τρέχουσες παρεμφερείς καταστάσεις. Οι τεχνικές της Τεχνητής Νοημοσύνης διαθέτουν κάποια σημαντικά χαρακτηριστικά, τα οποία στερούνται οι αντίστοιχες Στατιστικές τεχνικές. Αρκετές από τις νέες τεχνικές είναι ικανές να χειρίζονται θορυβώδη δεδομένα, δεδομένα δηλαδή που περιλαμβάνουν τυχαία κυμαινόμενες τιμές. Κατά περίπτωση δέχονται ως είσοδο και αριθμητικές και ονομαστικές τιμές, και μπορούν να χειριστούν δεδομένα στα οποία λείπουν τιμές. Οι ευφυείς τεχνικές προσφέρουν πολύπλευρη βοήθεια για τη λήψη αποφάσεων. Μπορούν να εντοπίζουν προβλήματα τα οποία χρίζουν προσοχής, να επιλύουν προβλήματα ή να συμβάλλουν στην επίλυση τους, καθώς και να παρέχουν βοήθεια με τη μορφή της συμβουλής, της ανάλυσης ή της αξιολόγησης. Υπό μια έννοια, συμβάλλουν στην υπέρβαση των ορίων της ανθρώπινης αντιληπτικής ικανότητας. Χάρη στα ιδιαίτερα χαρακτηριστικά τους και τις δυνατότητες τους, τα ΕΣΥΑ αυξάνουν την αποτελεσματικότητα των χρηστών και βοηθούν στη λήψη βελτιωμένων αποφάσεων. Αναλυτική αναφορά σε μεθόδους Τεχνητής Νοημοσύνης, όπως τα Νευρωνικά Δίκτυα, τα Δένδρα Αποφάσεων κλπ. γίνεται σε επόμενα κεφάλαια αυτού του βιβλίου.

Αναφορές / Βιβλιογραφία

Black, D. (1948). On the Rationale of Group Decision-Making. *Journal of Political Economy, 56*(1), 23-34.

Bonczek, R., Holsapple, C., & Whinston, A. (1980). The Evolving Roles of Models in Decision Support Systems. *Decision Sciences*, *11*(2), 337-356. doi: 10.1111/j.1540-5915.1980.tb01143.x

Davis, M. (1999). Smiling for the Camera. *Journal of Business Strategy, 20*(3), 20-24. doi: 10.1108/eb040002

Gorry, G. A., & Scott Morton, M. S. (1971). A Framework for Management Information Systems. *Sloan Management Review, 13*, 21-36.

Huber, G. P., Valacich, J. S., & Jessup, L. M. (1993). *A Theory of the Effects of Group Support Systems on an Organization's Nature and Decisions. Group Support Systems: New Perspectives.* New York, NY: Macmillan Publishing Co.

Keen, P. G. W. (1980). Adaptive Design for Decision Support Systems. *SIGMIS Database, 12*(1-2), 15-25. doi: 10.1145/1017654.1017659

Keen, P. G. W., & Scott-Morton, M. S. (1978). *Decision Support Systems: An Organizational Perspective.* Reading, MA: Addison-Wesley.

Laudon, K. C., & Laudon, J. P. (1998). *Information Systems and the Internet: A problem-solving approach*. New York, NY: Dryden Press.

Mintzberg, H. (1975). The Manager's Job: Folklore and Fact. *Harvard Business Review, 53*(4), 49-61.

Mintzberg, H. (1990). *Mintzberg on Management: Inside our Strange World of Organizations*. New York, NY: Free Press.

Miranda, S. M., & Bostrom, R. P. (1999). Meeting Facilitation: Process versus Content Interventions. *Journal of Management Information Systems, 15*(4), 89-114.

McLeod, R. (1990). *Management Information Systems*. New York, NY: Macmillan Publishing Company.

Power, D. J., Burstein, F., & Sharda, R. (2011). Reflections on the Past and Future of Decision Support Systems: Perspective of Eleven Pioneers. In D. Schuff, D. Paradice, F. Burstein, D.J. Power & R. Sharda (Eds.), *Decision Support Annals of Information Systems* (pp. 25-48). New York, NY: Springer.

Sharda, R., Delen, D., & Turban, E. (2015). *Business Intelligence and Analytics: Systems for Decision Support*. Upper Saddle River, NJ: Prentice Hall.

Sauter, V. (1997). *Decision Support Systems*. New York, NY: John Wiley and Sons Inc.

Simon, H. A. (1957). *Models of Man: Social and Mathematical Essays on Rational Human Behavior in a Social Setting*. New York, NY: John Wiley and Sons Inc.

Simon, H. A. (1977). *The New Science of Management Decision*. Englewood Cliffs, NJ: Prentice-Hall.

Smith, B. C., Gunther, D. P., Rao, B. V., & Ratlife, R. M. (2001). E-Commerce and Operations Research in Airline Planning, Marketing and Distribution. *Interfaces, 31*(2), 37-55. doi: 10.1287/inte.31.2.37.10627

Turban, E., Aronson, J. E., & Liang, T. P. (2005). *Decision Support Systems and Intelligence Systems*. New Jersey, NJ: Pearson Education Inc.

Watson, H. J., Carroll, A. B., & Mann, R. I. (1987). *Information Systems for Management*. Plano, TX: Business Publications Inc.

Watson, H. J., & Satzinger, J. (1994). Guidelines for Designing EIS Interfaces: Meeting Executive's Information Needs. *Information Systems Management, 11(4), 46-52. doi:10.1080/07399019408964670*

3 Μοντελοποίηση Προβλημάτων

Σύνοψη

Το παρόν Κεφάλαιο καλύπτει τη θεματική ενότητα της Μοντελοποίησης Προβλημάτων. Η χρήση μοντέλων για την αντιμετώπιση προβλημάτων, τη διεξαγωγή αναλύσεων και τη λήψη αποφάσεων είναι μια πολύ διαδεδομένη τακτική στην Οικονομία και τη Διοίκηση Επιχειρήσεων. Τα μοντέλα σήμερα αποτελούν βασικό εργαλείο των Συστημάτων Επιχειρηματικής Ευφυΐας. Στο παρόν κεφάλαιο γίνεται μια συνοπτική παρουσίαση των βασικών εννοιών και των πιο δημοφιλών τεχνικών, που αναφέρονται στη χρήση μοντέλων για τη λήψη επιχειρηματικών αποφάσεων. Αρχικά αναλύεται η έννοια του μοντέλου και αναφέρονται περιπτώσεις όπου ενδείκνυται η χρήση τους. Υπάρχουν πολλές κατηγορίες μοντέλων ανάλογα με το κριτήριο που χρησιμοποιείται κάθε φορά. Τα μοντέλα, ανάλογα με τα χαρακτηριστικά τους, χωρίζονται σε εικονικά, αναλογικά και συμβολικά. Ανάλογα με την πιθανοκρατική φύση τους, χωρίζονται σε αιτιοκρατικά ή ντετερμινιστικά και στοχαστικά. Τέλος, ανάλογα με την ικανότητα τους να αναπαριστούν χρονικά μεταβαλλόμενες καταστάσεις, χωρίζονται σε στατικά και δυναμικά. Όλες οι παραπάνω κατηγορίες μοντέλων παρουσιάζονται και αναλύονται. Βασικές έννοιες στη μοντελοποίηση προβλημάτων είναι η βεβαιότητα, η αβεβαιότητα και το ρίσκο. Οι έννοιες αυτές παρατίθενται και σχολιάζονται. Στη συνέχεια του κεφαλαίου εξετάζονται μερικές από τις πιο δημοφιλείς τεχνικές μοντελοποίησης. Η Ανάλυση Αποφάσεων είναι μια μέθοδος που εφαρμόζεται σε περιπτώσεις προβλημάτων με περιορισμένο αριθμό εναλλακτικών επιλογών. Δύο εργαλεία που χρησιμοποιούνται στην Ανάλυση Αποφάσεων είναι τα Διαγράμματα Επιρροής και οι Πίνακες Αποφάσεων ή Πίνακες Ανταμοιβών. Τα Διαγράμματα Επιρροής αναπαριστούν αποφάσεις, μεταβλητές και αποτελέσματα ως ένα κατευθυνόμενο ακυκλικό γράφο και χρησιμοποιούνται για τη λογική αναπαράσταση ενός μοντέλου. Οι Πίνακες Αποφάσεων οργανώνουν σε μορφή πίνακα συνδυασμούς καταστάσεων, αποφάσεων και αποτελεσμάτων, και καθορίζουν μια προτεινόμενη απόφαση με τη χρήση της Αναμενόμενης Τιμής. Ένα μαθηματικό μοντέλο αποτελείται από τις μεταβλητές απόφασης, τις μη ελεγχόμενες μεταβλητές, τις μεταβλητές παραμέτρων και τις μαθηματικές σχέσεις που τις συνδέουν. Οι έννοιες αυτές αναλύονται και παρουσιάζεται η διαγραμματική αναπαράσταση ενός μαθηματικού μοντέλου. Ο Γραμμικός Προγραμματισμός είναι ίσως η πλέον διαδεδομένη μέθοδος λήψης αποφάσεων στη Διοίκηση Επιχειρήσεων. Αντικείμενο του Γραμμικού Προγραμματισμού είναι η εύρεση της μικρότερης ή μεγαλύτερης τιμής μιας συνάρτησης, με επιλογή κατάλληλων τιμών για τις μεταβλητές εισόδου. Αναπτύσσονται ζητήματα σχετικά με τον Γραμμικό Προγραμματισμό και παρατίθεται παράδειγμα σχετικού προβλήματος, που επιλύεται με γεωμετρικό τρόπο. Με τη βοήθεια των αναλύσεων what-if, τα διοικητικά στελέχη μελετούν τι θα συμβεί εάν μεταβληθούν μεταβλητές απόφασης ή μη ελεγχόμενες μεταβλητές. Χρησιμοποιώντας ένα μοντέλο, πειραματίζονται μεταβάλλοντας κάποιες τιμές εισόδου και ελέγχοντας τις επιπτώσεις στη λύση του προβλήματος. Επίσης, πραγματοποιούν αναλύσεις αναζήτησης στόχου ορίζοντας τα αποτελέσματα και εξετάζοντας τις υπολογιζόμενες τιμές στις μεταβλητές εισόδου. Η Ανάλυση Ευαισθησίας μελετά το κατά πόσο οι μεταβολές στις τιμές των μεταβλητών εισόδου και των παραμέτρων επηρεάζουν το τελικό αποτέλεσμα. Οι ευρετικές μέθοδοι (heuristics) επιτρέπουν την ταχύτερη και οικονομικότερη αντιμετώπιση ασθενώς δομημένων ή περίπλοκων προβλημάτων. Οι Γενετικοί Αλγόριθμοι είναι μια από τις πιο γνωστές περιπτώσεις ευρετικών μεθόδων και μιμούνται τη διαδικασία της φυσικής επιλογής. Χρησιμοποιούνται σε περιπτώσεις πολυκριτήριας βελτιστοποίησης ή αντιμετώπισης περίπλοκων προβλημάτων. Τέλος, η προσομοίωση είναι μια τεχνική διεξαγωγής πειραμάτων, όπου ο χρήστης τροφοδοτεί με διάφορες τιμές μεταβλητών εισόδου ένα μοντέλο, το οποίο μιμείται τη συμπεριφορά του συστήματος και παρατηρεί τις αντιδράσεις του.

Προαπαιτούμενη Γνώση

Το παρόν Κεφάλαιο εισάγει τον αναγνώστη σε μια σειρά θεμάτων μοντελοποίησης αποφάσεων. Πολλές διαθέσιμες πηγές παρέχουν γνώσεις υποδομής, αλλά και περισσότερες λεπτομέρειες σχετικά με τις έννοιες και τις μεθόδους. Για θέματα μοντελοποίησης, ανάλυσης αποφάσεων, ανάλυσης ευαισθησίας, χρονοσειρών, καθώς και πολλών άλλων τεχνικών με χρήση Υπολογιστικών Φύλλων, ο αναγνώστης μπορεί να ανατρέξει στο βιβλίο του Ragsdale (2014). Στην εργασία του Dantzig (1951) παρουσιάστηκε η μέθοδος simplex. Πολλά σύγχρονα θέματα Γραμμικού Προγραμματισμού αναπτύσσονται στα βιβλία της σειράς International Series in Operations Research and Management Science. Στην εργασία των Gigerenzer and Gaissmaier (2011) γίνεται αναλυτική παρουσίαση των Ευρετικών Μεθόδων και περιπτώσεων εφαρμογής τους. Αναδρομή σε εργασίες ορόσημα σχετικά με τους Γενετικούς Αλγορίθμους υπάρχει στο Fogel (1998), ενώ γνώσεις υποδομής βρίσκονται στους Fraser (1957), Bremermann (1958) και Holland (1975). Σύγχρονα θέματα Γενετικών Αλγορίθμων παρουσιάζονται στο

Simon (2013). Γνώσεις υποδομής για την προσομοίωση παρέχονται στα βιβλία των Shannon (1975) και Law and Kelton (1991), ενώ σύγχρονα θέματα εφαρμογής προσομοίωσης για τη διαχείριση του χρηματοοικονομικού ρίσκου μπορεί να βρει ο αναγνώστης στο βιβλίο των Chan and Wong (2013).

3.1 Η έννοια του μοντέλου

Η χρήση μοντέλων είναι μια διαδεδομένη τακτική που έχει εφαρμοστεί επί μακρόν σε πολλούς επιστημονικούς κλάδους, όπως στη Φυσική, στην αεροναυπηγική, στην οικονομία κλπ. Τα πρώιμα Συστήματα Υποστήριξης Αποφάσεων στηριζόταν αποκλειστικά στη χρήση μοντέλων, ενώ τα σύγχρονα ΣΥΑ τα περιλαμβάνουν ως αναπόσπαστο τμήμα τους. Τα μοντέλα αποτελούν βασικό εργαλείο της Αναλυτικής των Επιχειρήσεων και των Συστημάτων Επιχειρηματικής Ευφυΐας. Καθημερινά, χιλιάδες αναλυτές τα χρησιμοποιούν για να αναλύουν επιχειρηματικά δεδομένα, να εξάγουν συμπεράσματα και να βελτιώνουν με τον τρόπο αυτό τη διαδικασία λήψης αποφάσεων.

Ένα μοντέλο είναι μια αφαιρετική αναπαράσταση ενός πραγματικού συστήματος. Αυτό σημαίνει ότι υλοποιεί ορισμένες μόνο ιδιότητες και χαρακτηριστικά του πραγματικού συστήματος, και απορρίπτει τις υπόλοιπες. Η ανάγκη για επιλογή προκύπτει από το γεγονός ότι η πραγματικότητα είναι εξαιρετικά περίπλοκη και δεν μπορεί να αναπαρασταθεί πλήρως. Το μοντέλο απλοποιεί την πραγματικότητα, επιλέγοντας ένα τμήμα της. Ένα συνηθισμένο παράδειγμα μοντέλου είναι ένα αεροπλάνο παιχνίδι. Το παιδί παίζοντας, κατανοεί βασικά χαρακτηριστικά του αεροπλάνου, όπως το ατρακτοειδές σχήμα του, την ύπαρξη φτερών και ουράς, την ύπαρξη ενός συστήματος πρόωσης (έλικες ή τουρμπίνες) κλπ.

Ένα μοντέλο δεν μπορεί να εκτελέσει όλες τις λειτουργίες του πραγματικού συστήματος. Στην περίπτωση του παιδικού αεροπλάνου το μοντέλο δεν εκτελεί τη βασική λειτουργία, δηλαδή δεν πετάει. Τίθεται επομένως το ερώτημα ποια από τα στοιχεία του πραγματικού συστήματος θα περιληφθούν και ποια θα αποκλειστούν. Η απάντηση στο ερώτημα αυτό δίνεται από τον σκοπό της κατασκευής του μοντέλου. Το μοντέλο κατασκευάζεται για να μελετηθούν ορισμένες λειτουργίες του πραγματικού συστήματος. Πρέπει λοιπόν να συμπεριληφθούν εκείνα τα στοιχεία του συστήματος που καθορίζουν τις συγκεκριμένες λειτουργίες. Άλλα στοιχεία, όσο σημαντικά και αν είναι για το πραγματικό σύστημα, αποκλείονται. Ένα μοντέλο αεροπλάνου, που κατασκευάζεται για να μελετηθεί η αεροδυναμική άντωση των πτερύγων μέσα σε αεροσήραγγα, δεν χρειάζεται να διαθέτει σύστημα πρόωσης, εφόσον η αεροσήραγγα εξασφαλίζει την απαραίτητη ροή αέρα.

Κατά την επιλογή, το κρίσιμο είναι να περιληφθούν όλα τα στοιχεία, τα χαρακτηριστικά και οι ιδιότητες του πραγματικού συστήματος, που επηρεάζουν σημαντικά το φαινόμενο που εξετάζεται. Όπως χαρακτηριστικά είπε ο Αϊνστάιν, «όλα πρέπει να είναι όσο πιο απλά γίνεται αλλά όχι απλούστερα». Ο σχεδιαστής πρέπει να μπορεί να περιλάβει εκείνο το τμήμα της πραγματικότητας που θα επιτρέψει στο μοντέλο να συμπεριφερθεί, σε σχέση με το φαινόμενο που εξετάζεται, όπως και το πραγματικό σύστημα. Είναι προφανές ότι η κατασκευή ενός μοντέλου δεν είναι τετριμμένη εργασία, ιδιαίτερα σε περιπτώσεις αυξημένης πολυπλοκότητας. Ένα καλό παράδειγμα μοντελοποίησης σύνθετου συστήματος είναι αυτό των Gabriel, Kydes, and Whitman (2001), οι οποίοι μοντελοποίησαν την ενεργειακή και οικονομική κατάσταση των ΗΠΑ.

Η βασική χρησιμότητα ενός μοντέλου είναι ότι επιτρέπει τη μελέτη της συμπεριφοράς του συστήματος, σε περιπτώσεις που ο πειραματισμός με το πραγματικό σύστημα δεν είναι εφικτός. Τέτοιες περιπτώσεις είναι εκείνες όπου:

- Πειραματισμός με το πραγματικό σύστημα είναι ιδιαίτερα δαπανηρός. Η μελέτη της πλευστότητας ενός υπερωκεάνιου, σε περίπτωση ρήγματος στα ύφαλα και εισροής υδάτων, δεν μπορεί να πραγματοποιηθεί με πρόκληση ρήγματος στο πραγματικό πλοίο.
- Πειραματισμός με το πραγματικό σύστημα είναι ιδιαίτερα επικίνδυνος. Ένα σχέδιο εκκένωσης αστικής περιοχής, σε περίπτωση πυρηνικού ατυχήματος, είναι μάλλον… απίθανο να δοκιμαστεί με πραγματικό πείραμα.
- Απαιτείται μελέτη φαινομένου με μεγάλη χρονική διάρκεια. Στην περίπτωση αυτή πειραματισμός με το πραγματικό σύστημα μπορεί να σημαίνει αναμονή αιώνων, μέχρι το πείραμα να αποδώσει αποτελέσματα.
- Απαιτούνται πάρα πολλές επαναλήψεις του πειράματος με εναλλακτικά σενάρια, διαφορετικές παραμέτρους κλπ. Κάτι τέτοιο θα ήταν ιδιαίτερα χρονοβόρο. Με τη βοήθεια ενός ηλεκτρονικού μοντέλου και χάρη στην ικανότητα των υπολογιστών να εκτελούν ταχύτατα υπολογισμούς, το πείραμα μπορεί να εκτελεστεί εκατομμύρια φορές.

3.2 Κατηγορίες μοντέλων

Η μεγάλη χρησιμότητα των μοντέλων τα έχει καταστήσει ιδιαίτερα δημοφιλή. Κατά καιρούς έχουν κατασκευαστεί διάφοροι τύποι μοντέλων που παρουσιάζουν ομοιότητες και διαφορές. Μπορεί κανείς να ορίσει διαφορετικές κατηγορίες μοντέλων χρησιμοποιώντας αντίστοιχα κριτήρια. Ο Vercellis (2009) κατηγοριοποιεί τα μοντέλα σύμφωνα με τα χαρακτηριστικά τους, τον βαθμό αβεβαιότητας και τη χρονική τους διάσταση.

Σύμφωνα με τα χαρακτηριστικά τους, τα μοντέλα χωρίζονται σε εικονικά, αναλογικά και συμβολικά:

- **Εικονικά μοντέλα.** Πρόκειται για υλικά μοντέλα, που αντιγράφουν το πρωτότυπο σύστημα και μιμούνται τη συμπεριφορά του. Οι μινιατούρες είναι παράδειγμα εικονικών μοντέλων.
- **Αναλογικά μοντέλα.** Πρόκειται για υλικά μοντέλα, που μπορεί να μη μοιάζουν με το πρωτότυπο σύστημα, συμπεριφέρονται όμως με ανάλογο τρόπο. Τμήματα αεροσκάφους που δοκιμάζονται σε αεροσήραγγα είναι παράδειγμα αναλογικών μοντέλων.
- **Συμβολικά μοντέλα.** Είναι μη υλικές αναπαραστάσεις συστημάτων. Συμβολικά μοντέλα είναι τα μαθηματικά μοντέλα, τα οποία, χρησιμοποιώντας σταθερές, μεταβλητές και μαθηματικές σχέσεις, αναπαριστούν τη συμπεριφορά του συστήματος. Τα μοντέλα που χρησιμοποιούνται στην Επιχειρηματική Ευφυΐα ανήκουν σε αυτήν την κατηγορία.

Μια δεύτερη κατηγοριοποίηση των μοντέλων τα διαχωρίζει σε αιτιοκρατικά ή ντετερμινιστικά και σε στοχαστικά:

- **Αιτιοκρατικά μοντέλα.** Σύμφωνα με τα αιτιοκρατικά ή ντετερμινιστικά μοντέλα, το αποτέλεσμα μιας διαδικασίας καθορίζεται επακριβώς από τις παραμέτρους και τις αρχικές συνθήκες, που καθορίζουν τις παραμέτρους. Συνεπώς, ένα πείραμα που θα επαναληφθεί με τις ίδιες ακριβώς συνθήκες, θα αποδώσει ακριβώς το ίδιο αποτέλεσμα. Ένα μοντέλο της μορφής $y=ax+b$ είναι ντετερμινιστικό, εφόσον η τιμή του y εξαρτάται μονοσήμαντα από τις τιμές των a,b,x.
- **Στοχαστικά μοντέλα.** Στα στοχαστικά μοντέλα εισάγεται ένας βαθμός αβεβαιότητας. Η ίδια διαδικασία δεν αποδίδει πάντα το ίδιο αποτέλεσμα, αλλά ένα σύνολο δυνατών αποτελεσμάτων, όπου για κάθε ένα από αυτά αντιστοιχεί μια πιθανότητα εμφάνισης. Οι τιμές των μεταβλητών εισόδου είναι πιθανολογικά ενδεχόμενα.

Τα προβλήματα του πραγματικού κόσμου είναι κατά κανόνα στοχαστικά. Ωστόσο, η δημιουργία στοχαστικών μοντέλων είναι σαφώς πιο δύσκολη. Για λόγους απλότητας, ένα στοχαστικό πρόβλημα μπορεί να προσεγγιστεί από ένα ντετερμινιστικό μοντέλο, όταν τα στοχαστικά στοιχεία επηρεάζουν οριακά.

Ένας τρίτος τρόπος κατηγοριοποίησης των μοντέλων συναρτάται με την ικανότητα τους να αναπαριστούν χρονικά μεταβαλλόμενες καταστάσεις. Σύμφωνα με αυτό το κριτήριο, τα μοντέλα χωρίζονται σε στατικά και δυναμικά:

- **Τα στατικά μοντέλα** αναπαριστούν ένα στιγμιότυπο του συστήματος. Το σύστημα εξετάζεται εντός ενός χρονικού διαστήματος και οι παράγοντες που το επηρεάζουν παραμένουν σταθεροί. Ένα μοντέλο κατανομής πόρων, πχ μηχανημάτων, σε διάφορες εργασίες, χωρίς χρονικές μεταβολές είναι ένα στατικό μοντέλο.
- **Τα δυναμικά μοντέλα** αναπαριστούν χρονικά μεταβαλλόμενες καταστάσεις του συστήματος, είναι δηλαδή μοντέλα χρονικά εξαρτώμενα. Ένα μοντέλο προσλήψεων προσωπικού σε ένα πολυκατάστημα είναι δυναμικό, εάν προβλέπει χρονική διακύμανση της πελατείας του, πχ σε περιόδους εορτών, και άρα χρονικά μεταβαλλόμενες ανάγκες σε προσωπικό. Ο χρόνος μπορεί να θεωρηθεί διακριτός ή διαρκώς μεταβαλλόμενος. Στην πρώτη περίπτωση η κατάσταση του συστήματος μεταβάλλεται βηματικά, δηλαδή σε συγκεκριμένες χρονικές στιγμές, ενώ παραμένει σταθερή στα ενδιάμεσα διαστήματα. Στη δεύτερη περίπτωση, η κατάσταση του συστήματος είναι συνεχής συνάρτηση του χρόνου.

3.3 Βεβαιότητα, Αβεβαιότητα, Ρίσκο.

Η λήψη αποφάσεων γίνεται στη βάση προσδοκώμενων αποτελεσμάτων που θα προκύψουν στο μέλλον. Το ερώτημα είναι εάν μπορεί να γνωρίζει κανείς τι θα συμβεί στο μέλλον. Υπάρχουν δύο ακραίες εκδοχές. Στη μια περίπτωση, κάποιος γνωρίζει με απόλυτη σιγουριά το αποτέλεσμα των ενεργειών του. Η απόφαση τότε λαμβάνεται σε συνθήκες βεβαιότητας. Στην αντίθετη ακραία εκδοχή υπάρχει απόλυτη άγνοια σχετικά με το

αποτέλεσμα και η απόφαση λαμβάνεται σε συνθήκες αβεβαιότητας. Οι ενδιάμεσες καταστάσεις είναι εκείνες όπου τα αποτελέσματα είναι πιθανολογικά ενδεχόμενα. Στην περίπτωση αυτή, η απόφαση λαμβάνεται σε συνθήκες ρίσκου. Τα διοικητικά στελέχη αντιμετωπίζουν προβλήματα όλων αυτών των κατηγοριών. Όμως για ένα δεδομένο πρόβλημα, ο βαθμός της αβεβαιότητας μεταβάλλεται μεταξύ διαφορετικών στελεχών, αναλόγως με τη γνώση που έχει ο καθένας για το συγκεκριμένο πρόβλημα. Αναλυτικότερα οι τρεις περιπτώσεις έχουν ως εξής:

Βεβαιότητα. Λήψη αποφάσεων σε συνθήκες βεβαιότητας σημαίνει ότι είμαστε σίγουροι (ή υποθέτουμε ότι είμαστε σίγουροι) για το τι θα συμβεί στο μέλλον. Σε αυτήν την περίπτωση επιλέγουμε να πράξουμε αυτό που θα μας αποδώσει το μέγιστο όφελος. Ένα πολύ συνηθισμένο παράδειγμα απόφασης με βεβαιότητα, που θα το συναντήσει ο αναγνώστης σε πολλά συγγράμματα, είναι ότι η επένδυση σε κρατικά ομόλογα είναι μια απόφαση με βεβαιότητα, αφού ο επενδυτής γνωρίζει με σιγουριά το ύψος της απόδοσης και τον χρόνο αποπληρωμής (μερικές φορές η ζωή είναι πολύ σκληρή με τις θεωρίες).

Σε πραγματικές συνθήκες, είναι λίγο δύσκολο να γνωρίζει κανείς με απόλυτη βεβαιότητα τι θα συμβεί στο μέλλον. Εάν όμως *υποθέσει* ότι η έκβαση είναι βέβαιη, τότε απλοποιείται σημαντικά το μοντέλο. Βεβαίως, αυτό δεν σημαίνει ότι τα μοντέλα για λήψη αποφάσεων σε συνθήκες βεβαιότητας είναι πάντα απλά. Προβλήματα Γραμμικού Προγραμματισμού, που παρουσιάζονται αργότερα σε αυτό το κεφάλαιο, ανήκουν στην κατηγορία των προβλημάτων με βεβαιότητα, το μοντέλο όμως μπορεί να περιλαμβάνει χιλιάδες μεταβλητές και περιορισμούς.

Ρίσκο. Αποφάσεις λαμβάνονται σε συνθήκες ρίσκου όταν στο μέλλον μπορεί να συμβεί μια σειρά από ενδεχόμενα γεγονότα. Η υπόθεση σε αυτήν την περίπτωση είναι ότι γνωρίζουμε εκ των προτέρων (ή μπορούμε να υπολογίσουμε) την πιθανότητα εμφάνισης του κάθε γεγονότος. Κάθε απόφαση μας θα αποδώσει ένα όφελος ή μια ζημία, ανάλογα με το τι θα συμβεί στο μέλλον. Οι αποφάσεις που λαμβάνουν τα στελέχη των επιχειρήσεων ανήκουν κατά κανόνα σε αυτήν την κατηγορία. Ένα παράδειγμα απόφασης σε συνθήκες ρίσκου είναι η επιλογή για ριψοκίνδυνες ή συντηρητικές επενδύσεις. Εάν υπάρχει οικονομική άνθηση, τότε οι ριψοκίνδυνες επενδύσεις θα αποδώσουν πολύ περισσότερο από τις συντηρητικές. Εάν όμως υπάρχει ύφεση, οι ριψοκίνδυνες επενδύσεις θα προκαλέσουν ζημίες ενώ οι συντηρητικές ίσως αποφέρουν κάποιο μικρό κέρδος. Το αποτέλεσμα κάθε απόφασης εξαρτάται από τις μελλοντικές εξελίξεις.

Αβεβαιότητα. Όπως και στην περίπτωση του ρίσκου, στο μέλλον μπορεί να συμβούν μια σειρά από γεγονότα. Η διαφορά έγκειται στο ότι η πιθανότητα εμφάνισης του κάθε ενδεχόμενου γεγονότος δεν είναι γνωστή ούτε μπορεί να υπολογιστεί. Τα στελέχη προσπαθούν να αποφύγουν την αβεβαιότητα, αναζητώντας περισσότερη πληροφόρηση για το πρόβλημα, ώστε να το χειριστούν ως περίπτωση ρίσκου ή βεβαιότητας. Ωστόσο, αυτό δεν είναι πάντα εφικτό. Αν η απόφαση με αβεβαιότητα είναι αναπόφευκτη, τότε ο τρόπος λήψης της απόφασης εξαρτάται από την υποκειμενική στάση του στελέχους απέναντι στον κίνδυνο. Έχουν προταθεί διάφορες μέθοδοι:

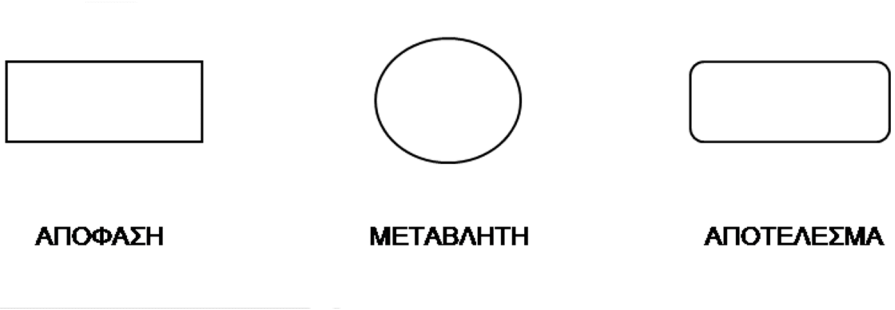
- Η αισιόδοξη στάση είναι να επιλέξει τη λύση που θα αποφέρει το μεγαλύτερο κέρδος σε περίπτωση θετικής έκβασης.
- Η απαισιόδοξη στάση είναι να επιλέξει τη λύση που θα αποφέρει το λιγότερο κακό αποτέλεσμα σε περίπτωση αρνητικής έκβασης.
- Σύμφωνα με το κριτήριο Hurwicz, ορίζεται συντελεστής βαρύτητας για την αισιοδοξία και την απαισιοδοξία και υπολογίζεται η καλύτερη λύση.
- Ένας άλλος τρόπος αντιμετώπισης είναι να θεωρήσει κανείς ότι όλα τα ενδεχόμενα γεγονότα του μέλλοντος έχουν ίση πιθανότητα εμφάνισης.
- Ένας πέμπτος τρόπος είναι να επιλέξει εκείνη τη λύση, η οποία προκαλεί τη μικρότερη διαφορά μεταξύ του καλύτερου και χειρότερου ενδεχόμενου (minimax regret).

3.4 Ανάλυση Αποφάσεων

Η Ανάλυση Αποφάσεων είναι μια μεθοδολογία που εφαρμόζεται για τη λήψη αποφάσεων σε προβλήματα με περιορισμένο αριθμό εναλλακτικών επιλογών. Περιλαμβάνει εργαλεία για την αξιολόγηση παραγόντων, που επηρεάζουν μια απόφαση και την εύρεση μιας συνιστώμενης λύσης. Δύο εργαλεία που χρησιμοποιούνται στην Ανάλυση Αποφάσεων είναι τα Διαγράμματα Επιρροής και οι Πίνακες Αποφάσεων ή Πίνακες Ανταμοιβών.

3.4.1 Διαγράμματα Επιρροής

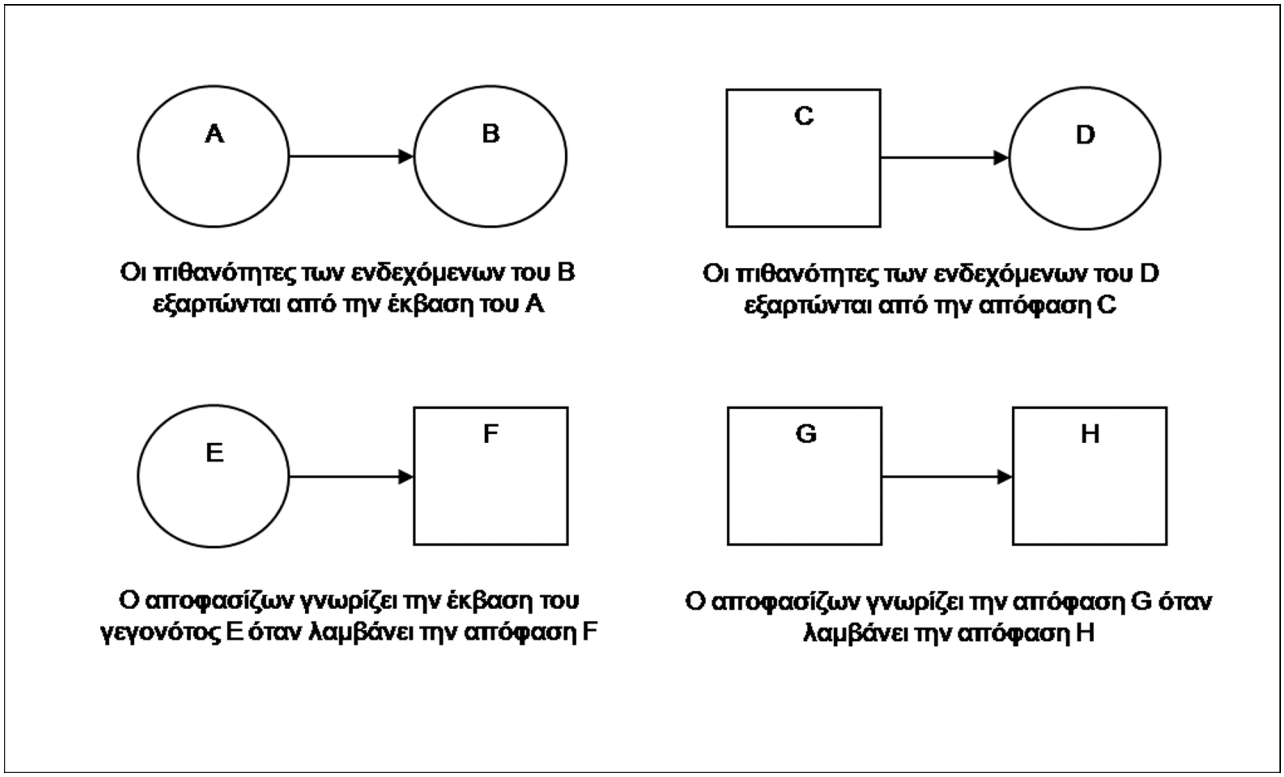
Τα Διαγράμματα Επιρροής (ΔΕ) (Influence Diagrams) είναι ένας γραφικός τρόπος αναπαράστασης ενός προβλήματος λήψης απόφασης. Ένα ΔΕ είναι ένας κατευθυνόμενος ακυκλικός γράφος, που αποτελείται από κόμβους και βέλη. Οι κόμβοι αναπαριστούν αποφάσεις, μεταβλητές (αβέβαια γεγονότα), ή αποτελέσματα (αξίες) και το σχήμα του κόμβου υποδηλώνει τον τύπο του, όπως φαίνεται και στο Σχήμα 3.1. Τα βέλη αναπαριστούν τις επιρροές μεταξύ των κόμβων. Με τον τρόπο αυτό, το ΔΕ αποτελεί μια πιθανολογική περιγραφή του προβλήματος.



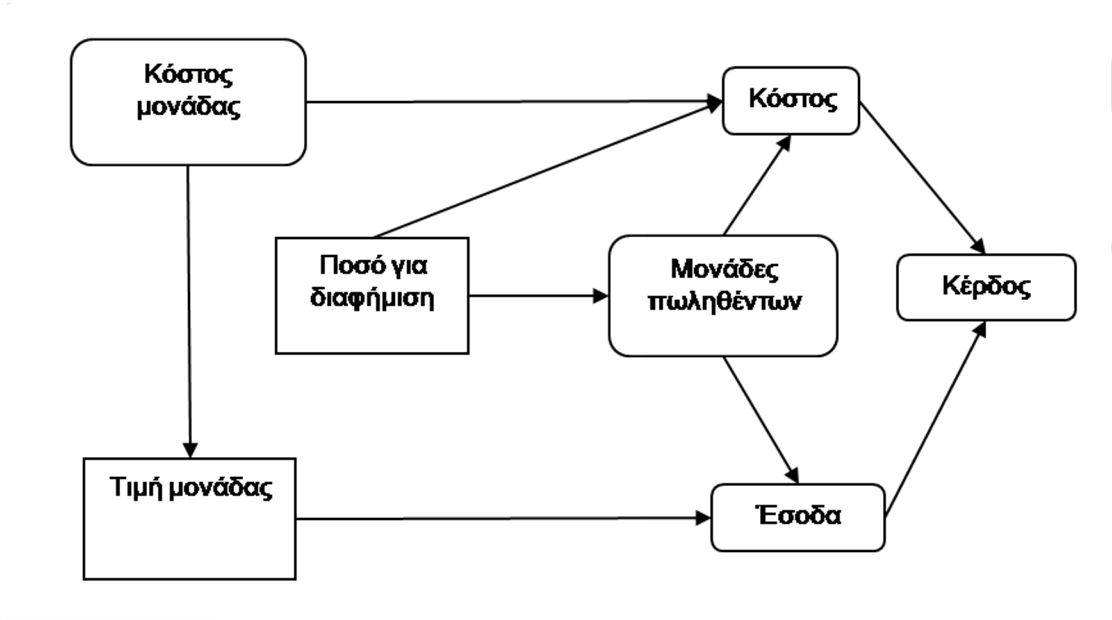
Σχήμα 3.1 *Κόμβοι Διαγράμματος Επιρροής*

Η ακριβής σημασία των βελών σχετίζεται με το είδος των κόμβων που συνδέουν, όπως φαίνεται και στο Σχήμα 3.2

Τα Διαγράμματα Επιρροής είναι ένα εξαιρετικό μέσο για τη λογική αναπαράσταση ενός μοντέλου, καθώς αποτυπώνει τις σχέσεις επιρροής μεταξύ των συστατικών μερών του μοντέλου. Είναι αρκετά ακριβές και περιγράφει τις πιθανολογικές εξαρτήσεις και τη ροή της πληροφορίας. Μπορεί να χρησιμοποιηθεί για τον σχεδιασμό του μοντέλου, αλλά και ως τρόπος συνεννόησης μεταξύ των αναλυτών, των στελεχών της επιχείρησης και όσων άλλων εμπλέκονται στην κατασκευή του μοντέλου. Στο Σχήμα 3.3 παρουσιάζεται ένα παράδειγμα Διαγράμματος Επιρροής. Η απόφαση για την τιμή της μονάδας ενός προϊόντος λαμβάνεται αφού γνωρίζουμε το κόστος μονάδας. Το ποσό για διαφήμιση, το πλήθος πωληθεισών μονάδων και το κόστος μονάδας επηρεάζουν το κόστος. Επίσης, το ποσό για διαφήμιση επηρεάζει το πλήθος των πωληθεισών μονάδων. Τα έσοδα επηρεάζονται από την τιμή μονάδας και το πλήθος πωληθεισών μονάδων.



Σχήμα 3.2 *Σχέσεις επιρροής σε Διαγράμματα Επιρροής*



Σχήμα 3.3 *Διάγραμμα Επιρροής.*

3.4.2 Πίνακες Αποφάσεων.

Οι Πίνακες Αποφάσεων είναι ένα εργαλείο στήριξης αποφάσεων, που οργανώνει και παρουσιάζει την πληροφορία με συστηματικό τρόπο. Η μέθοδος είναι κατάλληλη για προβλήματα αναζήτησης ενός στόχου ή κριτηρίου. Πρόκειται για έναν πίνακα, όπου οι στήλες αντιστοιχούν σε διάφορες εκδοχές μιας εξωτερικής κατάστασης και οι γραμμές σε διάφορες εναλλακτικές αποφάσεις. Κάθε εκδοχή της εξωτερικής κατάστασης έχει μια γνωστή πιθανότητα εμφάνισης. Όλες οι πιθανές εκδοχές εμφανίζονται ως στήλες στον πίνακα, και το άθροισμα των πιθανοτήτων τους ισούται με 1. Τα κελιά του πίνακα γεμίζουν με τιμές ανταμοιβών ή αποτελεσμάτων, που αντιστοιχούν στον συνδυασμό μιας κατάστασης με μια απόφαση.

Στο Σχήμα 3.4 παρουσιάζεται ένας Πίνακας Αποφάσεων. Υπάρχουν n δυνατές καταστάσεις K_1, K_2,…,K_n και m δυνατές εναλλακτικές E_1, E_2,…,E_m. Η πιθανότητα εμφάνισης της κατάστασης K_i είναι p_i και το άθροισμα των πιθανοτήτων ισούται με 1. Η ανταμοιβή για την απόφαση E_j, εάν έχει συμβεί η κατάσταση K_i, είναι A_{ji}.

	Καταστάσεις			
	K_1	K_2	…	K_n
Εναλλακτική	p1	p2	…	pn
E_1	A_{11}	A_{12}	…	A_{1n}
E_2	A_{21}	A_{22}	…	A2n
…	…	…	…	…
E_m	A_{m1}	A_{m2}	…	A_{mn}

Σχήμα 3.4 *Πίνακας Αποφάσεων.*

Για την εύρεση της προτεινόμενης απόφασης χρησιμοποιείται η **Αναμενόμενη Τιμή**. Η αναμενόμενη τιμή κάθε απόφασης ισούται με το άθροισμα των γινομένων της αμοιβής της εκάστοτε κατάστασης επί την πιθανότητα εμφάνισης της κατάστασης. Η Αναμενόμενη Τιμή υπολογίζεται σύμφωνα με την Εξίσωση 3.1

$$AT_j = \sum_{i=1}^{n} (piAji) \tag{3.1}$$

Ένας επενδυτής έχει τρεις δυνατές εναλλακτικές, να επενδύσει σε καταθέσεις με σταθερό επιτόκιο 3%, σε

καταθέσεις με μεταβαλλόμενο επιτόκιο ή να αγοράσει μετοχές. Η απόδοση της κατάθεσης με μεταβαλλόμενο επιτόκιο και της αγοράς μετοχών εξαρτώνται από την οικονομική κατάσταση. Οι ενδεχόμενες οικονομικές καταστάσεις είναι: Ανάπτυξη με πιθανότητα 35%, Στασιμότητα με πιθανότητα 40% και Ύφεση με πιθανότητα 25%. Οι αποδόσεις σε καταθέσεις με κυμαινόμενο επιτόκιο είναι 4%, 2% και 0% για την περίπτωση της Ανάπτυξης, της Στασιμότητας και της Ύφεσης αντίστοιχα. Οι αποδόσεις των μετοχών είναι 8%, 1% και -3%. Τα στοιχεία αυτά απεικονίζονται στο Σχήμα 3.5.

Η αναμενόμενη τιμή για το Σταθερό Επιτόκιο είναι (0,35*3)+(0,4*3)+(0,25*3)=3. Η αναμενόμενη τιμή για το Κυμαινόμενο Επιτόκιο είναι (0,35*4)+(0,4*1)+(0,25*0)=1,8. Η αναμενόμενη τιμή για τις Μετοχές είναι (0,35*8)+(0,4*2)+(0,25*-3)=2,85. Με βάση την Αναμενόμενη Τιμή η πλέον συμφέρουσα επένδυση είναι οι καταθέσεις σταθερού επιτοκίου.

	Καταστάσεις		
	Ανάπτυξη	Στασιμότητα	Ύφεση
Εναλλακτική	35	40	25
Σταθερό Επιτόκιο	3,0	3,0	3,0
Κυμαινόμενο Επιτόκιο	4,0	1,0	0,0
Μετοχές	8	2	-3

Σχήμα 3.5 *Αποδόσεις επενδύσεων σε διαφορετικές οικονομικές καταστάσεις*

3.5 Συστατικά μέρη Μαθηματικών Μοντέλων

Τα μοντέλα που χρησιμοποιούνται στην Αναλυτική των Επιχειρήσεων και στην Επιχειρηματική Ευφυΐα είναι κατά κανόνα μαθηματικά μοντέλα. Τα μαθηματικά μοντέλα είναι μη υλικές και συμβολικές αναπαραστάσεις πραγματικών συστημάτων. Χρησιμοποιώντας μαθηματικές έννοιες και σχέσεις, μοντελοποιούν ορισμένες από τις λειτουργίες του πραγματικού συστήματος. Έχουν κατασκευαστεί χιλιάδες μαθηματικά μοντέλα, από πολύ απλά μέχρι εξαιρετικά σύνθετα, τα οποία βρίσκουν εφαρμογή σε διάφορες επιστήμες, όπως στη Φυσική, στη Μηχανική, στην Ηλεκτρονική, στην Οικονομία κλπ. Ανεξαρτήτως του πεδίου εφαρμογής, του βαθμού πολυπλοκότητας ή του τύπου τους (πχ γραμμικά ή μη γραμμικά μοντέλα), τα μαθηματικά μοντέλα απαρτίζονται από μερικά βασικά συστατικά μέρη. Τα συστατικά μέρη των μαθηματικών μοντέλων είναι:

- οι μεταβλητές απόφασης,
- οι μη ελεγχόμενες μεταβλητές ή παράμετροι,
- οι μεταβλητές αποτελεσμάτων,
- οι μαθηματικές σχέσεις που συνδέουν τις παραπάνω μεταβλητές.

Το Σχήμα 3.6 παρουσιάζει ένα μαθηματικό μοντέλο και τα συστατικά του μέρη

Σχήμα 3.6 *Μαθηματικό μοντέλο*

Οι **Μεταβλητές Απόφασης** εκφράζουν εκείνα τα μεγέθη που ορίζονται με απόφαση του χρήστη του μοντέλου. Τα μεγέθη αυτά είναι ελεγχόμενα από τον χρήστη και η εκχώρηση τιμών σε αυτά αντιστοιχεί με τη λήψη διαφορετικών αποφάσεων. Ο χρήστης του μοντέλου πειραματίζεται με το σύστημα, εκχωρώντας τιμές στις μεταβλητές απόφασης και παρατηρώντας τα παραγόμενα αποτελέσματα στις μεταβλητές αποτελεσμάτων. Οι μεταβλητές απόφασης εκφράζουν τις δυνατότητες παρέμβασης του χρήστη στο πραγματικό σύστημα. Σε ένα πρόβλημα χαρτοφυλακίου επενδύσεων, το ποιές και πόσες μετοχές θα αγοραστούν είναι ένα παράδειγμα μεταβλητών απόφασης.

Οι **Μη Ελεγχόμενες Μεταβλητές ή Παράμετροι** εκφράζουν εκείνα τα μεγέθη των οποίων οι τιμές καθορίζονται από άλλους παράγοντες, ανεξέλεγκτους από τον χρήστη, ή είναι σταθερές. Κατά κανόνα, οι μεταβλητοί παράγοντες ανήκουν στο εξωτερικό περιβάλλον του συστήματος και επιδρούν σε αυτό. Οι σταθερές παράμετροι εκφράζουν μια σταθερή επιρροή στο σύστημα. Στο πρόβλημα του χαρτοφυλακίου επενδύσεων, μη ελεγχόμενες μεταβλητές είναι οι αποδόσεις των μετοχών (μεταβλητό μέγεθος) και το ποσοστό φορολόγησης (σταθερή παράμετρος). Και τα δύο επηρεάζουν το σύστημα με τρόπο ανεξάρτητο από τη βούληση του χρήστη.

Οι **Μεταβλητές Αποτελεσμάτων** καταγράφουν τα αποτελέσματα στην έξοδο του συστήματος. Οι τιμές τους καθορίζονται από τις τιμές των μεταβλητών απόφασης και των μη ελεγχόμενων μεταβλητών. Εκφράζουν τον βαθμό επιτυχίας ή αποτυχίας του συστήματος. Αποτελούν αντικείμενο παρατήρησης του χρήστη κατά τη διάρκεια των πειραμάτων και μέτρο επίτευξης των επιδιώξεων του. Η μεταβλητή που εκφράζει τα κέρδη ή τις ζημίες στο παράδειγμα του χαρτοφυλακίου, είναι μεταβλητή αποτελεσμάτων. Σε σύνθετα μαθηματικά μοντέλα, εκτός από την τελική μεταβλητή αποτελεσμάτων, είναι δυνατόν να υπάρχουν και ενδιάμεσες μεταβλητές αποτελεσμάτων, οι τιμές των οποίων υπολογίζονται από άλλες μεταβλητές.

Οι **Μαθηματικές Σχέσεις** συνδέουν τα υπόλοιπα συστατικά μέρη και καθορίζουν τον τρόπο με τον οποίο υπολογίζονται οι μεταβλητές αποτελεσμάτων από τις μεταβλητές απόφασης και τις μη ελεγχόμενες μεταβλητές. Οι μαθηματικές σχέσεις εκφράζουν τη λειτουργία του συστήματος.

3.6 Μαθηματική Βελτιστοποίηση και Γραμμικός Προγραμματισμός.

Στη Διοίκηση Επιχειρήσεων, η λήψη αποφάσεων απαιτεί την επιλογή μιας απόφασης μεταξύ άλλων εναλλακτικών. Για ένα μεγάλο ποσοστό τέτοιων προβλημάτων, λύση μπορούν να δώσουν οι μέθοδοι της Μαθηματικής Βελτιστοποίησης. Αντικείμενο της Μαθηματικής Βελτιστοποίησης είναι η εύρεση της μικρότερης ή μεγαλύτερης τιμής μιας συνάρτησης, με επιλογή κατάλληλων τιμών για τις μεταβλητές εισόδου. Ο Γραμμικός Προγραμματισμός είναι μια μέθοδος Μαθηματικής Βελτιστοποίησης. Έχει εφαρμοστεί επανειλημμένως για τη λύση πολλών προβλημάτων και αποτελεί ίσως τη δημοφιλέστερη μέθοδο στον χώρο της Επιχειρησιακής Έρευνας. Η ανάπτυξη του Γραμμικού Προγραμματισμού οφείλεται στη συμβολή του Kantorovich (1960), ο οποίος έθεσε τα θεμέλια του και κατάδειξε την πρακτική σημασία του, καθώς και του Dantzig (1951) ο οποίος ανακάλυψε τη μέθοδο simplex.

Ο Γραμμικός Προγραμματισμός (ΓΠ) επιλύει προβλήματα κατανομής πόρων με βέλτιστο τρόπο. Το σκεπτικό είναι ότι υπάρχει ένας περιορισμένος αριθμός πόρων (πχ κεφαλαίων, μηχανημάτων, εργατικού δυναμικού κλπ.). Οι πόροι διατίθενται και χρησιμοποιούνται για την επίτευξη ενός σκοπού. Το ζητούμενο είναι να κατανεμηθούν οι πόροι με τέτοιον τρόπο, ώστε να επιτευχθεί το βέλτιστο αποτέλεσμα. Συνήθως, το επιδιωκόμενο αποτέλεσμα είναι η μεγιστοποίηση του κέρδους ή η ελαχιστοποίηση του κόστους. Στον πραγματικό κόσμο, οι επιλογές δεν μπορεί να είναι αυθαίρετες. Μια σειρά από ιδιότητες, απαιτήσεις, κανονισμούς κλπ. επιβάλλουν περιορισμούς στη διάθεση των πόρων. Για παράδειγμα, οι πόροι είναι περιορισμένοι ή ορισμένες ποσότητες δεν μπορούν να πάρουν αρνητικές τιμές. Συνοψίζοντας λοιπόν το πρόβλημα, πρέπει να κατανεμηθούν πόροι έτσι ώστε να βελτιστοποιηθεί το αποτέλεσμα, και ταυτόχρονα η κατανομή των πόρων να γίνει με τρόπο που να μην παραβιάζει τους περιορισμούς.

Ένα παράδειγμα προβλήματος ΓΠ είναι το ακόλουθο. Μία εταιρεία παράγει δύο προϊόντα, το προϊόν Α και το προϊόν Β. Από την πώληση του προϊόντος Α η εταιρεία κερδίζει 150 ευρώ, ενώ από την πώληση του προϊόντος Β κερδίζει 100 ευρώ. Η εταιρεία διαθέτει 100 εργάσιμες ώρες ημερησίως και για την κατασκευή του προϊόντος Α χρειάζονται 2 ώρες, ενώ για την κατασκευή του προϊόντος Β χρειάζεται 1 ώρα. Λόγοι μάρκετινγκ απαιτούν η ποσότητα του προϊόντος Β που κατασκευάζεται να μην υπερβαίνει την ποσότητα του Α και η ποσότητα του προϊόντος Α πρέπει να υπερβαίνει τα 20 τεμάχια. Επίσης, ο μηχανολογικός εξοπλισμός δεν επιτρέπει την κατασκευή περισσότερων από 28 τεμάχια για το προϊόν Β. Το ζητούμενο είναι πόσες μονάδες πρέπει να κατασκευαστούν για το προϊόν Α και πόσες για το Β ώστε να μεγιστοποιηθεί το κέρδος.

Στο παραπάνω πρόβλημα πρέπει να ληφθεί απόφαση σχετικά με τις τιμές δύο μεταβλητών X και Y, όπου η X είναι η ποσότητα του προϊόντος Α που θα παραχθεί και Υ είναι η ποσότητα του προϊόντος Β. Αυτές οι δύο

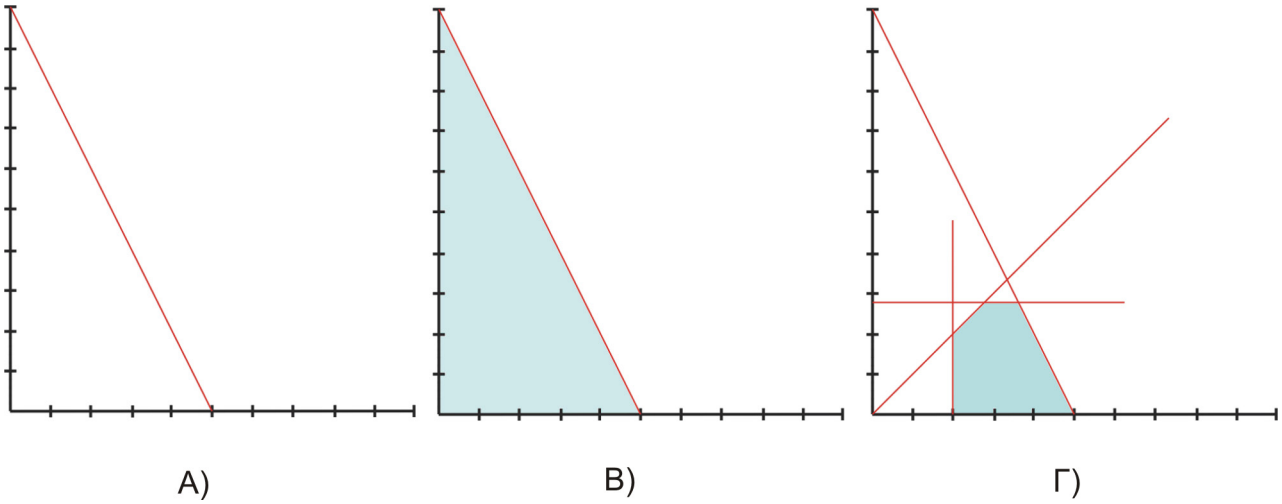
μεταβλητές είναι οι *μεταβλητές απόφασης*. Οι τιμές πρέπει να καθοριστούν έτσι ώστε να μεγιστοποιείται το κέρδος. Το κέρδος υπολογίζεται από τη συνάρτηση *Z = 150*X + 100*Y.*

Η συνάρτηση αυτή καλείται *αντικειμενική συνάρτηση* και είναι το μέγεθος που θα βελτιστοποιηθεί, δηλαδή θα μεγιστοποιηθεί ή θα ελαχιστοποιηθεί (στην περίπτωση του παραδείγματος πρέπει να μεγιστοποιηθεί). Ταυτόχρονα ισχύουν περιορισμοί. Ο χρόνος κατασκευής δεν μπορεί να υπερβαίνει τις 100 ώρες συνολικά. Μαθηματικά αυτό διατυπώνεται με την ανισότητα *2*X + 1*Y <= 100.*

Επίσης, το *Y* δεν μπορεί να υπερβαίνει το *X,* το *X* πρέπει να είναι μεγαλύτερο από 20 και τέλος το *Y* πρέπει να είναι μικρότερο ή ίσο από 28 και μεγαλύτερο από 0. Το πρόβλημα λοιπόν επαναδιατυπώνεται ως εξής:

Max[Z] = 150*X + 100*Y
υπό τους περιορισμούς
2X + 1Y <= 100
X>=Y
X >= 20
Y <= 28
Y >= 0

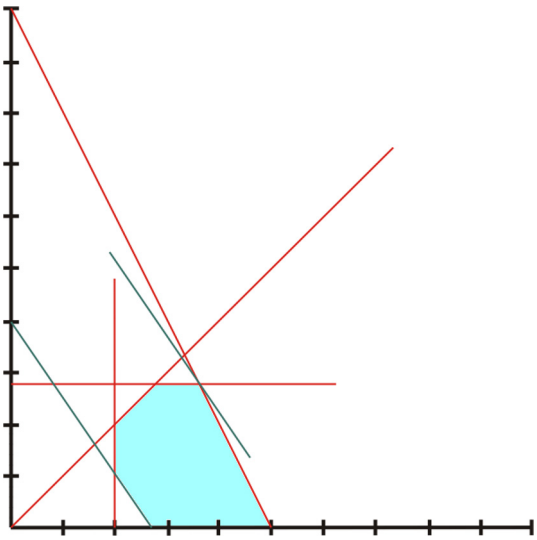
Για την καλύτερη κατανόηση της λύσης του προβλήματος, θα το αναπαραστήσουμε γραφικά. Στο Σχήμα 3.7.Α έχει χαραχθεί η καμπύλη *2X+Y=100.* Όλα τα σημεία αριστερά της καμπύλης ικανοποιούν την ανισότητα *2X+Y<=100.* Τα σημεία αυτά βρίσκονται στο σκιασμένο τμήμα του Σχήματος 3.7.Β. Εάν χαραχθούν και οι υπόλοιπες καμπύλες, *X=Y, X=20 και* Y=28, τότε τα σημεία που ικανοποιούν τους περιορισμούς βρίσκονται στο σκιασμένο τμήμα του Σχήματος 3.7.Γ. Ένα από τα σημεία των γωνιών του πολυγώνου είναι το βέλτιστο σημείο και μπορεί να βρεθεί με απλή σύγκριση.



Σχήμα 3.7 *Γραμμικός Προγραμματισμός Ι*

Ένας άλλος τρόπος για την εύρεση του βέλτιστου σημείου είναι να προστίθεται η καμπύλη της αντικειμενικής συνάρτησης για μια τυχαία τιμή του *Z*, πχ *Z=4000*. Ακολούθως, η καμπύλη αυτή μετατοπίζεται παράλληλα, μέχρι να συναντήσει την επάνω δεξιά γωνία του επιλεγμένου πολύεδρου. Οι τιμές *X* και *Y* του σημείου αυτού, δηλαδή *X=36* και *Y=28* είναι οι βέλτιστες τιμές για το πρόβλημα μας. Με αυτές τις τιμές *X* και *Y* το συνολικό κέρδος *Z* γίνεται *Z=8200*.

Το πρόβλημα φαίνεται σχετικά απλό γιατί αναζητά τιμές για δύο μόνο μεταβλητές απόφασης. Για πολλές μεταβλητές απόφασης το πρόβλημα γίνεται εξαιρετικά περίπλοκο. Η πιο διαδεδομένη μέθοδος για την επίλυση προβλημάτων ΓΠ είναι η μέθοδος simplex, η οποία ανακαλύφθηκε από τον G. Dantzig στις αρχές της δεκαετίας του 1940. Η μεγάλη της αξία συνίσταται στον πολύ αποτελεσματικό τρόπο αξιολόγησης των σημείων των γωνιών και εύρεσης του βέλτιστου σημείου. Ο αλγόριθμος ξεκινά από ένα γωνιακό σημείο, και στη συνέχεια αναζητά ένα παρακείμενο γωνιακό σημείο για το οποίο η τιμή της αντικειμενικής συνάρτησης είναι μεγαλύτερη. Η διαδικασία επαναλαμβάνεται μέχρις ότου δεν μπορεί να βρεθεί παρακείμενο σημείο, που να βελτιώνει το αποτέλεσμα της αντικειμενικής συνάρτησης.



Σχήμα 3.8 *Γραμμικός Προγραμματισμός ΙΙ*

Υπάρχουν τέσσερεις βασικές προϋποθέσεις που ισχύουν για τα προβλήματα Γραμμικού Προγραμματισμού:

- **Γραμμικότητα**. Η αντικειμενική συνάρτηση και οι περιορισμοί πρέπει να είναι γραμμικές συναρτήσεις. Αντικειμενικές συναρτήσεις όπως η $Z=AX^2+BY$ ή περιορισμοί όπως $X>Y^2$ δεν είναι γραμμικοί και δεν εμπίπτουν στην κατηγορία των προβλημάτων Γραμμικού Προγραμματισμού.
- **Βεβαιότητα**. Όλα τα στοιχεία του προβλήματος, όπως οι συντελεστές της αντικειμενικής συνάρτησης και οι περιορισμοί είναι εκ των προτέρων γνωστοί με απόλυτη βεβαιότητα.
- **Διαιρετότητα**. Όλες οι μεταβλητές απόφασης δεν είναι ακέραιοι αριθμοί, αλλά είναι πραγματικοί αριθμοί, οι οποίοι μπορούν να διαιρούνται επ' άπειρον.
- **Μονοδιάστατη**. Μπορεί να υπάρχει μόνο μια αντικειμενική συνάρτηση.

Ο Γραμμικός Προγραμματισμός έχει χρησιμοποιηθεί ευρύτατα στις επιχειρήσεις, για την επίλυση διαφόρων προβλημάτων. Μερικοί συνηθισμένοι τύποι προβλημάτων είναι οι ακόλουθοι:

- **Προβλήματα ανάμιξης συστατικών**. Στα προβλήματα αυτής της κατηγορίας πρέπει να επιτευχθεί η ιδανική σύνθεση με την ανάμιξη διαφόρων συστατικών. Για παράδειγμα, για τον προσδιορισμό μιας δίαιτας πρέπει να αναμιχθούν δύο είδη τροφής, κάθε μια από τις οποίες περιέχει ορισμένα θρεπτικά συστατικά. Η ανάμιξη των τροφών πρέπει να γίνει με τρόπο που να εξασφαλίζει την πρόσληψη των αναγκαίων θρεπτικών συστατικών, με ταυτόχρονη ελαχιστοποίηση του κόστους. Στην κατηγορία αυτή ανήκει και το κλασσικό πρόβλημα του επενδυτή, ο οποίος καλείται να επενδύσει ένα ποσό σε διάφορες επενδυτικές εναλλακτικές (μετοχές, ομόλογα κλπ.), επιτυγχάνοντας μέγιστη απόδοση και με τον περιορισμό του μεγέθους του ρίσκου που επιθυμεί να αναλάβει.
- **Πρόβλημα εκχώρησης**. Σε αυτήν την κατηγορία προβλημάτων υπάρχουν διάφοροι πόροι, όπως μηχανολογικός εξοπλισμός, εργατικό δυναμικό κλπ. στους οποίους πρέπει να ανατεθούν εργασίες με τέτοιον τρόπο ώστε να βελτιστοποιείται ένα μέγεθος, όπως πχ η μείωση του κόστους παραγωγής.
- **Προβλήματα μεταφορών**. Πρόκειται για πολύ συνηθισμένα και γνωστά προβλήματα ΓΠ. Σημειωτέον ότι ο Dantzig εργάστηκε για πολλά χρόνια στην πολεμική αεροπορία των ΗΠΑ, επιλύοντας τέτοια προβλήματα. Αντικείμενο των προβλημάτων αυτών είναι ελαχιστοποίηση του κόστους μεταφοράς εμπορευμάτων, συνήθως από τα εργοστάσια παραγωγής στις αποθήκες, με περιορισμούς που επιβάλλονται από τη δυνατότητα παραγωγής των εργοστασίων και τη δυνατότητα πώλησης από τις αποθήκες.

3.7 Αναλύσεις what – if και αναζήτησης στόχου

Τα στελέχη των επιχειρήσεων χρησιμοποιούν τα μοντέλα για να εκτελέσουν διάφορες αναλύσεις, να εξάγουν συμπεράσματα και να λάβουν αποφάσεις. Δύο από τους συνηθέστερους τύπους αναλύσεων που διεξάγουν είναι η ανάλυση what-if και η ανάλυση αναζήτησης στόχου.

Αναλύσεις what-if. Στη λήψη επιχειρηματικών αποφάσεων, ειδικά σε τακτικό και στρατηγικό επίπεδο, η ύπαρξη κάποιου βαθμού αβεβαιότητας είναι ο κανόνας. Συχνά τα στελέχη είναι υποχρεωμένα να κάνουν εκτιμήσεις και να προσπαθούν να προβλέψουν τις τιμές ορισμένων μεγεθών. Επίσης, συχνά θέλουν να πειραματιστούν με διάφορες εναλλακτικές αποφάσεις και να ελέγξουν τα αποτελέσματα τους. Για την αντιμετώπιση τέτοιων προβλημάτων, μπορούν να χρησιμοποιηθούν οι λεγόμενες αναλύσεις what-if. Η ανάλυση what-if εξετάζει τι θα συμβεί εάν μεταβληθούν κάποια ή κάποιες από τις μεταβλητές απόφασης, τις μη ελεγχόμενες μεταβλητές ή τις παραμέτρους ενός προβλήματος. Χρησιμοποιώντας ένα μοντέλο, τα στελέχη πειραματίζονται, μεταβάλλοντας κάποιες τιμές και ελέγχοντας τις επιπτώσεις που έχουν οι μεταβολές αυτές στη λύση του προβλήματος. Με τον τρόπο αυτό, μελετούν διαφορετικά σενάρια, διαφορετικές εναλλακτικές αποφάσεις και διαφορετικές συνθήκες του εξωτερικού περιβάλλοντος.

Παραδείγματα αναλύσεων what-if είναι τα παρακάτω:

- Τι θα συμβεί με τις δόσεις αποπληρωμής ενός δανείου, εάν τα επιτόκια δανεισμού μεταβληθούν κατά ορισμένα ποσοστά;
- Κατά πόσο θα αυξηθεί το μερίδιο της αγοράς που κατέχει η επιχείρηση, εάν το ποσό που διατίθεται για διαφήμιση μεταβληθεί κατά ορισμένα ποσοστά;

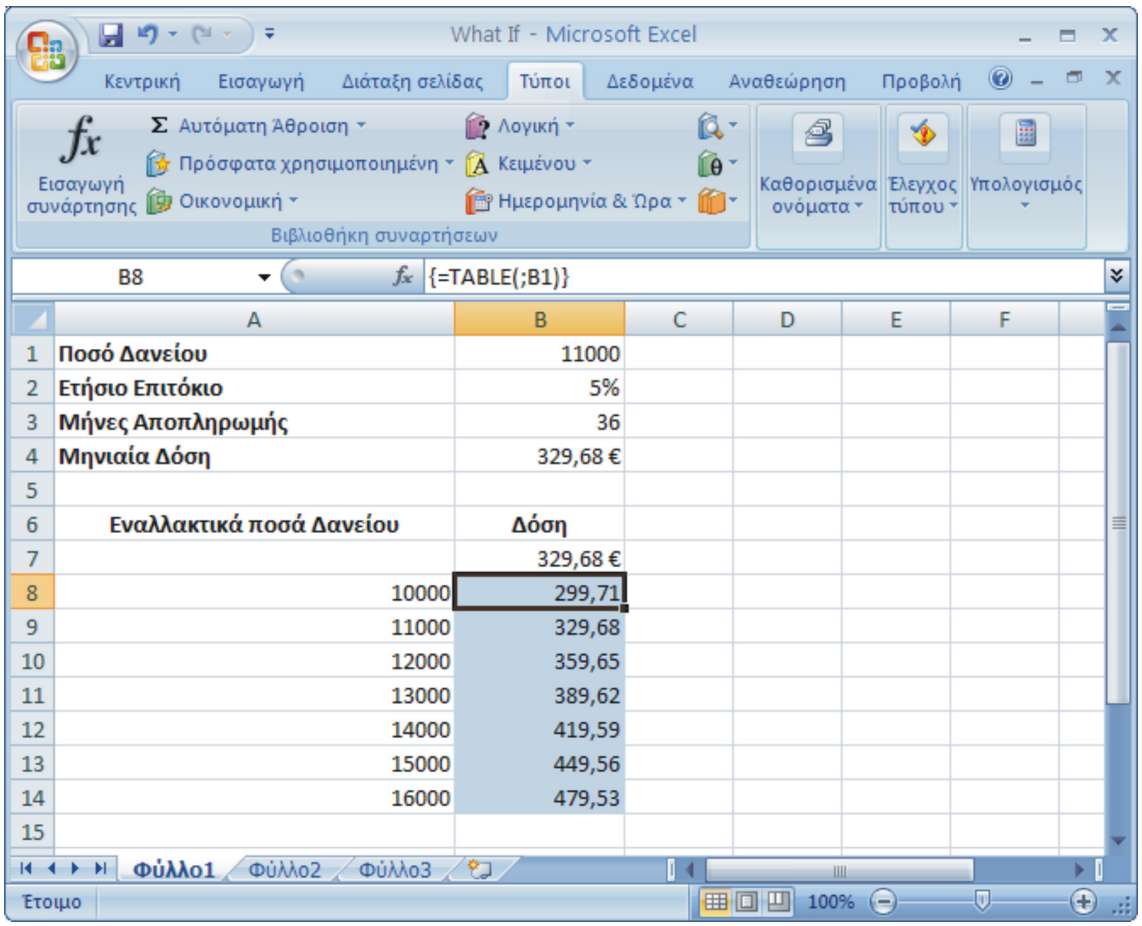
Αναλύσεις Αναζήτησης Στόχου. Με την ανάλυση what-if ο αναλυτής παρέχει τιμές σε μεταβλητές εισόδου του μοντέλου και λαμβάνει απαντήσεις σχετικά με τα αποτελέσματα. Πολλές φορές όμως, επιθυμεί να εκτελέσει την αντίστροφη διαδικασία, δηλαδή να ορίσει τα αποτελέσματα και να δει τι τιμές πρέπει να πάρουν οι μεταβλητές εισόδου για να επιτευχθούν αυτά τα αποτελέσματα. Η διαδικασία αυτή ονομάζεται ανάλυση Αναζήτησης Στόχου. Η ανάλυση Αναζήτησης στόχου είναι μια προς-τα-πίσω διαδικασία, όπου καθορίζονται οι τιμές των μεταβλητών αποτελεσμάτων και υπολογίζονται οι τιμές των μεταβλητών εισόδου.

Παραδείγματα αναλύσεων Αναζήτησης Στόχου είναι τα παρακάτω:

- Πόσο πρέπει να γίνει το επιτόκιο ή ο χρόνος αποπληρωμής ενός δανείου, ώστε η μηνιαία δόση να γίνει 2000 ευρώ;
- Πόσο πρέπει να αυξηθεί το ποσό που διατίθεται για διαφήμιση, ώστε το μερίδιο της αγοράς που κατέχει η επιχείρηση να αυξηθεί κατά 3%;

Τα Υπολογιστικά Φύλα (Spreadsheets) είναι το συνηθέστερο περιβάλλον για τη διεξαγωγή αναλύσεων what-if και αναλύσεων αναζήτησης στόχου. Το Excel διαθέτει έτοιμα εργαλεία για τη διεξαγωγή τέτοιου τύπου αναλύσεων. Ειδικότερα, για αναλύσεις what-if το Excel προσφέρει δύο εναλλακτικές. Η πρώτη εναλλακτική ονομάζεται «σενάρια». Ο χρήστης δημιουργεί το μοντέλο του με όλους του απαραίτητους υπολογισμούς. Αφού επιλέξει μια μεταβλητή εισόδου με την οποία θα πειραματιστεί, ο χρήστης χρησιμοποιώντας το κατάλληλο εργαλείο (tab Δεδομένα, επιλογή Ανάλυση πιθανοτήτων/Διαχείριση σεναρίων) δημιουργεί μια σειρά από «σενάρια», για κάθε ένα από τα οποία ορίζει εναλλακτικές τιμές για την ελεγχόμενη μεταβλητή. Στη συνέχεια, μπορεί, επιλέγοντας σενάριο, να δει την υπολογιζόμενη τιμή του μοντέλου για την τιμή της επιλεγμένης μεταβλητής εισόδου που αντιστοιχεί στο συγκεκριμένο σενάριο. Το Excel δίνει δυνατότητα ελέγχου για περισσότερες μεταβλητές.

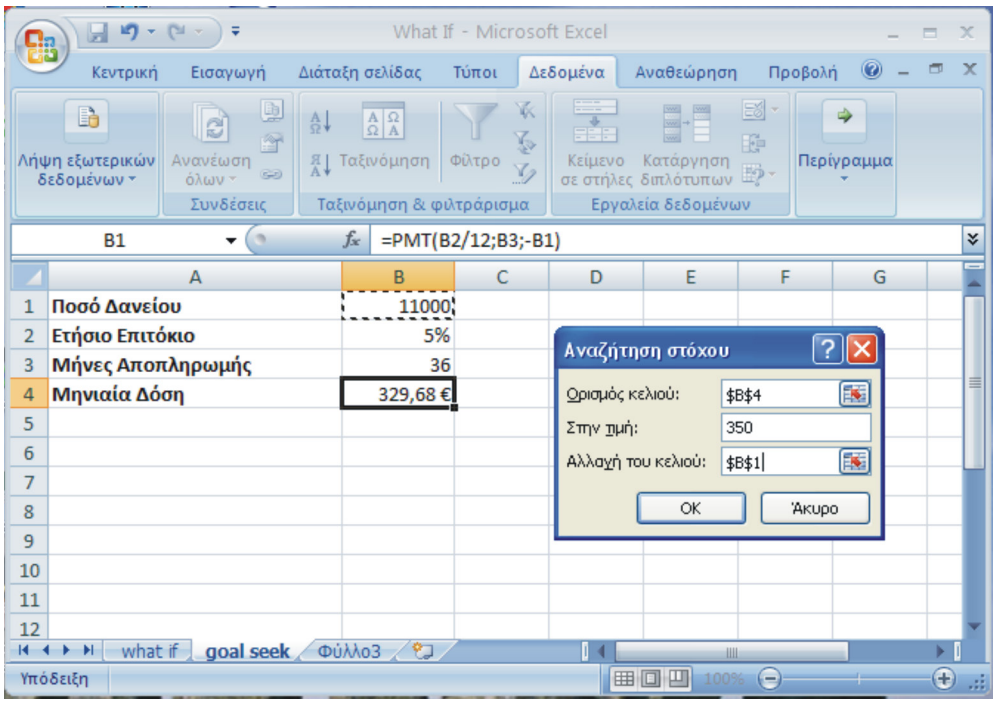
Ο δεύτερος τρόπος, για τη διεξαγωγή αναλύσεων what-if με το Excel, είναι με τη βοήθεια των πινάκων δεδομένων. Ο χρήστης μπορεί πάλι να ορίσει μια σειρά από δυνατές τιμές για κάποια μεταβλητή εισόδου και να δει τα αποτελέσματα που υπολογίζονται από το μοντέλο. Η διαφορά με τα «σενάρια» είναι ότι στα σενάρια παρουσιάζεται μόνο ένα αποτέλεσμα για μια τιμή εισόδου, ενώ με τον πίνακα δεδομένων παρουσιάζονται στην οθόνη ταυτόχρονα όλα τα αποτελέσματα. Στο Σχήμα 3.9 παρουσιάζεται παράδειγμα ανάλυσης what-if με τη βοήθεια πίνακα δεδομένων. Δανειολήπτης επιθυμεί να πάρει δάνειο 11000 ευρώ. Το ετήσιο επιτόκιο είναι 5% και σκοπεύει να αποπληρώσει το δάνειο σε χρονικό διάστημα 36 μηνών. Τα στοιχεία αυτά καταχωρούνται στα κελιά B1, B2 και B3 αντίστοιχα. Για τον υπολογισμό της μηνιαίας δόσης αποπληρωμής χρησιμοποιείται η έτοιμη συνάρτηση PMT, που περιλαμβάνεται στις οικονομικές συναρτήσεις του Excel. Το αποτέλεσμα της συνάρτησης PMT με τα προαναφερθέντα στοιχεία υπολογίζεται στο κελί B4. Επιπλέον, ο χρήστης επιθυμεί να ελέγξει ποια θα είναι η μηνιαία δόση για άλλα εναλλακτικά ποσά δανεισμού. Ο χρόνος αποπληρωμής και τα επιτόκιο παραμένουν σταθερά. Στα κελιά από A8 έως A14 τοποθετούνται εναλλακτικά ποσά δανεισμού, που κυμαίνονται από 10000 ευρώ έως 16000 ευρώ. Στα κελιά από B8 έως B14 υπολογίζονται οι αντίστοιχες μηνιαίες δόσεις. Με τη βοήθεια των πινάκων δεδομένων ο χρήστης μπορεί να πειραματιστεί με μέχρι δύο μεταβλητές. Σημειώνεται ότι τα χειριστικά ζητήματα του Excel βρίσκονται έξω από τα όρια του παρόντος συγγράμματος. Ο ενδιαφερόμενος αναγνώστης μπορεί να τα αναζητήσει σε βιβλία σχετικά με το Microsoft Office.



	A	B
1	**Ποσό Δανείου**	11000
2	**Ετήσιο Επιτόκιο**	5%
3	**Μήνες Αποπληρωμής**	36
4	**Μηνιαία Δόση**	329,68 €
5		
6	**Εναλλακτικά ποσά Δανείου**	**Δόση**
7		329,68 €
8	10000	299,71
9	11000	329,68
10	12000	359,65
11	13000	389,62
12	14000	419,59
13	15000	449,56
14	16000	479,53
15		

Σχήμα 3.9 *Ανάλυση what-if στο Excel με χρήση πίνακα δεδομένων.*

Για τη διεξαγωγή αναλύσεων Αναζήτησης Στόχου, το Excel περιλαμβάνει άλλο εργαλείο (tab Δεδομένα, επιλογή Ανάλυση πιθανοτήτων/αναζήτηση στόχου). Ο χρήστης ορίζει το κελί – στόχο, δηλαδή το κελί στο οποίο υπολογίζονται τα αποτελέσματα, ορίζει την επιθυμητή τιμή στόχο και τέλος, ορίζει τη μεταβλητή εισόδου η οποία πρέπει να τροποποιηθεί κατάλληλα. Στο Σχήμα 3.10 παρουσιάζεται παράδειγμα αναζήτησης στόχου στο Excel. Ο χρήστης επιθυμεί να πάρει δάνειο σύμφωνα με τα στοιχεία του προηγούμενου παραδείγματος. Η υπολογιζόμενη δόση είναι 329,68 ευρώ. Ο χρήστης θέλει να μάθει τι ποσό μπορεί να πάρει αν αυξήσει τη δόση στα 350 ευρώ, διατηρώντας σταθερό το επιτόκιο και τον χρόνο αποπληρωμής. Ορίζει στο πεδίο «Ορισμός κελιού» το κελί στόχο (B4, δηλαδή η δόση αποπληρωμής) και στο πεδίο «Αλλαγή του κελιού» τη μεταβλητή εισόδου που θα τροποποιηθεί (B1, δηλαδή το ποσό δανεισμού). Πατώντας το πλήκτρο «OK» θα υπολογιστεί στο κελί B1 το νέο ποσό που μπορεί να δανειστεί ο χρήστης, αν αυξήσει τη δόση στα 350 ευρώ. Το νέο ποσό δανεισμού είναι 11678 ευρώ.



Σχήμα 3.10 *Αναζήτηση στόχου στο Excel*

Διατίθενται μια σειρά από Πρόσθετα Προγράμματα (add – ins) για το Excel, όπως το λογισμικό Solver, τα οποία επεκτείνουν κατά πολύ τις δυνατότητες διεξαγωγής αναλύσεων.

3.8 Ανάλυση Ευαισθησίας

Έχει τονιστεί επανειλημμένως ότι η λήψη επιχειρηματικών αποφάσεων γίνεται κατά κανόνα με κάποιο βαθμό αβεβαιότητας. Τα στελέχη είναι υποχρεωμένα να εκτιμούν και να προβλέπουν τις μελλοντικές τιμές ορισμένων μεγεθών. Τα μεγέθη αυτά, υπό μορφή μη ελεγχόμενων μεταβλητών, αποτελούν μέρος των μοντέλων και συνεισφέρουν στον υπολογισμό του τελικού αποτελέσματος. Όταν κάποιος χρησιμοποιεί ένα μοντέλο, πειραματίζεται με τις μεταβλητές απόφασης και υπολογίζει αποτελέσματα, θεωρώντας ότι οι τιμές των μη ελεγχόμενων μεταβλητών και των παραμέτρων παραμένουν σταθερές. Τίθεται επομένως το ερώτημα, κατά πόσο τα αποτελέσματα του μοντέλου είναι ανθεκτικά σε διακυμάνσεις των τιμών των μη ελεγχόμενων μεταβλητών και των παραμέτρων. Ένα άλλο ερώτημα είναι, πόσο επηρεάζουν το αποτέλεσμα μικρές μεταβολές στις μεταβλητές απόφασης. Απάντηση σε αυτά τα ερωτήματα δίνει η Ανάλυση Ευαισθησίας. Η Ανάλυση Ευαισθησίας μελετά το κατά πόσο οι μεταβολές στις τιμές των μεταβλητών εισόδου και των παραμέτρων επηρεάζουν το τελικό αποτέλεσμα. Μοντέλα ευαίσθητα σε εξωτερικές συνθήκες εμπνέουν περιορισμένη εμπιστοσύνη, καθώς τα αποτελέσματα τους μπορεί να είναι άκυρα, αν οι εξωτερικές συνθήκες μεταβληθούν. Για τον λόγο αυτό, οι καλές πρακτικές σχεδιασμού μοντέλων επιβάλλουν τον έλεγχο της ευαισθησίας των μοντέλων. Ο σχεδιαστής πρέπει να ορίσει πόση είναι η αβεβαιότητα στις μεταβλητές εισόδου, και πόσο συνεισφέρει η κάθε μεταβλητή εισόδου στην αβεβαιότητα της εξόδου.

Η Ανάλυση Ευαισθησίας είναι πολλαπλά χρήσιμη:

- Επιτρέπει τον έλεγχο της ευρωστίας των αποτελεσμάτων του μοντέλου ή και του πραγματικού συστήματος σε μεταβολές του εξωτερικού κόσμου.
- Προσφέρει εκτίμηση της σημαντικότητας της κάθε μεταβλητής εισόδου για τον καθορισμό των αποτελεσμάτων. Η πληροφορία αυτή είναι ιδιαίτερα χρήσιμη γιατί επιτρέπει:
  - Την απλοποίηση του μοντέλου με την απομάκρυνση μη σημαντικών μεταβλητών.
  - Τη μείωση του βαθμού αβεβαιότητας των αποτελεσμάτων. Ο χρήστης, γνωρίζοντας ποιες είναι οι πιο σημαντικές μεταβλητές, μπορεί να αναζητήσει πρόσθετη πληροφόρηση για αυτές, να μειώσει τον βαθμό αβεβαιότητας τους και έτσι να μειώσει σημαντικά και τον συνολικό βαθμό αβεβαιότητας του μοντέλου.
- Προσφέρει καλύτερη κατανόηση της συμπεριφοράς του μοντέλου και του πραγματικού συστήματος.
- Βοηθά στη λήψη πιο ευέλικτων αποφάσεων. Ο χρήστης, έχοντας υπόψη του τις επιπτώσεις των μεταβολών του πραγματικού κόσμου, λαμβάνει τις κατάλληλες αποφάσεις, οι οποίες προβλέπουν εκ των προτέρων τις πιθανές αντιδράσεις σε περίπτωση μεταβολών των συνθηκών.

- Αυξάνει την κατανόηση των σχέσεων μεταξύ των μεταβλητών εισόδου και εξόδου.
- Βοηθά στην οικοδόμηση πιο αξιόπιστων μοντέλων.
- Βοηθά στον εντοπισμό λαθών. Αναπάντεχα αποτελέσματα στην έξοδο μπορεί να σηματοδοτούν την ύπαρξη λάθους στο μοντέλο.

Έχουν προταθεί διάφορες μέθοδοι για τη διεξαγωγή Αναλύσεων Ευαισθησίας. Η απλούστερη εκδοχή είναι να δοκιμάζει ο χρήστης διάφορες τιμές για τις μεταβλητές εισόδου και να παρατηρεί τα αποτελέσματα. Κάθε φορά μπορεί να αλλάζει τις τιμές μιας μόνο μεταβλητής. Η προσέγγιση αυτή δεν θα αποδώσει αντικειμενικά αποτελέσματα, εάν υπάρχει σχέση αλληλεξάρτησης μεταξύ δύο ή περισσότερων μεταβλητών εισόδου. Εναλλακτικά, ο χρήστης μπορεί να αλλάζει τις τιμές από περισσότερες μεταβλητές, Άλλες, πιο εξελιγμένες μέθοδοι προβλέπουν τον καθορισμό συναρτήσεων πυκνότητας πιθανοτήτων για κάθε μεταβλητή εισόδου, και τη δημιουργία ενός πίνακα τυχαίων τιμών για τις μεταβλητές εισόδου και αντίστοιχου πίνακα με τα αποτελέσματα στην έξοδο. Ο αναγνώστης μπορεί να βρει περισσότερες πληροφορίες σχετικά με τις μεθόδους διεξαγωγής Αναλύσεων Ευαισθησίας στην εργασία του Hamby (1994).

3.9 Ευρετικές Μέθοδοι - Γενετικοί Αλγόριθμοι

Για την αναζήτηση λύσεων σε ένα πρόβλημα, μια καλή και αρκετά διαδεδομένη επιλογή είναι η χρήση των λεγόμενων Ευρετικών Μεθόδων (ΕΜ) (Heuristics). Η λέξη «heuristics» προέρχεται από την ελληνική λέξη «εύρεσις». Ο όρος ευρετικές μέθοδοι αναφέρεται σε στρατηγικές ευκολότερης ανακάλυψης κανόνων, οι οποίοι βοηθούν στην επίλυση ενός προβλήματος. Κατά καιρούς έχουν προταθεί διάφοροι ορισμοί για τις ευρετικές μεθόδους. Οι Newell and Simon (1972), οι οποίοι εισήγαγαν τον όρο, τον χρησιμοποιούν για να περιγράψουν απλές διαδικασίες, που αντικαθιστούν περίπλοκους αλγορίθμους. Οι Kahneman and Frederick (2002) ορίζουν ότι οι ΕΜ αξιολογούν ένα χαρακτηριστικό, χρησιμοποιώντας ένα άλλο ευκολότερο χαρακτηριστικό. Οι Shah and Oppenheimer (2008) αναφέρουν ότι η λέξη «heuristic» έχει χάσει τη σημασία της, και επικεντρώνουν στη μείωση της προσπάθειας μέσω της εξέτασης λιγότερων στοιχείων, της μείωσης της προσπάθειας για συλλογή τιμών, της απλοποίησης της στάθμισης των στοιχείων, της ολοκλήρωσης λιγότερων πληροφοριών και της εξέτασης λιγότερων εναλλακτικών. Οι Gigerenzer and Gaissmaier (2011) ορίζουν τις ΕΜ ως μια στρατηγική, που αγνοεί μέρος της πληροφορίας για να επιτύχει τη λήψη αποφάσεων με μεγαλύτερη ταχύτητα, φειδώ η/και ακρίβεια, από ότι άλλες, περισσότερο σύνθετες μέθοδοι. Ο ορισμός αυτός καινοτομεί, εισάγοντας το στοιχείο της αυξημένης ακρίβειας. Στο σημείο αυτό αντιδιαστέλλεται με την κλασσική αντίληψη περί ΕΜ, που θεωρεί ότι υπάρχει ένα ισοζύγιο μεταξύ ακρίβειας και ευκολίας, στο οποίο θυσιάζεται ένα αποδεκτό μέρος της ακρίβειας, προς όφελος της ταχύτητας και της ευκολίας.

Για την καλύτερη κατανόηση των ΕΜ, οι Gigerenzer and Gaissmaier (2011) αναφέρουν ένα απλό αλλά πολύ χαρακτηριστικό παράδειγμα. Οι επιχειρήσεις ενδιαφέρονται να εντοπίσουν τους λεγόμενους «ενεργούς πελάτες» τους, εκείνους τους πελάτες δηλαδή οι οποίοι είναι πιθανόν να πραγματοποιήσουν νέες αγορές στο αμέσως επόμενο χρονικό διάστημα. Για τον εντοπισμό αυτών των πελατών μπορεί να χρησιμοποιηθούν περίτεχνες στατιστικές μέθοδοι, όπως η παλινδρόμηση. Ωστόσο, οι μάνατζερ χρησιμοποιούν ένα πολύ πιο απλό κανόνα και θεωρούν ότι εάν ο πελάτης πραγματοποίησε αγορά το αμέσως προηγούμενο χρονικό διάστημα, τότε θεωρείται ενεργός, διαφορετικά θεωρείται ανενεργός. Αναφέρουν μάλιστα περιπτώσεις όπου ο απλός ορισμός των μάνατζερ αποδείχθηκε πιο ακριβής από τα στατιστικά μοντέλα.

Η ιδέα της αυξημένης ακρίβειας είναι καινοτόμα και παραμένει προς διερεύνηση. Σύμφωνα πάντως με την ευρύτερα αποδεκτή άποψη, οι ΕΜ είναι ένα τρόπος ταχύτερης και οικονομικότερης επίλυσης προβλημάτων, ο οποίος αποδίδει «αρκετά καλές» λύσεις. Οι ΕΜ χρησιμοποιούν προηγούμενη γνώση. Είναι μια επαναληπτική διαδικασία, που περιλαμβάνει την αναζήτηση, την αξιολόγηση και τη μάθηση. Η γνώση που αποκτήθηκε θα χρησιμοποιηθεί στην επόμενη επανάληψη. Στο τέλος της διαδικασίας θα έχει βρεθεί μια λύση, για την οποία δεν υπάρχει κάποια εγγύηση ότι είναι η καλύτερη δυνατή, θα είναι όμως μια λύση αρκετά καλή. Στο παράδειγμα με τους ενεργούς πελάτες, εάν αναπτυχθεί ένα μοντέλο παλινδρόμησης, τότε θα παραχθεί μια λύση που θα είναι η καλύτερη δυνατή, τουλάχιστον σύμφωνα με τη μέθοδο της παλινδρόμησης. Στις ΕΜ δεν υπάρχει ένα παρόμοιο σκεπτικό, που να τεκμηριώνει με κάποιο τρόπο ότι η λύση που βρέθηκε είναι η καλύτερη δυνατή.

Οι ΕΜ ενδείκνυνται για την αντιμετώπιση ασθενώς δομημένων προβλημάτων, για τα οποία δεν μπορεί να αναπτυχθεί μια αλγοριθμική λύση. Επίσης, μπορούν να χρησιμοποιηθούν σε περιπτώσεις εξαιρετικά περίπλοκων προβλημάτων, για τα οποία άλλου τύπου λύσεις θα απαιτούσαν απαγορευτικά πολύ χρόνο ή πολλούς υπολογιστικούς πόρους. Σύμφωνα με την έννοια της *οριοθετημένης λογικής (bounded rationality)*, που εισήγαγε ο Simon και που αναλύεται στο Κεφάλαιο 2, οι αποφάσεις στον πραγματικό κόσμο λαμβάνονται υπό

περιορισμούς, τους οποίους επιβάλλουν τόσο τα όρια των ανθρώπινων γνωστικών ικανοτήτων όσο και τα όρια που θέτει το περιβάλλον. Ειδικά για την περίπτωση των επιχειρηματικών αποφάσεων, όπου τα στελέχη τελούν υπό τη συνεχή πίεση του χρόνου, η βασική ιδιότητα των ΕΜ να επιλύουν προβλήματα εξοικονομώντας προσπάθεια και χρόνο, τις καθιστά πολύ δελεαστικές. Σε πραγματικές συνθήκες, το κόστος που απαιτείται για την ανακάλυψη της καλύτερης δυνατής λύσης πιθανώς να υπερβαίνει το κέρδος που αποδίδει η μέγιστη ακρίβεια. Οι ΕΜ σε αρκετές περιπτώσεις μπορούν να αποδώσουν πολλαπλές αποδεκτές λύσεις. Βασικό μειονέκτημα τους είναι ότι δεν αποδίδουν τη βέλτιστη λύση. Επίσης, ισχύουν για συγκεκριμένες περιπτώσεις και δεν έχουν τη γενικότερη ισχύ των αλγορίθμων. Ένα πολύ συνηθισμένο πεδίο εφαρμογής των ΕΜ είναι τα λογισμικά αντιμετώπισης ιών των υπολογιστών (antivirus). Τα λογισμικά αυτά ελέγχουν τη δομή και τη λογική των προγραμμάτων, τις εντολές προς τον υπολογιστή, δεδομένα που υπάρχουν στο εκτελέσιμο αρχείο και άλλα στοιχεία, και στη συνέχεια αξιολογούν την πιθανότητα να έχουν προσβληθεί τα εκτελέσιμα αρχεία από ιούς.

3.9.1 Γενετικοί Αλγόριθμοι

Οι Γενετικοί Αλγόριθμοι (ΓΑ) είναι μια από τις πιο γνωστές περιπτώσεις Ευρετικών Μεθόδων. Ανήκουν στην κατηγορία των Εξελικτικών Αλγορίθμων, οι οποίοι μιμούνται διαδικασίες της φυσικής εξέλιξης. Ειδικότερα, οι Γενετικοί Αλγόριθμοι είναι εμπνευσμένοι από τη θεωρία της Εξέλιξης του Δαρβίνου. Σύμφωνα με τη Βιολογία, κάθε οργανισμός διαθέτει χαρακτηριστικά, τα οποία είναι αποθηκευμένα ως γενετική πληροφορία στα γονίδια. Τα γονίδια συνδέονται μεταξύ τους σε μακριές σειρές και δημιουργούν τα χρωμοσώματα. Όταν δύο οργανισμοί ζευγαρώνουν και αναπαράγονται, ο απόγονος κληρονομεί ορισμένα γονίδια από τον ένα γονέα και τα υπόλοιπα από τον άλλο γονέα. Σε σπάνιες περιπτώσεις ένα γονίδιο μπορεί να μεταβληθεί μόνο του και η περίπτωση αυτή ονομάζεται μετάλλαξη. Εάν ο συνδυασμός χαρακτηριστικών του απογόνου είναι επιτυχημένος, εάν δηλαδή ο οργανισμός είναι καλύτερα προσαρμοσμένος στο περιβάλλον, τότε έχει περισσότερες πιθανότητες να επιβιώσει και να αναπαραχθεί, κληροδοτώντας τα χαρακτηριστικά του στους απογόνους του. Αντιθέτως, εάν ο συνδυασμός χαρακτηριστικών δεν είναι επιτυχημένος, τότε έχει λιγότερες πιθανότητες να επιβιώσει και να αναπαραχθεί. Η διαδικασία αυτή ονομάζεται φυσική επιλογή και αποτελεί τον μηχανισμό διαρκούς εξέλιξης και βελτίωσης των ειδών.

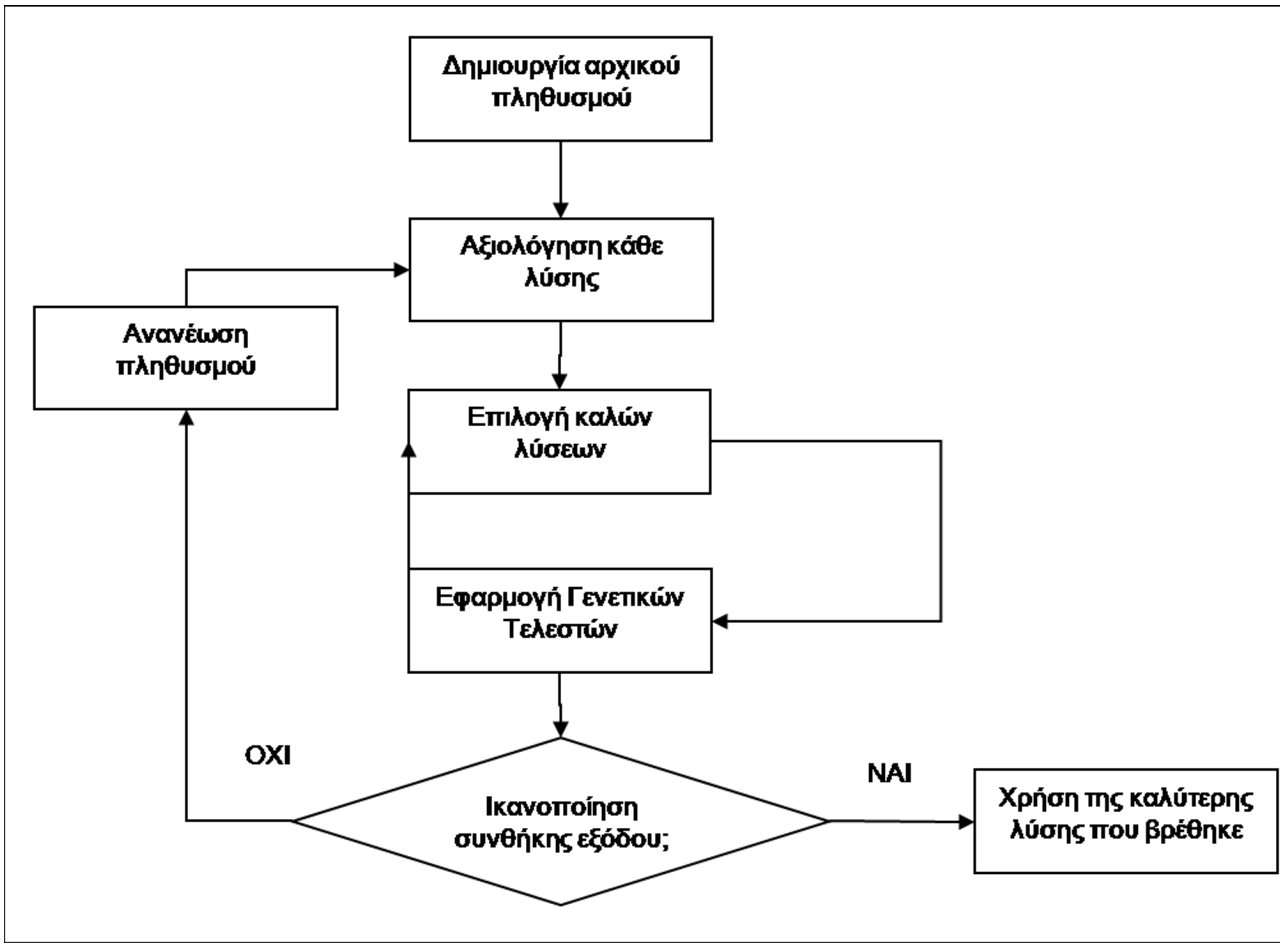
Οι Γενετικοί Αλγόριθμοι μιμούνται τη διαδικασία της φυσικής επιλογής. Αρχικά κωδικοποιούνται οι λύσεις ενός προβλήματος. Ο συνηθέστερος τρόπος κωδικοποίησης είναι μια ακολουθία από bits. Για παράδειγμα, ένας κανόνας χορήγησης δανείων που εφαρμόζει μια τράπεζα δηλώνει ότι «ΕΑΝ η ηλικία είναι μικρότερη ή ίση των 40 ΚΑΙ το εισόδημα είναι έως και μέτριο ΤΟΤΕ το δάνειο εγκρίνεται». Ο κανόνας αυτός μπορεί να κωδικοποιηθεί ως εξής: Η συνθήκη «η ηλικία είναι μικρότερη των 40» κωδικοποιείται με ένα bit που έχει την τιμή 0. Η συνθήκη «το εισόδημα είναι έως και μέτριο» κωδικοποιείται με την τιμή 1. Η απόφαση «το δάνειο εγκρίνεται» κωδικοποιείται με την τιμή 1. Με τον τρόπο αυτό ο συνολικός κανόνας κωδικοποιείται ως 011. Αντίστοιχα, ο κανόνας «ΕΑΝ η ηλικία είναι μεγαλύτερη των 40 ΚΑΙ το εισόδημα είναι έως και μέτριο ΤΟΤΕ το δάνειο δεν εγκρίνεται» θα κωδικοποιηθεί ως 110. Οι ακολουθίες των bit αποκαλούνται Χρωμοσώματα, ενώ τα τμήματα τους που κωδικοποιούν ένα χαρακτηριστικό ονομάζονται Γονίδια. Οι συνήθεις κωδικοποιήσεις λύσεων πραγματικών προβλημάτων περιέχουν αρκετά περισσότερα bits. Ένα τυπικό χρωμόσωμα μπορεί να έχει τη μορφή «10001110111001110001111101011».

Για την εύρεση λύσεων στο πρόβλημα εφαρμόζεται ο παρακάτω αλγόριθμος:

- Αρχικά δημιουργείται ένας μεγάλος πληθυσμός από τυχαία χρωμοσώματα., δηλαδή χρωμοσώματα με τυχαίες τιμές 0 και 1. Κάθε ένα από αυτά τα χρωμοσώματα, εάν αποκωδικοποιηθεί, συμβολίζει μια ενδεχόμενη λύση του προβλήματος.
- Στη συνέχεια, κάθε χρωμόσωμα ελέγχεται ως προς το εάν αποτελεί μια καλή λύση για το πρόβλημα. Για τον έλεγχο αυτό χρησιμοποιείται η λεγόμενη Συνάρτηση Καταλληλότητας (Fitness Function). Η Συνάρτηση Καταλληλότητας υπολογίζει για κάθε χρωμόσωμα μια τιμή καταλληλότητας.
- Ακολούθως, τα χρωμοσώματα που έχουν υψηλή τιμή καταλληλότητας επιλέγονται για αναπαραγωγή. Η πιθανότητα επιλογής είναι ανάλογη με την τιμή καταλληλότητας.
- Τα χρωμοσώματα που έχουν επιλεγεί για αναπαραγωγή διασταυρώνονται, δηλαδή οργανώνονται σε ζευγάρια και τα ζευγάρια ανταλλάσσουν μεταξύ τους γονίδια δημιουργώντας νέους απογόνους. Ο ρυθμός αναπαραγωγής είναι μια παράμετρος που ορίζεται από τον χρήστη.
- Σε σπανιότερες περιπτώσεις ένα από τα bits ενός χρωμοσώματος αλλάζει τιμή και μετατρέπεται από 0 σε 1 ή αντιστρόφως. Το φαινόμενο αυτό ονομάζεται μετάλλαξη και η συχνότητα του ορίζεται από τον ρυθμό μετάλλαξης. Η Διασταύρωση και η Μετάλλαξη ονομάζονται γενετικοί τελεστές.

- Ο πληθυσμός ανανεώνεται. Προστίθενται οι απόγονοι και απομακρύνονται τα χρωμοσώματα με χαμηλή τιμή καταλληλότητας.
- Επαναλαμβάνονται τα βήματα από το βήμα 2 και μετά μέχρις ότου ικανοποιηθεί η συνθήκη εξόδου.

Με τον τερματισμό του αλγορίθμου έχει δημιουργηθεί ένας πληθυσμός που περιλαμβάνει «καλές» λύσεις. Ο βασικός αλγόριθμος των ΓΑ αναπαριστάται διαγραμματικά στο Σχήμα 3.11.



Σχήμα 3.11 *Αλγόριθμος ΓΑ*

Οι Γενετικοί Αλγόριθμοι έχουν εφαρμοστεί με επιτυχία σε πάρα πολλούς τομείς. Ενδεικτικά αναφέρονται η ρομποτική για την εύρεση διαδρομών, η ασφάλεια, ο σχεδιασμός δικτύων, κεραιών και κυκλωμάτων, η οικονομία για τη διαχείριση χαρτοφυλακίου κλπ. Πολλές φορές χρησιμοποιούνται σε συνδυασμό με άλλες μεθόδους, επιτυγχάνοντας βελτίωση των αποτελεσμάτων. Ένα παράδειγμα συνδυασμού των ΓΑ με άλλες μεθόδους είναι τα υβριδικά μοντέλα κατηγοριοποίησης, τα οποία αναπτύσσονται στο Κεφάλαιο 10. Η επιτυχία των ΓΑ οφείλεται στα πλεονεκτήματα τους, ορισμένα από τα οποία είναι τα εξής:

- Μπορούν να παράγουν καλές λύσεις, οι οποίες μάλιστα βελτιώνονται με την πάροδο του χρόνου.
- Μπορούν και αξιοποιούν προηγούμενες ή εναλλακτικές λύσεις.
- Είναι ικανοί για πολυκριτήρια βελτιστοποίηση, δηλαδή ταυτόχρονη βελτιστοποίηση πολλαπλών στόχων.
- Είναι κατάλληλα για περίπλοκα προβλήματα με συναρτήσεις, οι οποίες έχουν πολλαπλές μέγιστες τιμές ή είναι διακριτές ή έχουν πολλές διαστάσεις ή υπάρχει μη γραμμική σχέση μεταξύ των μεταβλητών.
- Χρησιμοποιούν έννοιες εύκολα κατανοητές.

Μειονεκτήματα των Γενετικών Αλγορίθμων είναι ότι:

- Δεν υπάρχει εγγύηση ότι βρέθηκε η βέλτιστη λύση.
- Η κωδικοποίηση των λύσεων δεν είναι πάντα εύκολη και προφανής.
- Υπάρχουν δυσκολίες στον καθορισμό της Συνάρτησης Καταλληλότητας.
- Δεν παρέχουν κάποια ερμηνεία για τις λύσεις που εντοπίζουν.

3.10 Προσομοίωση

Υπάρχουν περιπτώσεις, όπου το πραγματικό σύστημα είναι τόσο περίπλοκο ή χαρακτηρίζεται από τέτοια στοιχεία αβεβαιότητας, ώστε είναι αδύνατον ή είναι πάρα πολύ δύσκολο να αποτυπωθεί πλήρως σε μαθηματικές σχέσεις και σε μοντέλα βελτιστοποίησης. Σε αυτές τις περιπτώσεις κατασκευάζεται ένα λογισμικό, το οποίο μιμείται τη συμπεριφορά του πραγματικού συστήματος, αναπαριστώντας με μαθηματικό τρόπο τη λειτουργία τμημάτων του πραγματικού συστήματος, καθώς και την αλληλεπίδραση αυτών των τμημάτων. Το λογισμικό αυτό αποτελεί ένα μοντέλο, που προσομοιάζει τη συμπεριφορά του συστήματος και μπορεί να χρησιμοποιηθεί για τη διεξαγωγή πειραμάτων. Αυτή η τεχνική πειραματισμού ονομάζεται προσομοίωση. Σύμφωνα με τον ορισμό του Shannon (1975), προσομοίωση είναι η διαδικασία σχεδιασμού ενός μοντέλου ενός πραγματικού συστήματος και διεξαγωγής πειραμάτων για την κατανόηση της συμπεριφοράς του συστήματος ή για την αξιολόγηση στρατηγικών. Ο χρήστης τροφοδοτεί το μοντέλο με διάφορες τιμές για τις μεταβλητές εισόδου και παρατηρεί τις αντιδράσεις και τη συμπεριφορά του μοντέλου. Το συνηθέστερο παράδειγμα προσομοίωσης είναι οι προσομοιωτές πτήσης, στους οποίους ο χρήστης βλέπει ένα σύστημα πλοήγησης όμοιο με το κόκπιτ ενός αεροπλάνου, εκτελεί χειρισμούς χρησιμοποιώντας τα διαθέσιμα όργανα και το μοντέλο αντιδρά με τρόπο όμοιο με ένα πραγματικό αεροπλάνο.

Ας θεωρήσουμε την περίπτωση ενός στοχαστικού μοντέλου, ενός μοντέλου δηλαδή που χαρακτηρίζεται από κάποιο βαθμό αβεβαιότητας και που οι τιμές των μεταβλητών εισόδου είναι πιθανολογικά ενδεχόμενα. Έχει αναφερθεί στην αρχή του κεφαλαίου ότι η δημιουργία ενός στοχαστικού μοντέλου είναι δυσκολότερη από ότι η δημιουργία ενός ντετερμινιστικού μοντέλου. Μια επιχείρηση επιθυμεί να προβλέψει το ύψος των μελλοντικών κερδών. Τα κέρδη θα προκύψουν από τις πωλήσεις, το κόστος παραγωγής και τις πάγιες δαπάνες. Αν κάποιος γνώριζε με ακρίβεια τις τιμές αυτών των μεταβλητών, θα μπορούσε να υπολογίσει το κέρδος. Όμως το ύψος των πωλήσεων δεν είναι σταθερό, αλλά είναι ένα πιθανολογικό ενδεχόμενο. Μπορούμε να υποθέσουμε ότι το ύψος των πωλήσεων ακολουθεί μια συνεχή συνάρτηση πυκνότητας πιθανότητας, όπως είναι η κανονική κατανομή. Το ίδιο μπορούμε να υποθέσουμε και για τις άλλες μεταβλητές εισόδου. Χρησιμοποιώντας τις κατανομές πιθανοτήτων, επιλέγονται τιμές για τις μεταβλητές εισόδου και υπολογίζεται το αποτέλεσμα, δηλαδή το αναμενόμενο κέρδος. Εάν το πείραμα επαναληφθεί πολλές φορές, με διάφορες τιμές για τις μεταβλητές εισόδου, τότε θα προκύψει ένα σύνολο τιμών για τη μεταβλητή αποτελεσμάτων, που θα είναι ενδεικτικό της κατανομής πιθανοτήτων για το ύψος των κερδών.

Η προσομοίωση είναι μια περιγραφική και όχι κανονιστική μέθοδος. Περιγράφει τα αποτελέσματα ενός συστήματος υπό διαφορετικές συνθήκες. Ουσιαστικά πρόκειται για μια τεχνική διεξαγωγής πειραμάτων. Στην προσομοίωση δεν υπάρχει αυτόματη αναζήτηση λύσεων. Τα στελέχη των επιχειρήσεων τροφοδοτούν το μοντέλο με δεδομένα εισόδου και παρατηρούν τα αποτελέσματα στην έξοδο του συστήματος. Με τον τρόπο αυτό διεξάγουν διαδοχικές αναλύσεις what – if και μελετούν τη συμπεριφορά του συστήματος. Η διαδικασία αυτή επαναλαμβάνεται πολλές φορές. Οι συνδυασμοί συνθηκών και αποτελεσμάτων δημιουργούν μια τεχνητή ιστορία του συστήματος. Η μελέτη αυτής της ιστορίας προσφέρει γνώση για το σύστημα και κατανόηση της συμπεριφοράς του.

Η προσομοίωση έχει εφαρμοστεί σε διάφορα προβλήματα και έχει αποδώσει ικανοποιητικά αποτελέσματα. Μπορεί να χρησιμοποιηθεί σε περιπτώσεις όπου:

- Το πρόβλημα είναι υπερβολικά περίπλοκο για να χρησιμοποιηθεί αριθμητική βελτιστοποίηση. Αυτό μπορεί να σημαίνει ότι δεν υπάρχει κατάλληλος αλγόριθμος για το πρόβλημα ή ότι η φύση του προβλήματος είναι στοχαστική ή ότι υπάρχουν σχέσεις αλληλεξάρτησης μεταξύ των μεταβλητών ή ότι το μοντέλο που θα δημιουργούταν θα ήταν εξαιρετικά περίπλοκο.
- Ο χρήστης επιθυμεί να ελέγξει τις επιπτώσεις των μεταβολών του περιβάλλοντος στα αποτελέσματα του συστήματος.
- Ο χρήστης επιθυμεί να μελετήσει τις αλληλεπιδράσεις μεταξύ υποσυστημάτων και του πλήρους συστήματος.
- Ο στόχος είναι η απόκτηση γνώσης για τη βελτίωση του πραγματικού συστήματος.
- Το ζητούμενο είναι η εύρεση εκείνων των μεταβλητών εισόδου, που διακυμάνσεις στις τιμές τους

έχουν μεγάλες επιπτώσεις στα αποτελέσματα του συστήματος.
- Ο χρήστης επιθυμεί να ελέγξει τα αποτελέσματα πολιτικών πριν την εφαρμογή τους.
- Το μοντέλο θα χρησιμοποιηθεί για εκπαίδευση χρηστών.

Η προσομοίωση δεν είναι πάντοτε η καλύτερη επιλογή. Η εφαρμογή της δεν προκρίνεται όταν:

- Το πρόβλημα μπορεί να λυθεί με αναλυτικές μεθόδους.
- Το κόστος κατασκευής του μοντέλου υπερβαίνει το κέρδος.
- Δεν υπάρχουν οι αναγκαίοι πόροι ή ο χρόνος.

Ορισμένα από τα βασικά πλεονεκτήματα της προσομοίωσης είναι τα ακόλουθα:

- Μπορεί να χρησιμοποιηθεί σε μη δομημένα ή περίπλοκα προβλήματα.
- Είναι ιδιαίτερα αποτελεσματική για τη διεξαγωγή αναλύσεων what – if.
- Επιτρέπει την κατανόηση της επίδρασης των μεταβλητών στα αποτελέσματα.
- Δίνει τη δυνατότητα εικονικού πειραματισμού χωρίς να διαταράσσεται η λειτουργία του πραγματικού συστήματος.
- Επιτρέπει τον πειραματισμό με πόρους, οι οποίοι είναι προς το παρόν ανύπαρκτοι και μπορεί να αποκτηθούν στο μέλλον.
- Καθιστά εφικτή την επιτάχυνση ή την επιβράδυνση φαινομένων.

Μειονεκτήματα της προσομοίωσης είναι τα παρακάτω:

- Κόστος κατασκευής μοντέλου, το οποίο σε ορισμένες περιπτώσεις μπορεί να είναι πολύ υψηλό.
- Δεν προσφέρει κάποια εγγύηση ότι στο τέλος της διαδικασίας θα έχει βρεθεί η βέλτιστη λύση.
- Πολλές φορές η ερμηνεία των αποτελεσμάτων είναι δύσκολη.
- Τα μοντέλα της προσομοίωσης και τα αποτελέσματα των πειραμάτων αφορούν συγκεκριμένο σύστημα ή πρόβλημα και συνήθως δεν μπορούν να γενικευθούν.
- Η διαδικασία ανάπτυξης του μοντέλου και η διαδικασία των πειραμάτων μπορεί να είναι χρονοβόρα.
- Σε αρκετές περιπτώσεις υπάρχει ανάγκη εκπαίδευσης των χρηστών στο μοντέλο.

Η διεξαγωγή πειραμάτων με τη μέθοδο της προσομοίωσης απαιτεί μια δομημένη διαδικασία που αποτελείται από διακριτά βήματα. Έχουν προταθεί διάφορες παραλλαγές ως προς τα βήματα αυτής της διαδικασίας. Οι Turban, Aronson and Liang (2005) αναφέρουν τα παρακάτω στάδια:

- Ορισμός του προβλήματος. Το πρόβλημα μελετάται, οριοθετείται και καθορίζονται διάφορες πλευρές του.
- Κατασκευή του μοντέλου. Καθορίζονται οι μεταβλητές και οι σχέσεις τους και κατασκευάζεται το απαραίτητο λογισμικό.
- Έλεγχος και επικύρωση του μοντέλου. Επιβεβαιώνεται ότι το μοντέλο αντιπροσωπεύει το σύστημα.
- Σχεδιασμός του πειράματος. Ορίζονται οι τιμές για τις μεταβλητές εισόδου και το πλήθος των επαναλήψεων.
- Διεξαγωγή του πειράματος.
- Αξιολόγηση των αποτελεσμάτων. Γίνεται ερμηνεία των αποτελεσμάτων. Μπορεί να εφαρμοστεί ανάλυση ευαισθησίας.
- Υλοποίηση. Γίνεται εφαρμογή των λύσεων στο πραγματικό σύστημα.

Βιβλιογραφία / Αναφορές

Bremermann, H. J. (1958). The Evolution of Intelligence: The Nervous System as a Model of its Environment. *Technical Report No. 1*. Seattle, WA: Department of Mathematics, University of Washington.

Chan, N. H., & Wong, H. Y. (2013). *Handbook of Financial Risk Management: Simulations and Case Studies*. Hoboken, NJ: John Wiley & Sons Inc.

Dantzig, G. (1951). Application of the Simplex Method to a Transportation Problem. In T. Koopmans (Ed.), *Activity Analysis of Production and Allocation* (pp. 359–373). New York, NY: John Wiley and Sons.

Fogel, D. B. (1998). *Evolutionary Computation*. Piscataway, NJ: The Fossil Record, IEEE.

Fraser, A. S. (1957). Simulation of genetic systems by automatic digital computers. II: Effects of linkage on rates under selection. *Australian Journal of Biological Sciences, 10*(4), 492–500. doi: 10.1071/BI9570492

Gabriel, S. A., Kydes, A. S., & Whitman, P. (2001). The National Energy Modeling System: A Large-Scale Energy-Economic Equilibrium Model. *Operations Research, 49*(1), 14-45.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*, 451-482. doi: 10.1146/annurev-psych-120709-145346

Hamby, D. M. (1994). A Review of Techniques for Parameter Sensitivity Analysis of Environmental Models. *Environmental Monitoring and Assessment*, *32*, 135-154. doi: 10.1007/BF00547132

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 49–81). New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511808098.004

Kantorovich, L. (1960). Mathematical methods in the organization and planning of production. *Management Science*, *6*, 550–559.

Law, A. M., & Kelton, W. D. (1991). *Simulation Modeling and Analysis*. New York, NY: McGraw-Hill.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ragsdale, C. (2014). *Spreadsheet Modeling and Decision Analysis: A partial introduction to Business Analytics*. Stamfort, CT: Cengage Learning.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics Made Easy: An Effort-Reduction Framework. *Psychological Bulletin*, *137*(2), 207–222. doi: 10.1037/0033-2909.134.2.207

Shannon, R. E. (1975). *Systems Simulation – The Art and Science*. Englewood Cliffs, NJ: Prentice-Hall.

Simon, D. (2013). *Evolutionary Optimization Algorithms*. Hoboken, NJ: John Wiley & Sons Inc.

Turban, E., Aronson, J. E., & Liang, T. P. (2005). *Decision Support Systems and Intelligence Systems*. New Jersey, NJ: Pearson Education Inc.

Vercellis, C. (2009). *Business Intelligence. Data Mining and Optimization for Decision Making*. Chichester, UK: John Wiley and Sons Ltd.

4 Πολυδιάστατη Ανάλυση και Αποθήκες Δεδομένων

Σύνοψη

Οι σύγχρονες επιχειρήσεις κατακλύζονται από ένα πακτωλό δεδομένων, τα οποία προέρχονται από εσωτερικές και εξωτερικές πηγές. Τα δεδομένα αυτά, αν και πολύτιμα για την εξαγωγή πληροφορίας και τη λήψη αποφάσεων, είναι υπερβολικά λεπτομερή, διασκορπισμένα σε διάφορες πηγές, ανομοιογενή, έχουν βραχύ χρονικό ορίζοντα και πάσχουν από σφάλματα, ελλείψεις, ασυμβατότητες κλπ. Οι Αποθήκες Δεδομένων (ΑΔ) αποτελούν την απάντηση σε αυτά τα προβλήματα και βρίσκονται στο επίκεντρο πληροφοριακών συστημάτων επιφορτισμένων με τη συγκέντρωση, αποθήκευση και ανάλυση των δεδομένων. Το παρόν κεφάλαιο καλύπτει τη θεματική ενότητα των ΑΔ. Αρχικά, παρατίθεται και σχολιάζεται ο ορισμός του Inmon, που προσδιορίζει τα τέσσερα βασικά χαρακτηριστικά των ΑΔ, τα οποία είναι ο θεματικός προσανατολισμός, η ολοκλήρωση, η χρονική διαφοροποίηση και η μη συχνή μεταβολή των δεδομένων. Στη συνέχεια, παρουσιάζεται η βασική αρχιτεκτονική των ΑΔ και τα συστατικά τμήματα τους, όπως η καθεαυτό ΑΔ, το υποσύστημα μεταδεδομένων, η data staging area για τον μετασχηματισμό και φόρτωση των δεδομένων και οι τελικές εφαρμογές που χρησιμοποιεί ο χρήστης. Τα συστήματα επεξεργασίας συναλλαγών με άμεση επικοινωνία (OLTP) χρησιμοποιούνται για την τήρηση των καθημερινών συναλλαγών του οργανισμού, ενώ τα συστήματα αναλυτικής επεξεργασίας άμεσης επικοινωνίας (OLAP) χρησιμοποιούνται για τη διεξαγωγή αναλύσεων. Γίνεται αναλυτική αναφορά στους δύο τύπους συστημάτων και παρατίθενται οι βασικές διαφορές τους. Οι ΑΔ είναι συστήματα ειδικού σκοπού και γι' αυτό ισχύουν διαφορετικές σχεδιαστικές αρχές από ότι στις σχεσιακές βάσεις δεδομένων. Πολύ συνοπτικά παρουσιάζεται το σχεσιακό μοντέλο, και στη συνέχεια αναλύονται οι βασικές έννοιες και τα σχήματα που εφαρμόζονται στις ΑΔ, δηλαδή το σχήμα Αστέρα, το σχήμα Χιονονιφάδας και το σχήμα Αστερισμού. Το πολυδιάστατο μοντέλο που ισχύει στις ΑΔ καλείται κύβος και μπορεί να έχει ν διαστάσεις. Ο κύβος συνίσταται σε ένα πλέγμα κυβοειδών. Αφετηριακό σημείο του πλέγματος είναι το βασικό κυβοειδές, το οποίο καθορίζεται από το σύνολο των διαστάσεων. Με επιλογή διαστάσεων προκύπτουν άλλα κυβοειδή, τα οποία συγκροτούν το πλέγμα. Κατά μήκος μιας διάστασης σημαντικές είναι οι ιεραρχίες εννοιών, οι οποίες συνιστούν μια διάταξη εννοιών σύμφωνα με τον βαθμό γενίκευσης. Χρησιμοποιώντας τον κύβο και τις πράξεις OLAP, ο χρήστης μπορεί να αναλύσει τα δεδομένα και να εκτελέσει τη Συναθροιστική Άνοδο (Roll up), την Αναλυτική Κάθοδο (Drill down), τον Οριζόντιο ή Κάθετο Τεμαχισμό (Slice, Dice) και την Περιστροφή (Pivot). Εκτός της πρωτόβουλης ανάλυσης, που βασίζεται στις υποθέσεις του χρήστη, κατάλληλο λογισμικό μπορεί να σαρώσει τα κελιά του κύβου, να εντοπίσει αποκλίνουσες τιμές και να τις υποδείξει στον χρήστη, ώστε να διενεργήσει περαιτέρω ανάλυση. Εκτεταμένη αναφορά γίνεται στις εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης των δεδομένων στην ΑΔ (εργασίες ETL), καθώς αυτές, λόγω της πολυπλοκότητας και του κόστους τους, αποτελούν ένα διακριτό αντικείμενο μελέτης στον χώρο των ΑΔ. Σημαντικός είναι και ο ρόλος του υποσυστήματος μεταδεδομένων, το οποίο τηρεί πληροφορίες για την ΑΔ, για τις πηγές και τους μετασχηματισμούς των δεδομένων, για κανόνες και στοιχεία πρόσβασης καθώς και για πολλά άλλα. Το υποσύστημα μεταδεδομένων μπορεί να χρησιμεύσει ως οδηγός χρήσης και συντήρησης, αλλά και ως κανονιστικό πλαίσιο λειτουργίας της ΑΔ. Τέλος γίνεται αναφορά στις εφαρμογές των ΑΔ στη σύγχρονη επιχείρηση.

Προαπαιτούμενη γνώση

Το παρόν Κεφάλαιο εισάγει τον αναγνώστη στις Αποθήκες Δεδομένων. Για την καλύτερη κατανόηση εννοιών που σχετίζονται με τον σχεδιασμό των ΑΔ χρειάζονται προηγούμενες γνώσεις για αντίστοιχες έννοιες του σχεσιακού μοντέλου. Στο Κεφάλαιο γίνεται μια επιγραμματική παρουσίαση του σχεσιακού μοντέλου. Ο αναγνώστης, που δεν έχει προηγούμενη εμπειρία στο αντικείμενο και θεωρεί απαραίτητες πρόσθετες πηγές, μπορεί να αναζητήσει γνώσεις υποδομής σε ένα από τα πολλά βιβλία που αναφέρονται στις Σχεσιακές Βάσεις Δεδομένων, όπως το Date (2012) ή το Coronel and Morris (2014). Το πεδίο των ΑΔ έχει καθοριστεί από το έργο δύο κορυφαίων ειδικών, του Bill Inmon και του Ralph Kimball. Το δημοσιευμένο έργο αυτών των δύο συγγραφέων αντανακλά την εξέλιξη των ΑΔ και προσφέρει πολύτιμες λεπτομέρειες. Δυο βιβλία σταθμοί είναι το Inmon (1996) (τελευταία έκδοση 2005) και το Kimball and Ross (2013). Επίσης,, η ιστοσελίδα του Ralph Kimball Group (Kimball Group, 2015) είναι μια βασική πηγή πληροφοριών για τις ΑΔ. Οι εργασίες ETL είναι ένα σημαντικό και περίπλοκο αντικείμενο. Γενικά ζητήματα ETL καλύπτονται στο Kimball and Caserta (2004).Ωστόσο, οι εργασίες ETL είναι ένα πολύ τεχνικό αντικείμενο και πολλοί συγγραφείς το αντιμετωπίζουν σε συνάρτηση με συγκεκριμένα λογισμικά, όπως οι Dupupet and Grays (2013),οι οποίοι αναφέρονται σε εργαλεία της Oracle και οι Knight, Knight, Moss, Davis and Rock (2014), οι οποίοι αναφέρονται σε εργαλεία του Microsoft SQL Server.

Τέλος, πάροχοι λογισμικού Επιχειρηματικής Ευφυΐας παρουσιάζουν ειδικότερα πεδία εφαρμογής των ΑΔ στη σύγχρονη επιχείρηση. Ενδεικτικά αναφέρουμε την ιστοσελίδα Business Intelligence Tools & Data Warehousing Applications της Teradata («Business Intelligence Tools & Data Warehousing Applications - Teradata,» n.d.).

4.1 Εισαγωγή - Ορισμός

Θα ήταν ίσως κοινοτυπία, εάν επαναλαμβάναμε ότι η σύγχρονη επιχείρηση του 21ου αιώνα έχει ενσωματώσει την τεχνολογία της πληροφορικής και την έχει καταστήσει βασική διάσταση της λειτουργίας της. Η παραγωγή και ροή της πληροφορίας μέσα στην επιχείρηση πραγματοποιείται με τη βοήθεια μηχανογραφικών συστημάτων, τα οποία καταγράφουν κάθε συναλλαγή. Τα συστήματα αυτά, γνωστά ως Συστήματα Επεξεργασίας Συναλλαγών Άμεσης Επικοινωνίας (On Line Transaction Processing Systems (OLTP)), ξεκίνησαν πριν από μερικές δεκαετίες, μηχανοργανώνοντας βασικές καθημερινές λειτουργίες της επιχείρησης, όπως η παρακολούθηση της αποθήκης, των πωλήσεων, των αγορών, των πελατών, των χρηματοοικονομικών στοιχείων. Με την πάροδο του χρόνου εξελίχθηκαν στα σύγχρονα ολοκληρωμένα συστήματα διαχείρισης επιχειρησιακών διαδικασιών, όπως είναι τα συστήματα Σχεδιασμού Επιχειρηματικών Πόρων (Enterprise Resources Planning (ERP)). Απαραίτητα για την καθημερινή λειτουργία της επιχείρησης, τα συστήματα αυτά ήταν τα πρώτα που αναπτύχθηκαν στα πλαίσια της εφαρμογής της μηχανοργάνωσης και αποτελούν σήμερα όρο λειτουργίας για τους μεγάλους οργανισμούς. Η έλευση του Ηλεκτρονικού Εμπορίου δημιούργησε νέες ανάγκες και έδωσε πρόσθετη ώθηση στη μηχανογραφική παρακολούθηση των συναλλαγών.

Πέραν όμως των διαδικασιών καθημερινής λειτουργίας της επιχείρησης, υπάρχουν πρόσθετα ζητήματα, που αφορούν τη διοίκηση των οργανισμών. Η αντιμετώπιση τέτοιων ζητημάτων περιλαμβάνει την πρόβλεψη, τον σχεδιασμό και τη λήψη αποφάσεων. Τα σύγχρονα στελέχη επιχειρήσεων, τα οποία λειτουργούν σε συνθήκες αβεβαιότητας και ρίσκου, είναι υποχρεωμένα να αναζητούν τη μέγιστη δυνατή πληροφόρηση, ώστε να περιορίσουν το ρίσκο και να λάβουν κατά το δυνατόν λογικές αποφάσεις. Τα δεδομένα όμως των Συστημάτων Επεξεργασίας Συναλλαγών, αν και ζωτικής σημασίας για τη λειτουργία της επιχείρησης, δεν είναι τα πλέον κατάλληλα για ανάλυση και λήψη αποφάσεων. Ειδικότερα, τα δεδομένα αυτά έχουν τα εξής χαρακτηριστικά:

- Ο όγκος τους είναι τέτοιος, που προκαλεί προβλήματα στην επεξεργασία.
- Ο βαθμός λεπτομέρειας τους είναι πολύ μεγάλος. Για τη λήψη αποφάσεων χρειάζεται περιληπτική και συγκεντρωτική πληροφόρηση.
- Τα δεδομένα είναι διάσπαρτα και αποθηκευμένα σε πολλές διαφορετικές πηγές.
- Κατά κανόνα τα λειτουργικά δεδομένα έχουν βραχύ ιστορικό ορίζοντα. Για την ανάλυση, τη σύγκριση και τη λήψη αποφάσεων πιθανότατα θα χρειαστεί επεξεργασία παλαιότερων δεδομένων, ίσως και σε βάθος δεκαετίας.
- Η ποιότητα των δεδομένων είναι χαμηλή, καθώς περιέχουν ελλείψεις, αντιφάσεις, διαφορετικές κωδικοποιήσεις κλπ. Η ύπαρξη προβλημάτων είναι ο κανόνας στα δεδομένα του πραγματικού κόσμου.

Εκτός από τα δεδομένα των Συστημάτων Επεξεργασίας Συναλλαγών, η σύγχρονη επιχείρηση αντλεί πληροφόρηση και από άλλες πηγές. Οι διαδικτυακοί επιχειρηματικοί σέρβερς καταγράφουν το ρεύμα κλικ των πελατών, την πλοήγηση δηλαδή των πελατών στην ιστοθέση του οργανισμού και στον ιστό γενικότερα. Τρίτοι φορείς όπως τράπεζες, κυβερνητικοί οργανισμοί κλπ. παρέχουν τα δικά τους δεδομένα. Επίσης, το Web 2.0, με τα ιστολόγια και τις ιστοθέσεις κοινωνικής δικτύωσης, μπορεί να αξιοποιηθεί για μια διαρκή σφυγμομέτρηση της κοινής γνώμης. Τα δεδομένα, τα οποία προέρχονται από αυτές τις πηγές, είναι κατά κανόνα, σε αντίθεση με τα λειτουργικά δεδομένα, αδόμητα, γεγονός που θέτει ειδικές απαιτήσεις επεξεργασίας. Οι οργανισμοί των αρχών του 21ου αιώνα τροφοδοτούνται με έναν πακτωλό δεδομένων. Οι επιχειρήσεις εκείνες που θα μπορέσουν να αξιοποιήσουν αυτά τα δεδομένα, θα βελτιώσουν την ποιότητα των αποφάσεων τους και θα αποκτήσουν συγκριτικό πλεονέκτημα έναντι των ανταγωνιστών τους. Προκύπτει λοιπόν η ανάγκη για εξειδικευμένα πληροφοριακά συστήματα, προσανατολισμένα στη συγκέντρωση, αποθήκευση και ανάλυση των απαραίτητων δεδομένων. Τα συστήματα αυτά θα χρησιμοποιηθούν για την εξαγωγή συμπερασμάτων και τη λήψη αποφάσεων. Οι Αποθήκες Δεδομένων βρίσκονται στο επίκεντρο αυτών των πληροφορικών συστημάτων.

Μια **Αποθήκη Δεδομένων** (ΑΔ) (Data Warehouse (DW)) είναι μια βάση δεδομένων διαφορετική από τις βάσεις δεδομένων που τηρούν τα λειτουργικά δεδομένα του οργανισμού. Στις ΑΔ μεταφέρονται και συλλέγονται δεδομένα από άλλες πηγές. Τα δεδομένα αυτά απαλλάσσονται από προβλήματα, ομογενοποιούνται, αποθηκεύονται σε συγκεντρωτική μορφή και χρησιμοποιούνται για ανάλυση, εξαγωγή συμπερασμάτων και λήψη αποφάσεων. Σύμφωνα με τον πολύ δημοφιλή ορισμό του Inmon (1996), μια Αποθήκη Δεδομένων είναι

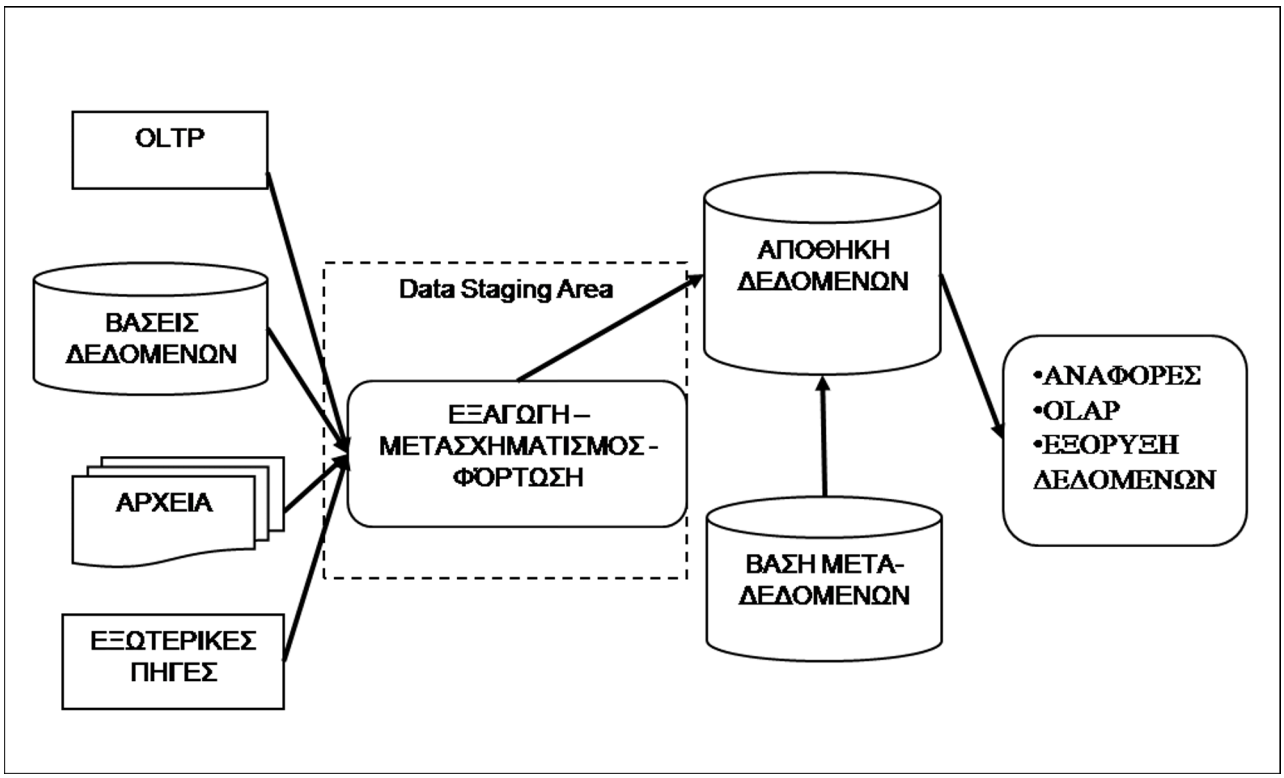
μια θεματικά προσανατολισμένη, ολοκληρωμένη, χρονικά διαφοροποιούμενη και μη ευμετάβλητη συλλογή δεδομένων, που χρησιμοποιείται για την υποστήριξη της διαδικασίας λήψης αποφάσεων. Ειδικότερα, τα τέσσερα βασικά χαρακτηριστικά της ΑΔ έχουν ως εξής:

- **Θεματικός Προσανατολισμός.** Στις ΑΔ η πληροφορία είναι οργανωμένη με βάση κάποιες κεντρικές έννοιες, όπως ο πελάτης, το προϊόν ή οι πωλήσεις. Επιδίωξη είναι η συγκέντρωση, οργάνωση και παρουσίαση της πληροφορίας, που σχετίζεται με αυτές τις έννοιες και μάλιστα με τρόπο που να διευκολύνεται η διαδικασία λήψης αποφάσεων. Αντιθέτως, στα Συστήματα Επεξεργασίας Συναλλαγών, επιδίωξη είναι η καταγραφή των καθημερινών συναλλαγών μεταξύ οντοτήτων. Οι ΑΔ επιτυγχάνουν να παρέχουν στοχευμένη πληροφόρηση για συγκεκριμένα ζητήματα, περιλαμβάνοντας και οργανώνοντας κατάλληλα τις σχετικές πληροφορίες και αποκλείοντας τις μη χρήσιμες πληροφορίες.
- **Ολοκλήρωση.** Στις ΑΔ μεταφέρονται δεδομένα από πολλές διαφορετικές πηγές, όπως συστήματα επεξεργασίας συναλλαγών, ανεξάρτητες βάσεις δεδομένων κλπ. Δεν είναι καθόλου σπάνιο αυτά τα δεδομένα να πάσχουν από προβλήματα, όπως διαφορετικές ονοματοδοσίες, διαφορετικές κωδικοποιήσεις, διαφορετικές μονάδες μέτρησης κλπ. Για παράδειγμα, σε ορισμένες βάσεις δεδομένων το μήκος μπορεί να τηρείται χρησιμοποιώντας μέτρα και εκατοστά, ενώ σε άλλες να χρησιμοποιούνται γιάρδες και ίντσες. Επίσης, το ανθρώπινο φύλο μπορεί να κωδικοποιείται ως «Α» και «Θ», ενώ αλλού ως «1» και «0». Κατά τη μεταφορά τους στην ΑΔ, τα δεδομένα «καθαρίζονται», ομογενοποιούνται και στη συνέχεια αποθηκεύονται απαλλαγμένα από προβλήματα.
- **Χρονική Διαφοροποίηση.** Τα Συστήματα Επεξεργασίας Συναλλαγών τηρούν τρέχουσα πληροφορία, αποτυπώνουν δηλαδή την τωρινή κατάσταση του οργανισμού. Αντιθέτως, οι Αποθήκες Δεδομένων τηρούν ιστορική πληροφορία, που μπορεί να αναφέρεται σε βάθος χρόνου μέχρι και δεκαετίας. Με τον τρόπο αυτό αποτυπώνουν πολλαπλά ιστορικά στιγμιότυπα του οργανισμού. Οι έννοιες μιας ΑΔ έχουν μια χρονική διάσταση. Με την καταγραφή της χρονικής εξέλιξης καθίσταται εφικτή η διεξαγωγή συγκρίσεων και η αναγνώριση τάσεων.
- **Μη ευμετάβλητα δεδομένα.** Τα δεδομένα των Συστημάτων Επεξεργασίας Συναλλαγών τελούν υπό συνεχή ανανέωση, καθώς οι χρήστες διαρκώς εισάγουν, τροποποιούν και διαγράφουν δεδομένα. Αντιθέτως, στις ΑΔ τα δεδομένα μεταφέρονται μαζικά σε συγκεκριμένες χρονικές στιγμές, και στη συνέχεια προσπελαύνονται με σκοπό την ανάλυση τους, αλλά δεν τροποποιούνται.

Όπως ήδη αναφέρθηκε, μια Αποθήκη Δεδομένων είναι μια διαφορετική βάση δεδομένων από αυτή στην οποία καταγράφονται οι συναλλαγές αναλυτικά. Εκτός από τους λόγους συγχώνευσης και καθαρισμού των δεδομένων, αυτό επιβάλλεται και για λόγους ταχύτητας λειτουργίας των συστημάτων. Εάν χρησιμοποιούνταν το σύστημα OLTP για διεξαγωγή αναλύσεων, μια σύνθετη ερώτηση με περίπλοκους υπολογισμούς, η οποία απαιτεί το «κλείδωμα» πινάκων, θα μπλόκαρε τη δυνατότητα μεταβολής των δεδομένων και θα καθυστερούσε την εργασία όλων των χρηστών για μεγάλο χρονικό διάστημα. Φυσικά μια τέτοια καθυστέρηση όλου του συστήματος θεωρείται απαράδεκτη.

4.2 Αρχιτεκτονική Αποθήκης Δεδομένων

Στο Σχήμα 4.1 παρουσιάζεται η γενική αρχιτεκτονική μιας Αποθήκης Δεδομένων. Όπως φαίνεται στο Σχήμα 4.1, μια Αποθήκη Δεδομένων αντλεί δεδομένα από πολλές διαφορετικές πηγές. Σε αυτές συμπεριλαμβάνονται συστήματα επεξεργασίας συναλλαγών, όπως συστήματα ERP, συστήματα SCM και CRM, άλλες βάσεις λειτουργικών δεδομένων, αρχεία και λοιπές εξωτερικές πηγές. Τα δεδομένα συγκεντρώνονται σε ένα ενδιάμεσο χώρο, που ονομάζεται Data Staging Area και εκεί υφίστανται επεξεργασία, ώστε να απαλλαγούν από διαφόρων ειδών προβλήματα. Ειδικότερα, αντιμετωπίζονται προβλήματα διαφορετικών ονομασιών, μονάδων μέτρησης, κωδικοποιήσεων κλπ. Επίσης, τα δεδομένα συναθροίζονται σύμφωνα με έννοιες που ενδιαφέρουν τους αναλυτές και σε κατάλληλο βαθμό λεπτομέρειας. Για παράδειγμα, μπορεί να υπολογίζονται συγκεντρωτικά στοιχεία πωλήσεων ανά πελάτη. Τα δεδομένα, αφού υποστούν αυτήν την επεξεργασία, αποθηκεύονται στην ΑΔ. Η Βάση Μεταδεδομένων τηρεί πληροφορίες σχετικά με τη διαδικασία φόρτωσης της ΑΔ, λεπτομέρειες για τη δομή της και άλλες βοηθητικές πληροφορίες. Τέλος, τα δεδομένα της ΑΔ είναι διαθέσιμα για ανάλυση. Η ανάλυση μπορεί να περιλαμβάνει την υποβολή ερωτημάτων (queries) στην ΑΔ και τη σύνταξη αναφορών, τη διεξαγωγή πράξεων Αναλυτικής Επεξεργασίας Άμεσης Επικοινωνίας (OLAP) ή την εφαρμογή μεθόδων Εξόρυξης Δεδομένων.



Σχήμα 4.1 *Αρχιτεκτονική Αποθήκης Δεδομένων*

Η αρχιτεκτονική που παρουσιάζεται στο Σχήμα 4.1 είναι γενική και δεν είναι απόλυτη. Για παράδειγμα, ένας οργανισμός μπορεί να διατηρεί παράλληλα και έναν αριθμό Πρατηρίων Δεδομένων (Data Marts). Τα Πρατήρια Δεδομένων είναι μικρές και εξειδικευμένου σκοπού ΑΔ. Μπορεί να αντλούν δεδομένα από την κεντρική ΑΔ ή να τροφοδοτούνται από τις εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης.

4.3 OLTP και OLAP

Έχουμε ήδη αναφερθεί στα Συστήματα Επεξεργασίας Συναλλαγών (OLTP). Πρόκειται για πληροφοριακά συστήματα που χρησιμοποιούνται για την καταγραφή και επεξεργασία των συναλλαγών μιας επιχείρησης. Οι ΑΔ σχετίζονται με τα συστήματα Αναλυτικής Επεξεργασίας των δεδομένων (On Line Analytical Processing (OLAP)). Τα συστήματα OLAP χρησιμοποιούνται από τους αναλυτές και τα υψηλόβαθμα στελέχη των επιχειρήσεων για τη διεξαγωγή αναλύσεων και τη λήψη αποφάσεων. Τους εξασφαλίζουν ταχεία και ευέλικτη πρόσβαση σε μεγάλους όγκους δεδομένων και τους επιτρέπουν την πολυδιάστατη επεξεργασία τους. Ο όρος πολυδιάστατη επεξεργασία περιγράφει τη δυνατότητα συνολικοποίησης και παρουσίασης των δεδομένων σε διαφορετικό βαθμό αφαίρεσης ή σύμφωνα με διαφορετικές έννοιες. Για παράδειγμα, ένας χρήστης θα μπορούσε να υποβάλλει ένα ερώτημα σε ένα σύστημα OLAP, ζητώντας να πληροφορηθεί το συνολικό ύψος πωλήσεων ανά γεωγραφική περιοχή ή ανά κατηγορία προϊόντος. Στη συνέχεια, θα μπορούσε να εξειδικεύσει περαιτέρω το ερώτημα του, ζητώντας τα σύνολα πωλήσεων ανά κατηγορία προϊόντος για τη γεωγραφική περιοχή της Πελοποννήσου το πρώτο τρίμηνο του τρέχοντος έτους. Γενικώς ο χρήστης έχει τη δυνατότητα να υποβάλλει ελεύθερα μη προκαθορισμένες ερωτήσεις, επικεντρώνοντας σε ζητήματα που τον ενδιαφέρουν και αυξομειώνοντας τον βαθμό γενίκευσης. Σε συνδυασμό με τις ΑΔ, οι οποίες τηρούν τα δεδομένα σε κατάλληλη μορφή, τα συστήματα OLAP μετατρέπουν τα ακατέργαστα δεδομένα σε στρατηγικά αξιοποιήσιμη πληροφορία, οργανώνοντας και παρουσιάζοντας τα με τρόπο που αντανακλά επιχειρηματικά ζητήματα, όπως αυτά γίνονται κατανοητά από τα επιχειρηματικά στελέχη. Βασικό χαρακτηριστικό των συστημάτων OLAP είναι ότι μπορούν να αποδώσουν με μεγάλη ταχύτητα πληροφορία, η οποία πηγάζει από την επεξεργασία μεγάλου όγκου δεδομένων. Επίσης, μπορούν να δώσουν απαντήσεις σε πιο περίπλοκα ερωτήματα από ότι μια παραδοσιακή βάση δεδομένων. Άλλο σημαντικό χαρακτηριστικό τους είναι ότι ο χρήστης χειρίζεται απευθείας τα δεδομένα χωρίς τη μεσολάβηση κάποιας εφαρμογής, όπως πχ ενός λογισμικού CRM.

Είναι σαφές ότι η φύση και η αποστολή των συστημάτων OLTP και OLAP είναι διαφορετική. Για τον λόγο αυτό, παρουσιάζουν πολλές διαφορές στα χαρακτηριστικά τους. Οι Han and Kamber (2001) παραθέτουν

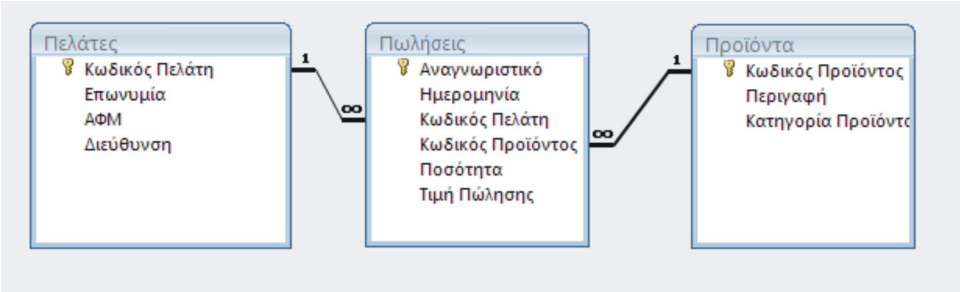
διαφορές μεταξύ των δύο συστημάτων. Επίσης, στο Διαδίκτυο ποικίλοι φορείς αναφέρονται στο ίδιο θέμα, αναδεικνύοντας πρόσθετες διαφοροποιήσεις. Στον Πίνακα 4.1 συνοψίζονται οι βασικότερες διαφορές μεταξύ των συστημάτων OLTP και OLAP.

Χαρακτηριστικό	**OLTP**	**OLAP**
Χρήστες	Χαμηλόβαθμοι υπάλληλοι	Αναλυτές και υψηλόβαθμα στελέχη
Σκοπός	Τήρηση λειτουργικών δεδομένων και παρακολούθηση των καθημερινών συναλλαγών	Εξαγωγή πληροφοριών για λήψη αποφάσεων
Επίπεδο αναφοράς	Λειτουργικό - τακτικό	Στρατηγικό – τακτικό
Σχεδιασμός ΒΔ	Καθοδηγούμενος από την εφαρμογή	Καθοδηγούμενος από τα εξεταζόμενα ζητήματα
Σχήμα βάσης δεδομένων	Σχεσιακό κανονικοποιημένο	Αστέρα ή χιονονιφάδας
Μοντέλο δεδομένων	Οντότητας –σχέσης	Πολυδιάστατο
Πλεονασμός δεδομένων	Υψηλός βαθμός κανονικοποίησης δεδομένων με τον χαμηλότερο δυνατό πλεονασμό και πολλούς πίνακες	Μη κανονικοποιημένα δεδομένα που επιτρέπουν τον πλεονασμό
Τάξη μεγέθους δεδομένων	Gigabyte	Terabyte
Βαθμός σύνοψης δεδομένων	Υψηλός βαθμός λεπτομέρειας	Υψηλός βαθμός σύνοψης – αθροιστικά δεδομένα
Χρονική διάσταση δεδομένων	Τρέχοντα, δυναμικά	Ιστορικά, στατικά
Τρόπος ενημέρωσης των δεδομένων	Διαρκής μεταβολή	Περιοδική ανανέωση
Συνήθεις πράξεις στη ΒΔ	Εισαγωγή, διαγραφή, τροποποίηση, ανάγνωση	Ανάγνωση
Ερωτήματα στη ΒΔ	Συνήθως τυποποιημένα και απλά ερωτήματα που επιστρέφουν μερικές εγγραφές	Μη τυποποιημένα και συνήθως σύνθετα ερωτήματα που απαιτούν συναθροίσεις.

Πίνακας 4.1 *Διαφορές OLTP - OLAP*

4.4 Σχεδιασμός Αποθήκης Δεδομένων

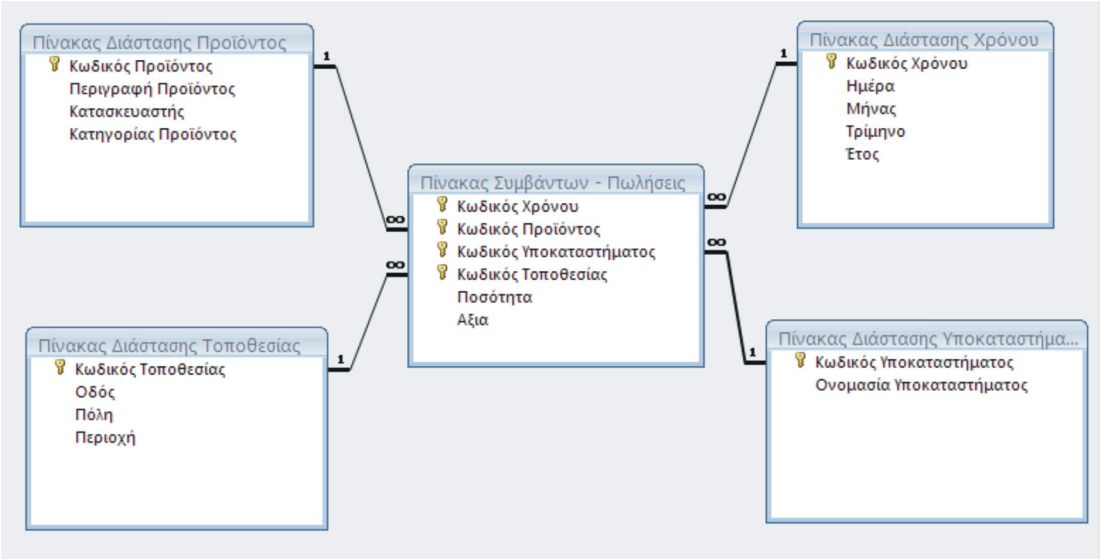
Όπως έχει αναφερθεί και παραπάνω, οι Αποθήκες Δεδομένων είναι βάσεις δεδομένων ειδικού σκοπού και καλούνται να εξυπηρετήσουν ειδικές ανάγκες. Ως αποτέλεσμα αυτού, για τις Αποθήκες Δεδομένων ισχύουν διαφορετικές σχεδιαστικές αρχές από αυτές των Σχεσιακών Βάσεων Δεδομένων. Ο τρόπος σχεδιασμού και δόμησης των Σχεσιακών Βάσεων Δεδομένων δεν αποτελεί αντικείμενο του παρόντος συγγράμματος, ωστόσο θα επιχειρήσουμε μια επιγραμματική παρουσίαση του, με σκοπό την καλύτερη κατανόηση των βασικών εννοιών και των διαφορών. Οι κλασικές σχεσιακές βάσεις δεδομένων είναι οργανωμένες ως ένα σύνολο πινάκων, ο καθένας από τους οποίους αναφέρεται σε μια οντότητα ή αντικείμενο ή γεγονός, και αποτελείται από στήλες, που αντιστοιχούν στα χαρακτηριστικά της οντότητας, και γραμμές, που αντιστοιχούν σε ένα στιγμιότυπο της οντότητας. Για παράδειγμα, τα στοιχεία των πελατών θα τηρούνται σε έναν πίνακα όπου οι στήλες θα είναι ο κωδικός του πελάτη, η επωνυμία, το ΑΦΜ, η διεύθυνση κλπ. ενώ σε μία γραμμή καταχωρούνται τα στοιχεία ενός συγκεκριμένου πελάτη. Σε κάθε πίνακα υπάρχει μια στήλη που ονομάζεται πρωτεύων κλειδί και έχει μοναδική τιμή σε κάθε γραμμή. Στο παράδειγμα με τον πίνακα στοιχείων πελατών, κάθε πελάτης θα έχει τον δικό του μοναδικό κωδικό και ο κωδικός είναι το πρωτεύων κλειδί. Οι πίνακες σχετίζονται μεταξύ τους και οι σχέσεις καταγράφονται με τη συμμετοχή των πρωτευόντων κλειδιών ενός πίνακα σε έναν άλλο πίνακα. Τα πρωτεύοντα κλειδιά, που συμμετέχουν σε έναν άλλο πίνακα για να καταγράψουν μια σχέση, καλούνται ξένα κλειδιά. Η επιχείρηση του παραδείγματος μας, για να τηρήσει τα στοιχεία πωλήσεων, θα κατασκεύαζε τον πίνακα των πελατών όπως αναφέρθηκε παραπάνω, έναν πίνακα προϊόντων με στήλες τον κωδικό του προϊόντος, την περιγραφή του και την κατηγορία του, και έναν πίνακα πωλήσεων με στήλες ένα πρωτεύων κλειδί, την ημερομηνία, τον κωδικό πελάτη (ξένο κλειδί), τον κωδικό προϊόντος (δεύτερο ξένο κλειδί), την ποσότητα και την τιμή πώλησης. Οι πίνακες με τις στήλες και τις σχέσεις τους παρουσιάζονται διαγραμματικά στο Σχήμα 4.2.



Σχήμα 4.2 *Σχεσιακό μοντέλο*

Στη βάση δεδομένων μπορεί να προστίθενται, να διαγράφονται και να τροποποιούνται στοιχεία για πελάτες, προϊόντα και πωλήσεις, η δομή της όμως παραμένει αμετάβλητη. Ο όρος Σχήμα Βάσης Δεδομένων αναφέρεται στον καθορισμό της αμετάβλητης λογικής δομής της βάσης δεδομένων, δηλαδή στους πίνακες που την αποτελούν, στις στήλες τους, στα πεδία ορισμού και στις ιδιότητες των στηλών, στις σχέσεις μεταξύ των πινάκων, καθώς και σε περιορισμούς ακεραιότητας των δεδομένων. Βασική επιδίωξη του σχεσιακού μοντέλου *είναι* ο περιορισμός *του πλεονασμού* των δεδομένων (data redundancy). Πλεονασμός των δεδομένων με απλά λόγια σημαίνει ότι η ίδια πληροφορία καταχωρείται πολλές φορές. Ο πλεονασμός των δεδομένων σπαταλά αποθηκευτικό χώρο, κυρίως όμως είναι αιτία ασυνέπειας των δεδομένων, ύπαρξης δηλαδή αντιφατικής πληροφορίας. Αν για παράδειγμα, τα πλήρη στοιχεία του πελάτη επαναλαμβάνονταν σε κάθε εγγραφή πωλήσεων, είναι πιθανόν σε διαφορετικές εγγραφές πωλήσεων ο ίδιος πελάτης να αναφερόταν με διαφορετικό ΑΦΜ, γεγονός παράλογο και παράνομο, τεχνικά όμως δυνατό. Ο περιορισμός του πλεονασμού των δεδομένων καλείται *κανονικοποίηση*, και έχουν οριστεί κανονικές μορφές, που πρέπει να τηρούνται κατά τον σχεδιασμό μιας σχεσιακής βάσης δεδομένων. Η κανονικοποίηση προκαλεί τον κατατεμαχισμό των δεδομένων σε πολλούς αλληλοσυνδεόμενους πίνακες. Το σχεσιακό μοντέλο επιτρέπει την τήρηση δεδομένων με τον μέγιστο βαθμό λεπτομέρειας και είναι ιδανικό για την καταγραφή των καθημερινών συναλλαγών.

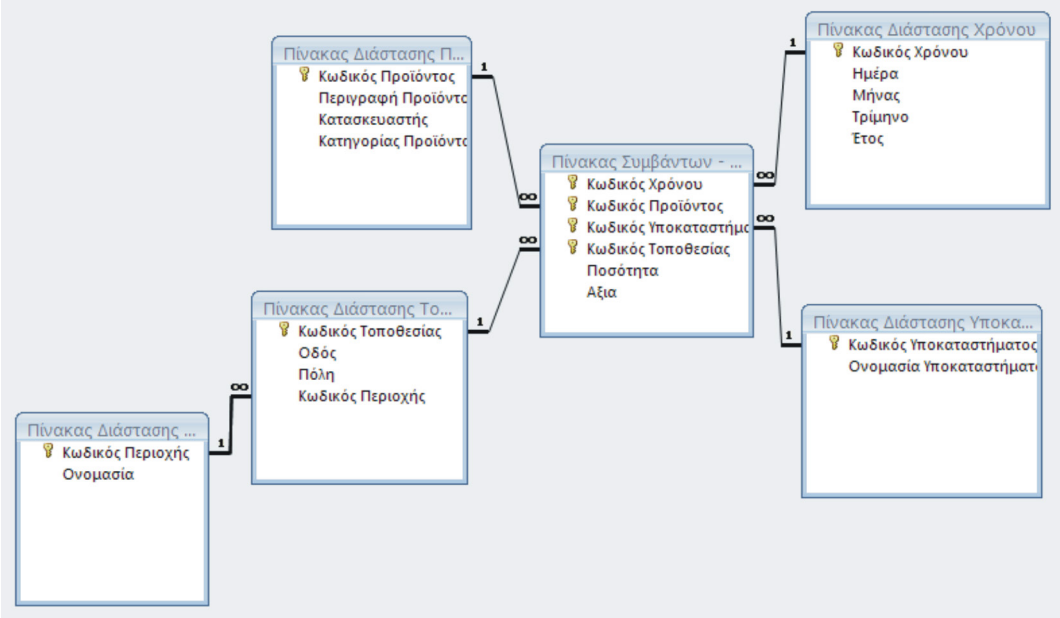
Για τις Αποθήκες Δεδομένων έχει προταθεί ένα διαφορετικό Σχήμα, που προσανατολίζεται σε αντικείμενα ενδιαφέροντος και που διευκολύνει τη διεξαγωγή αναλύσεων OLAP. Οι βασικές έννοιες του σχήματος προτάθηκαν από τον Ralph Kimball (Kimball Group, 1995). Καταρχήν, ορίζεται ένας πίνακας που περιέχει μεγάλο όγκο δεδομένων. Ο πίνακας αυτός ονομάζεται **Πίνακας Συμβάντων** (Fact Table) και αναφέρεται στο αντικείμενο το οποίο θα αναλυθεί. Αν για παράδειγμα, ενδιαφερόμαστε να αναλύσουμε τα στοιχεία πωλήσεων, τότε δημιουργούμε έναν Πίνακα Συμβάντων για τις πωλήσεις, ενώ αν ενδιαφερόμαστε να αναλύσουμε τα στοιχεία των αγορών, δημιουργούμε έναν Πίνακα Συμβάντων για τις αγορές. Επιπλέον του Πίνακα Συμβάντων, δημιουργούνται οι **Πίνακες Διαστάσεων** (Dimension Tables). Οι πίνακες διαστάσεων αναφέρονται σε ιδιότητες των συμβάντων και αντιστοιχούν στις παραμέτρους με βάση τις οποίες θα γίνει η ανάλυση. Σε έναν πίνακα συμβάντων λιανικών πωλήσεων μιας αλυσίδας καταστημάτων, πίνακες διαστάσεων μπορεί να είναι το προϊόν, η χρονική στιγμή πώλησης, το υποκατάστημα που έγινε η πώληση και η τοποθεσία. Μπορούμε να πούμε ότι ο πίνακας συμβάντων σχετίζεται με το *τι πληροφορία* θα αναλυθεί, ενώ οι πίνακες διαστάσεων σχετίζονται με το *πώς και ως προς τι θα αναλυθεί* η πληροφορία. Η βασική εκδοχή σχήματος μιας Αποθήκης Δεδομένων είναι το **Σχήμα Αστέρα** (Star Schema), σύμφωνα με το οποίο τοποθετείται κεντρικά ο πίνακας συμβάντων και περιμετρικά, σαν ακτίνες αστεριού, οι πίνακες διαστάσεων. Παράδειγμα Σχήματος Αστέρα παρουσιάζεται στο Σχήμα 4.3



Σχήμα 4.3 *Σχήμα Αστέρα*

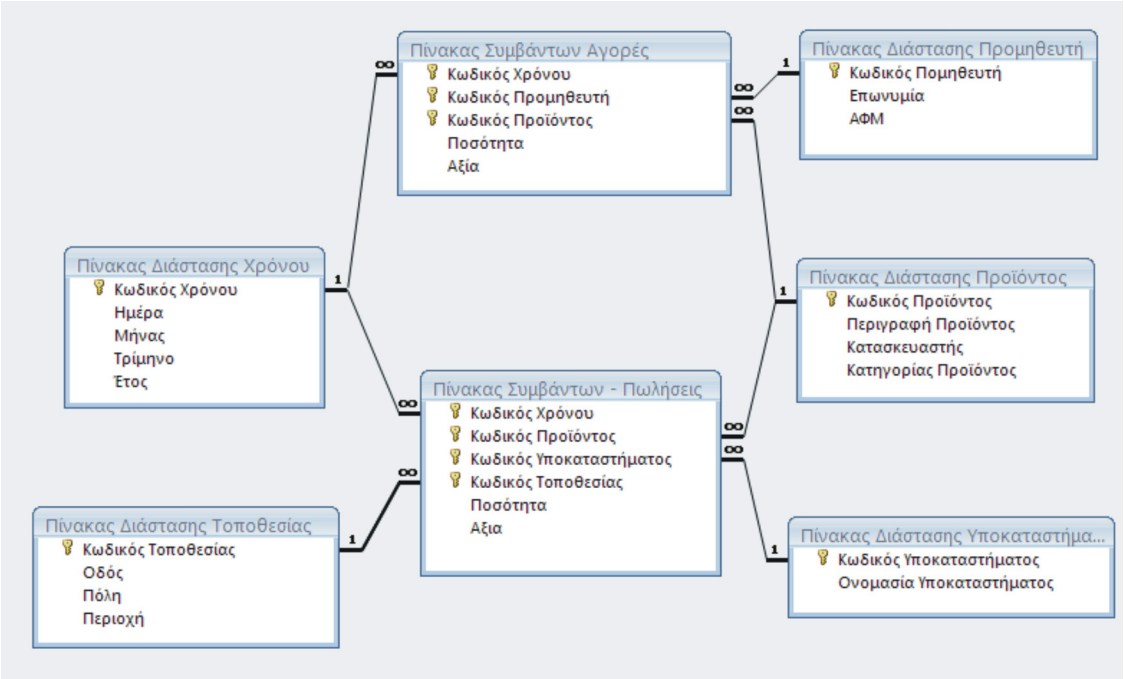
Όπως παρατηρούμε στο Σχήμα 4.3, Ο πίνακας συμβάντων περιλαμβάνει σαν στήλες τα κλειδιά των διαστάσεων και δύο επιπλέον αριθμητικά πεδία, την ποσότητα του εμπορεύματος και την αξία της πώλησης. Τα δύο αυτά πεδία καλούνται **μέτρα** (measures) και είναι τα μεγέθη τα οποία θα αναλυθούν. Με το παραπάνω σχήμα μπορούμε εύκολα να παρακολουθούμε την ποσότητα και την αξία των πωλήσεων ανά χρονική στιγμή, υποκατάστημα, τοποθεσία και προϊόν σε διάφορους συνδυασμούς. Παρατηρούμε επίσης ότι στον πίνακα συμβάντων δεν υπάρχει πλεονασμός δεδομένων. Αντιθέτως, στους πίνακες διαστάσεων υπάρχει πλεονασμός των δεδομένων. Για παράδειγμα, στον πίνακα τοποθεσία η περιοχή είναι πλεονάζων στοιχείο, αφού θα μπορούσε να συναχθεί με τη βοήθεια ενός πρόσθετου πίνακα, που θα τηρούσε τα στοιχεία των περιοχών.

Μια παραλλαγή του σχήματος Αστέρα είναι το **Σχήμα Χιονονιφάδας** (Snowflake Schema). Στο Σχήμα Χιονονιφάδας γίνεται κανονικοποίηση των πινάκων διαστάσεων έτσι ώστε να επιτευχθεί μείωση του πλεονασμού των δεδομένων. Το Σχήμα 4.4 παρουσιάζει το Σχήμα Χιονονιφάδας.



Σχήμα 4.4 *Σχήμα Χιονονιφάδας.*

Χαρακτηριστικό των σχημάτων Αστέρα και Χιονονιφάδας είναι ότι υπάρχει μόνο ένας πίνακας Συμβάντων. Σε μια Αποθήκη Δεδομένων όμως μπορεί να υπάρχουν περισσότεροι πίνακες Συμβάντων. Αυτοί οι πίνακες Συμβάντων μπορεί να έχουν κοινές διαστάσεις. Προκύπτει τότε ένα πιο περίπλοκο σχήμα που ονομάζεται **Σχήμα Αστερισμού** (Constellation Schema). Στο Σχήμα 4.5 απεικονίζεται το Σχήμα Αστερισμού.



Σχήμα 4.5 *Σχήμα Αστερισμού*

Γιατί δεν ήταν ικανοποιητικό το σχεσιακό μοντέλο και χρειάστηκε να προταθεί άλλο σχήμα, ειδικά για τις Αποθήκες Δεδομένων, θα μπορούσε να αναρωτηθεί κανείς. Η απάντηση σχετίζεται με το ζήτημα των επιδόσεων του συστήματος. Το σχεσιακό μοντέλο, για να κανονικοποιήσει τα δεδομένα, δημιουργεί πολλούς αλληλοσυνδεόμενους πίνακες. Η ανάκτηση της πληροφορίας απαιτεί διαδοχικές πράξεις συνένωσης πινάκων (joins), οι οποίες καθυστερούν το σύστημα. Το σχήμα Αστέρα επιτρέπει τον πλεονασμό των δεδομένων, ώστε να επιτύχει ταχύτερη ανάκτηση της πληροφορίας. Επιπλέον, τοποθετώντας το αντικείμενο έρευνας στο επίκεντρο του σχεδιασμού, δομούνται τα δεδομένα με βάση τις ανάγκες της ανάλυσης, καθίσταται δυνατή η απευθείας πρόσβαση στα δεδομένα, αποκλείονται μη ενδιαφέροντα δεδομένα και απλοποιούνται τα ερωτήματα. Να σημειωθεί επίσης ότι το σχήμα Αστέρα είναι προτιμότερο από το σχήμα Χιονονιφάδας, ακριβώς διότι επιτρέπει την ταχύτερη ανάκτηση της πληροφορίας.

Σύμφωνα με τον Inmon (1996), το σημαντικότερο ζήτημα κατά τον σχεδιασμό μιας Αποθήκης Δεδομένων είναι ο καθορισμός του βαθμού κόκκωσης των δεδομένων. Ο όρος **κόκκωση** (granularity) σημαίνει τον βαθμό λεπτομέρειας που τηρείται στο σύστημα. Χαμηλός βαθμός κόκκωσης σημαίνει ότι τα δεδομένα είναι πολύ λεπτομερή. Αντιθέτως, υψηλός βαθμός κόκκωσης σημαίνει ότι τηρούνται πιο γενικευμένα δεδομένα. Η τήρηση πολύ λεπτομερών δεδομένων αυξάνει τον όγκο της αποθηκευμένης πληροφορίας, απαιτεί αυξημένη υπολογιστική ισχύ και προκαλεί καθυστερήσεις στο σύστημα. Ως αντιστάθμισμα, προσφέρει αυξημένες δυνατότητες ανάλυσης, αφού η επεξεργασία των δεδομένων μπορεί να προχωρήσει μέχρι τις λεπτομέρειες. Αντιθέτως, ο υψηλός βαθμός κόκκωσης εξοικονομεί αποθηκευτικό χώρο και υπολογιστική ισχύ, επιταχύνει τη λειτουργία του συστήματος, θέτει όμως περιορισμούς στις αναλυτικές δυνατότητες, εφόσον διατίθεται λιγότερη λεπτομέρεια για επεξεργασία. Αν μια επιχείρηση τηρεί ημερήσια στοιχεία πωλήσεων, τότε θα διογκωθούν τα δεδομένα και θα επιβραδυνθεί το σύστημα, θα μπορεί όμως να αναλύσει τις πωλήσεις ανά ημέρα, εβδομάδα, μήνα κλπ. Εάν η επιχείρηση τηρεί μηνιαία στοιχεία πωλήσεων, θα περιορίσει τον όγκο των αποθηκευμένων δεδομένων και θα επιταχύνει το σύστημα, δεν θα μπορεί όμως να αναλύσει τη διακύμανση των πωλήσεων ανά εβδομάδα. Για τον λόγο αυτό, το ισοζύγιο μεταξύ βαθμού λεπτομέρειας των δεδομένων και αποδοτικής λειτουργίας του συστήματος πρέπει να οριστεί προσεκτικά. Διάφοροι παράγοντες πρέπει να ληφθούν υπόψη. Την Αποθήκη Δεδομένων χρησιμοποιούν στελέχη από διαφορετικά επίπεδα διοίκησης. Τα στελέχη αυτά έχουν διαφορετικές ανάγκες πληροφόρησης, και συνήθως τα ανώτερα διοικητικά επίπεδα χρειάζονται πιο γενικευμένη πληροφορία. Επίσης, διαφορετικά τμήματα του οργανισμού χρειάζονται διαφορετική πληροφόρηση. Για το τμήμα προμηθειών τα μηνιαία στοιχεία πωλήσεων μπορεί να είναι αρκετά, ενώ το τμήμα μάρκετινγκ πιθανώς να χρειάζεται εβδομαδιαία στοιχεία, ώστε να εκτιμά τη διακύμανση της ζήτησης σε περιόδους εορτών. Επίσης, οι γενικές απαιτήσεις του συστήματος δεν είναι σταθερές και μπορεί να μεταβληθούν με τον χρόνο. Μια αλλαγή στις κανονιστικές διατάξεις που διέπουν τη λειτουργία του οργανισμού, πχ. νέα νομοθεσία, πιθανώς να απαιτεί την τήρηση πληροφορίας με μεγαλύτερο βαθμό λεπτομέρειας. Όλοι οι παραπάνω παράγοντες πρέπει

να συνυπολογιστούν ώστε να επιλεγεί το επίπεδο κόκκωσης των δεδομένων που είναι κατάλληλο για την περίσταση. Ο Inmon αναφέρει και την πιθανή λύση να τηρούνται τα δεδομένα ταυτόχρονα σε δύο διαφορετικούς βαθμούς γενίκευσης.

4.5 Πολυδιάστατο Μοντέλο Δεδομένων - Κύβοι

Ο συνήθης χρήστης είναι εξοικειωμένος με τα υπολογιστικά φύλλα και τους πίνακες των σχεσιακών βάσεων δεδομένων. Οι πίνακες αυτοί είναι δισδιάστατοι. Οι Αποθήκες Δεδομένων και οι πράξεις OLAP βασίζονται σε ένα πολυδιάστατο μοντέλο δεδομένων, πολυδιάστατες δηλαδή δομές που αποκαλούνται **κύβοι**. Οι κύβοι είναι ένας τρόπος προβολής των δεδομένων με διάφορα κριτήρια. Ένας κύβος ορίζεται από τις διαστάσεις του και τις τιμές που βρίσκονται στα κελιά του. Οι διαστάσεις του κύβου αναφέρονται στα κριτήρια και αντιστοιχούν στους πίνακες διαστάσεων του σχήματος Αστέρα. Το περιεχόμενο των κελιών του κύβου είναι οι τιμές οι οποίες προέρχονται από τα μέτρα του πίνακα Συμβάντων. Με τον όρο «κύβος» εννοούμε συνήθως τον τρισδιάστατο γεωμετρικό κύβο. Στην περίπτωση όμως των Αποθηκών Δεδομένων εννοούμε δομές, οι οποίες έχουν ν διαστάσεις. Βεβαίως, η ανθρώπινη νόηση δυσκολεύεται να αντιληφθεί χώρους με περισσότερες από τρεις διαστάσεις.

Για την καλύτερη κατανόηση του θέματος ας θεωρήσουμε αρχικά έναν κύβο 2 διαστάσεων. Αυτό είναι ισοδύναμο με τον γνωστό δισδιάστατο πίνακα. Μία επιχείρηση, για παράδειγμα, θέλει να αναλύσει τις πωλήσεις της ανά γεωγραφική περιοχή και ανά τρίμηνο. Ο κύβος που θα προκύψει παρουσιάζεται στο Σχήμα 4.6

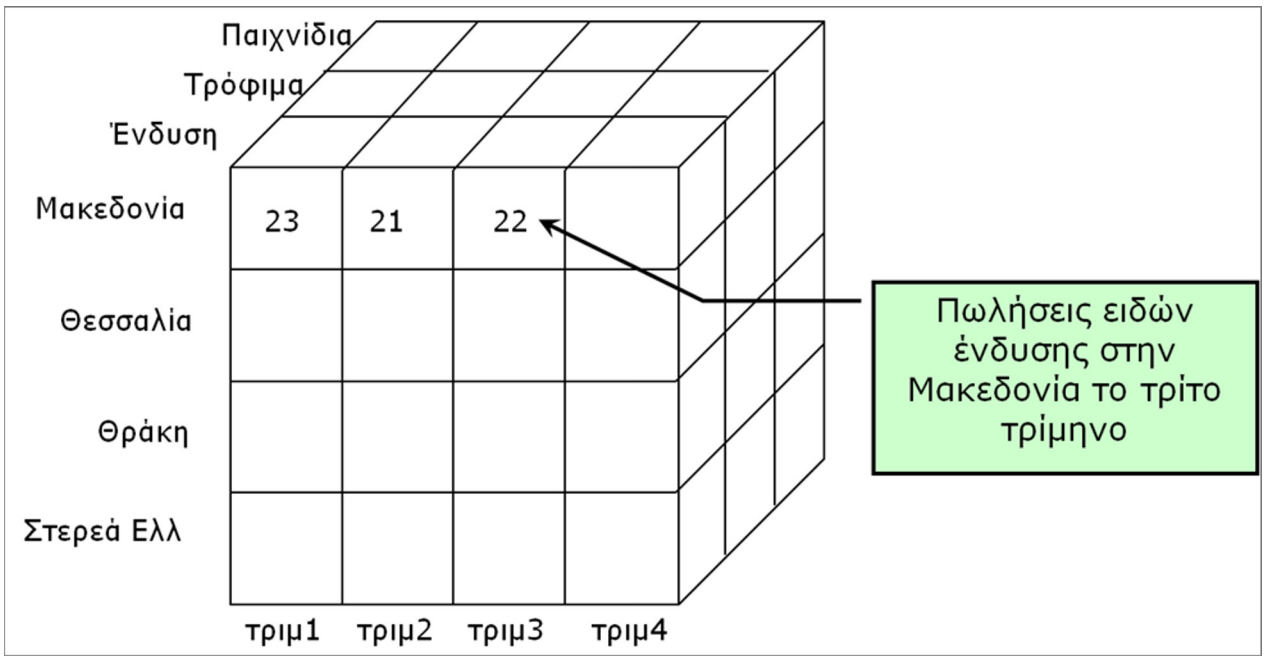
	Μακεδονία	Θεσσαλία	Θράκη
Τρίμηνο 1	124	70	47
Τρίμηνο 2	110	65	42
Τρίμηνο 3	134	64	44
Τρίμηνο 4	140	89	52

Σχήμα 4.6 *Κύβος πωλήσεων με 2 διαστάσεις.*

Όπως φαίνεται στο Σχήμα 4.6, οι στήλες αντιστοιχούν στη γεωγραφική περιοχή, ενώ οι γραμμές στο τρίμηνο. Οι τιμές που βρίσκονται στα κελιά δηλώνουν το ύψος των πωλήσεων για το συγκεκριμένο ζευγάρι τιμών περιοχής και τριμήνου. Έτσι το κελί που προκύπτει από την τομή της στήλης «Μακεδονία» και της γραμμής «Τρίμηνο 1» και ισούται με «124» είναι το ύψος των πωλήσεων στη Μακεδονία το πρώτο τρίμηνο του έτους.

Ας υποθέσουμε τώρα ότι θέλουμε να αναλύσουμε το ύψος των πωλήσεων με βάση τρία κριτήρια, δηλαδή το τρίμηνο, την περιοχή και την κατηγορία προϊόντος. Θα προκύψει τότε ένας τρισδιάστατος κύβος, που για διαστάσεις θα έχει τα τρία κριτήρια, και για περιεχόμενα κελιών θα έχει τις πωλήσεις για τους εκάστοτε συνδυασμούς τιμών των διαστάσεων. Ο τρισδιάστατος κύβος παρουσιάζεται στο Σχήμα 4.7. Η δομή αυτή μπορεί να γίνει κατανοητή ως μια στοίβα δισδιάστατων πινάκων κατά μήκος της τρίτης διάστασης. Βασιζόμενοι στο προηγούμενο παράδειγμα του δισδιάστατου πίνακα, μπορούμε να θεωρήσουμε τον τρισδιάστατο κύβο ως μια στοίβα δισδιάστατων πινάκων, με διαστάσεις την περιοχή και το τρίμηνο κατά μήκος της διάστασης κατηγορία προϊόντος. Φυσικά, τα ποσά στα κελιά επιμερίζονται ανάλογα. Ο χρήστης του συστήματος χρησιμοποιεί τον τρισδιάστατο κύβο υποβάλλοντας ερωτήσεις και λαμβάνοντας απαντήσεις. Εάν για παράδειγμα, ζητήσει να δει το ύψος των πωλήσεων για τα είδη ένδυσης στη Μακεδονία το τρίτο τρίμηνο, θα λάβει ως απάντηση τον αριθμό «22».

Οι κύβοι των Αποθηκών Δεδομένων μπορούν να έχουν ν διαστάσεις. Αναφέρθηκε ότι ο ανθρώπινος εγκέφαλος αδυνατεί να κατανοήσει χώρους τεσσάρων ή περισσοτέρων διαστάσεων. Σύμφωνα όμως με το σκεπτικό που αναπτύχθηκε παραπάνω, μπορούμε να θεωρήσουμε έναν κύβο ν διαστάσεων ως μια στοίβα κύβων *ν-1* διαστάσεων διατεταγμένων κατά μήκος της νιοστής διάστασης. Ένας κύβος τεσσάρων διαστάσεων μπορεί να γίνει κατανοητός ως μια στοίβα τρισδιάστατων κύβων.

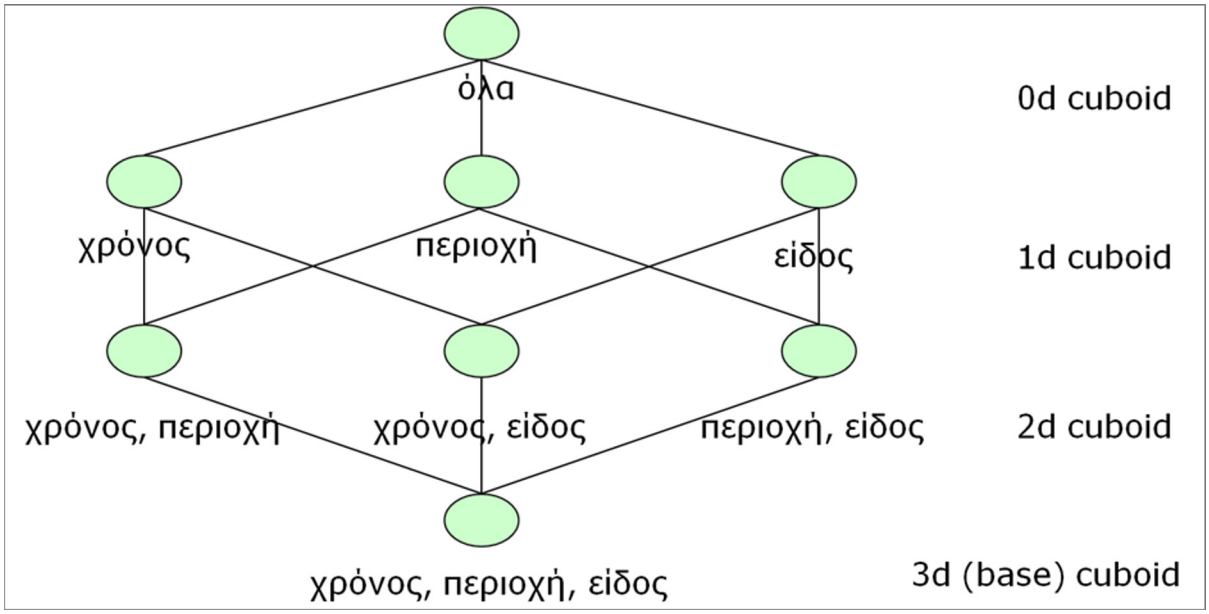


Σχήμα 4.7 *Κύβος πωλήσεων με 3 διαστάσεις*

4.5.1 Κυβοειδή

Όπως έγινε σαφές με τα παραπάνω παραδείγματα, μπορούμε σε ένα πολυδιάστατο μοντέλο να ενεργοποιούμε ή να απενεργοποιούμε διαστάσεις και να δημιουργούμε δομές που περιλαμβάνουν διαφορετικό επίπεδο γενίκευσης. Το δισδιάστατο μοντέλο, με διαστάσεις την περιοχή και το τρίμηνο, μπορεί να προκύψει από το τρισδιάστατο μοντέλο με διαστάσεις την περιοχή, το τρίμηνο και την κατηγορία προϊόντος, αν συναθροίσουμε τα δεδομένα ως προς την κατηγορία προϊόντος, αν δηλαδή υπολογίσουμε για κάθε ζευγάρι μιας περιοχής και ενός τριμήνου το άθροισμα των πωλήσεων για τις τρεις κατηγορίες προϊόντων. Το τρισδιάστατο μοντέλο μπορεί βεβαίως να συναθροιστεί και ως προς την περιοχή ή τον χρόνο, δημιουργώντας άλλες δισδιάστατες δομές.

Γενικώς από ένα μοντέλο ν διαστάσεων μπορούμε εύκολα να κατασκευάσουμε άλλα μοντέλα με λιγότερες διαστάσεις. Τα μοντέλα αυτά ονομάζονται **κυβοειδή** (cuboids). Το μοντέλο που έχει όλες τις διαστάσεις ονομάζεται **βασικό κυβοειδές** (base cuboid) και είναι αυτό που προσφέρει τον μέγιστο βαθμό λεπτομέρειας. Μπορούμε να θεωρήσουμε ένα μοντέλο με μηδενικές διαστάσεις. Το μοντέλο αυτό ονομάζεται κορυφαίο κυβοειδές (apex cuboid) και προσφέρει τον μέγιστο βαθμό γενίκευσης. Στην περίπτωση του παραδείγματος μας, το **κορυφαίο κυβοειδές** μας δίνει το σύνολο των πωλήσεων, το άθροισμα δηλαδή των πωλήσεων για όλες τις περιοχές, όλα τα τρίμηνα και όλες τις κατηγορίες προϊόντος. Μεταξύ του κορυφαίου και του βασικού κυβοειδούς, δημιουργείται ένα πλέγμα κυβοειδών με ενδιάμεσο αριθμό διαστάσεων και σε διάφορους συνδυασμούς. Το πλέγμα των κυβοειδών αποτελεί τον κύβο των δεδομένων. Στο Σχήμα 4.8 παρουσιάζεται το πλέγμα των κυβοειδών, όπου βασικό κυβοειδές είναι το τρισδιάστατο μοντέλο του παραδείγματος μας.



Σχήμα 4.8 *Πλέγμα κυβοειδών*

Ο χρήστης πλοηγείται μέσα στο πλέγμα, προβάλλει κυβοειδή και λαμβάνει πληροφόρηση. Μερικά παραδείγματα ερωτήσεων, που αντλούν πληροφόρηση από τα κυβοειδή, παρατίθενται αμέσως μετά:

- Παρουσίασε τα σύνολα πωλήσεων ανά περιοχή, τρίμηνο και κατηγορία προϊόντων (κυβοειδές τριών διαστάσεων).
- Παρουσίασε τα σύνολα πωλήσεων ανά περιοχή και τρίμηνο (κυβοειδές δύο διαστάσεων (πίνακας) – συνάθροιση στοιχείων ως προς το προϊόν).
- Παρουσίασε τα σύνολα πωλήσεων ανά περιοχή και κατηγορία προϊόντων (κυβοειδές δύο διαστάσεων (πίνακας) - συνάθροιση στοιχείων ως προς τη χρονική περίοδο).
- Παρουσίασε τα σύνολα πωλήσεων ανά τρίμηνο (κυβοειδές μιας διάστασης (γραμμή) - συνάθροιση στοιχείων ως προς το προϊόν και την περιοχή).
- Παρουσίασε τα σύνολα πωλήσεων ανά κατηγορία προϊόντων (κυβοειδές μίας διάστασης (γραμμή) - συνάθροιση στοιχείων ως προς την περιοχή και τη χρονική περίοδο).
- Παρουσίασε το σύνολο πωλήσεων (κυβοειδές 0 διαστάσεων (κελί) – σύνολο πωλήσεων).

Ένα αντικείμενο που χρησιμοποιείται στις ΑΔ είναι οι λεγόμενες **Υλοποιημένες Όψεις** (Materialized Views). Ο όγκος των δεδομένων σε μια ΑΔ είναι πολύ μεγάλος και ο υπολογισμός συγκεντρωτικών τιμών μπορεί να αποδειχθεί εξαιρετικά χρονοβόρος. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι να αποθηκευτούν εκ των προτέρων οι συγκεντρωτικές τιμές. Οι σύγχρονες βάσεις δεδομένων διαθέτουν τις Υλοποιημένες Όψεις, που είναι ουσιαστικά αποθηκεύσεις των αποτελεσμάτων ερωτήσεων (queries) σε ειδικούς πίνακες. Οι Υλοποιημένες Όψεις αναφέρονται και ως περιλήψεις (summaries), γιατί περιέχουν περιληπτικά στοιχεία.

Δεδομένου ότι από ένα βασικό κυβοειδές μπορούν να παραχθούν πολλά άλλα κυβοειδή, όπως φαίνεται και στο Σχήμα 4.8, γεννάται το ερώτημα ποια από αυτά πρέπει να υπολογιστούν. Πλήρης υλοποίηση (full materialization) σημαίνει ότι υπολογίζονται και αποθηκεύονται όλα τα κυβοειδή. Η προσέγγιση αυτή εξασφαλίζει αυξημένες επιδόσεις πρόσβασης στην πληροφορία, απαιτεί όμως πολύ μεγάλους αποθηκευτικούς χώρους. Μερική υλοποίηση (partial materialization) σημαίνει ότι υπολογίζονται ορισμένα μόνο από τα κυβοειδή του πλέγματος. Η επιλογή των κυβοειδών που θα υλοποιηθούν μπορεί να γίνει εμπειρικά και να βασιστεί στη συχνότητα χρήσης του κυβοειδούς ή στον όγκο των δεδομένων που απαιτείται για τον υπολογισμό τους. Ένας εναλλακτικός τρόπος είναι να επιλεγούν οι υλοποιημένες όψεις με τη βοήθεια κάποιου αλγορίθμου. Έχουν προταθεί διάφοροι αλγόριθμοι για την επιλογή των υλοποιημένων όψεων, όπως ο αλγόριθμος των Harinaryan, Rajaraman and Ulman (1996), καθώς και ο αλγόριθμος των Gupta and Mumick (2005).

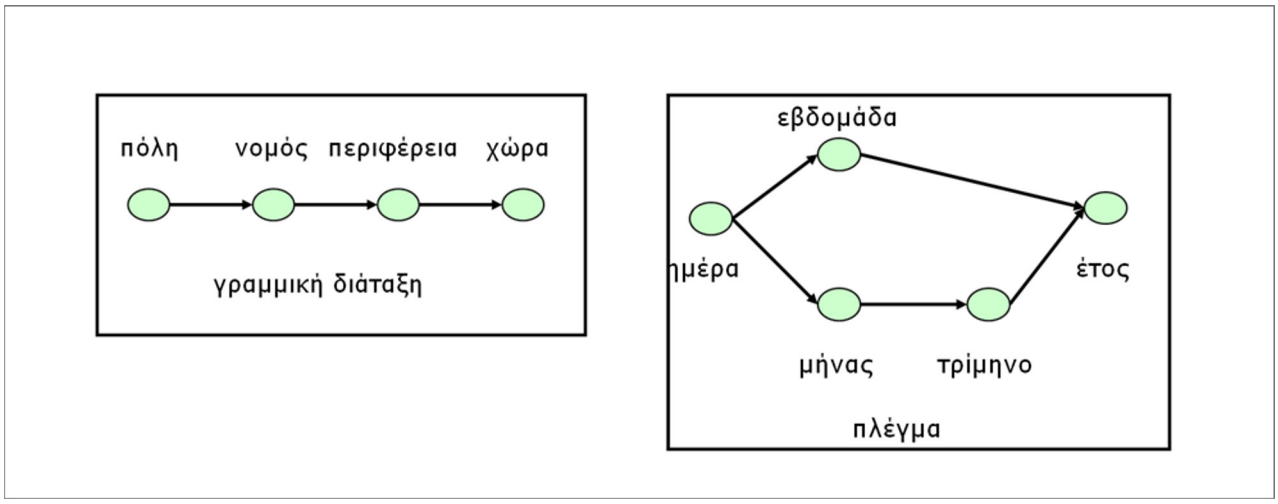
Οι αποθηκευμένες όψεις είναι πολλαπλώς χρήσιμες. Καταρχήν, εξασφαλίζουν γρήγορη πρόσβαση στην πληροφορία για ερωτήματα, που η εκτέλεση τους θα απαιτούσε υπερβολικά πολύ χρόνο. Επίσης, μπορούν να χρησιμοποιηθούν για τη δημιουργία άλλων ερωτημάτων και όψεων. Ένα πρόβλημα σχετικό με τις όψεις είναι

η συντήρηση τους. Οι Αποθήκες Δεδομένων εμπλουτίζονται σε τακτά χρονικά διαστήματα με νέα δεδομένα. Κατά πάσα πιθανότητα τα νέα δεδομένα έχουν σχέση με τα σύνολα που υπάρχουν στις όψεις, οπότε οι όψεις πρέπει να επαναϋπολογιστούν. Οι σύγχρονες βάσεις δεδομένων δίνουν τη δυνατότητα αυτόματης ενημέρωσης των όψεων, όταν τροποποιούνται εγγραφές στους πηγαίους πίνακες.

4.6 Ιεραρχίες Εννοιών

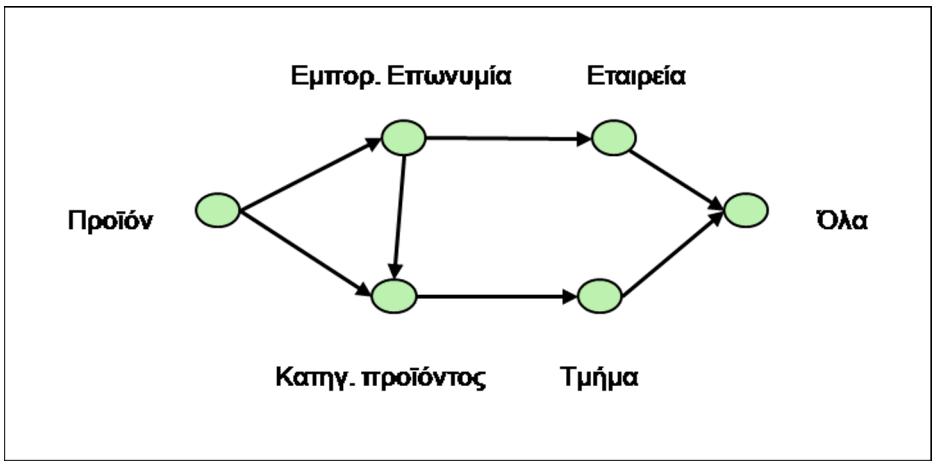
Σημαντικό ρόλο στις Αποθήκες Δεδομένων και στο πολυδιάστατο μοντέλο παίζουν οι λεγόμενες ιεραρχίες εννοιών. Μια **ιεραρχία εννοιών** είναι μια διάταξη εννοιών σύμφωνα με τον βαθμό γενίκευσης, από το ειδικότερο προς το γενικότερο. Ένα κλασικό παράδειγμα είναι η τοποθεσία, η οποία μπορεί να οριστεί μέσω της πόλης, του νομού, της περιφέρειας και της χώρας. Προφανώς η πόλη εντάσσεται στον νομό, ο νομός στην περιφέρεια κλπ. και με τον τρόπο αυτό υπάρχει μια κλιμάκωση του βαθμού γενίκευσης. Η πόλη, ο νομός, η περιφέρεια και η χώρα συγκροτούν μια ιεραρχία εννοιών.

Οι ιεραρχίες εννοιών μπορεί να είναι γραμμικά διατεταγμένες. Στη γεωγραφική ιεραρχία πόλη-χώρα υπάρχει μια μοναδική διαδρομή από τον αρχικό στον τελικό κόμβο. Ωστόσο, υπάρχουν περιπτώσεις όπου ορίζονται περισσότερες από μια ανεξάρτητες διαδρομές από την αρχή μέχρι το τέλος, και οι οποίες διαμορφώνουν ένα πλέγμα. Ας θεωρήσουμε ως παράδειγμα τη χρονική διάσταση. Υπάρχει μια ιεραρχία από την ημέρα έως το έτος, η διαδρομή όμως δεν είναι μοναδική. Η ημέρα εντάσσεται στον μήνα, ο μήνας στο τρίμηνο και το τρίμηνο στο έτος, δημιουργώντας μια διαδρομή. Ταυτόχρονα όμως, η μέρα εντάσσεται στην εβδομάδα, η εβδομάδα όμως δεν εντάσσεται στον μήνα. Δημιουργείται έτσι μια δεύτερη διαδρομή ημέρα < εβδομάδα < έτος και οι δύο διαδρομές διαμορφώνουν ένα πλέγμα. Τα παραπάνω απεικονίζονται διαγραμματικά στο Σχήμα 4.9



Σχήμα 4.9 *Ιεραρχίες Εννοιών*

Οι ιεραρχίες εννοιών μπορεί να είναι ακόμη πιο περίπλοκες. Ας θεωρήσουμε μια ιεράρχηση στην κατηγοριοποίηση των προϊόντων. Το υποθετικό προϊόν MiracleTV ανήκει στην κατηγορία προϊόντων Τηλεόραση Πλάσμα, η οποία κατηγορία πωλείται στο τμήμα Ηλεκτρικές Συσκευές. Ορίζεται λοιπόν μια ιεραρχία προϊόν < κατηγορία προϊόντος < τμήμα < όλα τα προϊόντα. Ταυτόχρονα, η MiracleTV είναι μοντέλο της εμπορικής επωνυμίας SuperBrand, που ανήκει στο βιομηχανικό συγκρότημα General Industries. Μια νέα ιεραρχία για τα προϊόντα είναι προϊόν < εμπορική επωνυμία < εταιρεία < όλα τα προϊόντα. Με δεδομένο όμως ότι μια κατηγορία προϊόντων περιλαμβάνει προϊόντα πολλών εμπορικών επωνυμιών, νοείται και μια ιεραρχία προϊόν < εμπορική επωνυμία < κατηγορία προϊόντος < τμήμα < όλα τα προϊόντα. Τα παραπάνω απεικονίζονται διαγραμματικά στο Σχήμα 4.10 και αποτελούν μια περίπτωση ετερογένειας διάστασης, που απαιτεί ειδική αντιμετώπιση. Ο αναγνώστης μπορεί να βρει λεπτομέρειες στο Hurtado and Gutierrez (2007).



Σχήμα 4.10 *Ετερογένεια Διάστασης*

Ιεραρχίες εννοιών μπορεί να προκύψουν και από αριθμητικά πεδία, εάν τα δεδομένα διακριτοποιηθούν, και για κάθε περιοχή τιμών οριστούν υποπεριοχές. Οι ιεραρχίες εννοιών που αφορούν μια διάσταση, μπορούν να χρησιμοποιηθούν για τη συνάθροιση και την προβολή των δεδομένων σε διαφορετικά επίπεδα γενίκευσης. Αυτό επιτυγχάνεται με τη βαθμονόμηση του αντίστοιχου άξονα σύμφωνα με το επιθυμητό επίπεδο της ιεραρχίας εννοιών. Στο παράδειγμα του Σχήματος 4.11 παρουσιάζονται οι πωλήσεις κατά τρίμηνο ή κατά μήνα.

Τρίμηνο 1	Τρίμηνο 2	Τρίμηνο 3	Τρίμηνο 4
80	70	80	90

Ιαν	Φεβ	Μαρ	Απρ	Μάι	Ιουν	Ιουλ	Αυγ	Σεπ	Οκτ	Νοε	Δεκ
30	30	20	20	20	30	20	30	30	20	30	40

Σχήμα 4.11 *Συνάθροιση δεδομένων με ιεραρχία εννοιών.*

4.7 Πράξεις OLAP

Όπως έχει αναφερθεί μέχρι τώρα, το πολυδιάστατο μοντέλο δεδομένων οργανώνει κεντρικές έννοιες σύμφωνα με διαστάσεις. Επίσης, κάθε διάσταση μπορεί να περιλαμβάνει ιεραρχίες εννοιών, οι οποίες επιτρέπουν την αναδιοργάνωση των δεδομένων σε διάφορα επίπεδα γενίκευσης. Το πολυδιάστατο μοντέλο είναι κατάλληλο για διαδραστική ανάλυση των δεδομένων. Ο χρήστης, χρησιμοποιώντας τις διαστάσεις και τις ιεραρχίες εννοιών, ανασυγκροτεί τα δεδομένα, τα συναθροίζει ή τα επιμερίζει, τα προβάλλει με διαφορετικούς τρόπους, αποκόπτει τμήματα τους, και γενικώς τα διερευνά, ώστε να πληροφορηθεί σχετικά με το ύψος διαφορετικών συνόλων, να κάνει συγκρίσεις, να εντοπίσει ακραίες ή ιδιόμορφες τιμές και γενικώς να αντλήσει πληροφορία, που θα τη χρησιμοποιήσει για τη λήψη αποφάσεων.

Η διεξαγωγή αναλύσεων OLAP γίνεται με τη βοήθεια κατάλληλων πράξεων, που επιτρέπουν την πλοήγηση στο πλέγμα των κυβοειδών και την υλοποίηση συναθροίσεων, επιμερισμών και άλλων τρόπων προβολής των δεδομένων. Οι βασικές πράξεις OLAP είναι οι:

- Συναθροιστική Άνοδος (Roll up),
- Αναλυτική Κάθοδος (Drill down),
- Οριζόντιος Τεμαχισμός (Slice),
- Κάθετος Τεμαχισμός (Dice),
- Περιστροφή (Pivot).

4.7.1 Συναθροιστική Άνοδος

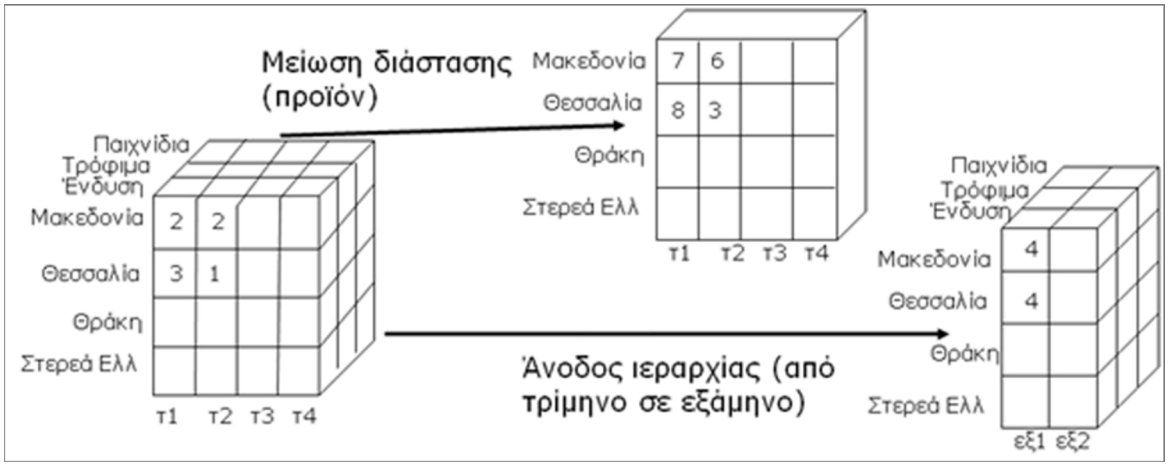
Η Συναθροιστική Άνοδος είναι η αναδιοργάνωση των δεδομένων σε μεγαλύτερο επίπεδο γενίκευσης, η συνάθροιση τους δηλαδή σε ευρύτερα σύνολα. Στο πολυδιάστατο μοντέλο, αυτό μπορεί να επιτευχθεί με δύο τρόπους:

- Με άνοδο στην ιεραρχία εννοιών μιας διάστασης,
- Με μείωση των διαστάσεων.

Με την άνοδο στην ιεραρχία εννοιών, ο χρήστης επιλέγει να βαθμονομήσει τον άξονα της διάστασης με ένα γενικότερο μέτρο, οπότε τα δεδομένα ομαδοποιούνται και συναθροίζονται σε ευρύτερες έννοιες από αυτές του τρέχοντος κυβοειδούς. Για παράδειγμα, αν στο αρχικό κυβοειδές η τοποθεσία ορίζεται με βάση την πόλη, με μια πράξη roll up μπορούμε να ορίσουμε την τοποθεσία με βάση τον νομό και να λάβουμε συγκεντρωτικά στοιχεία πωλήσεων ανά νομό. Στο παράδειγμα του Σχήματος 4.12 εκτελούμε συναθροιστική άνοδο με άνοδο στην ιεραρχία εννοιών, αλλάζοντας το μέτρο της διάστασης του χρόνου από τρίμηνο σε εξάμηνο και υπολογίζοντας συγκεντρωτικά στοιχεία πωλήσεων ανά εξάμηνο. Σε μορφή ψευδογλώσσας, η αντίστοιχη εντολή είναι:

- Roll-up on Time (from quarter to half-year).

Με τη μείωση διαστάσεων ο χρήστης αφαιρεί από το κυβοειδές μια από τις διαστάσεις. Τα δεδομένα τότε ομαδοποιούνται και συναθροίζονται για όλες τις τιμές της διάστασης που αφαιρέθηκε. Στο παράδειγμα του Σχήματος 4.12 αφαιρείται η διάσταση «κατηγορία προϊόντων» από τον αρχικό κύβο. Ο νέος κύβος παρουσιάζει τα σύνολα πωλήσεων ανά γεωγραφική περιοχή και τρίμηνο. Σε κάθε κελί του κύβου, το οποίο αντιστοιχεί σε μια περιοχή και ένα τρίμηνο, παρουσιάζονται τα σύνολα πωλήσεων για όλες τις κατηγορίες προϊόντων.



Σχήμα 4.12 *Συναθροιστική Άνοδος*

Η εντολή σε μορφή ψευδογλώσσας είναι

- Roll-up on Product-Category (from individual to all).

4.7.2 Αναλυτική Κάθοδος

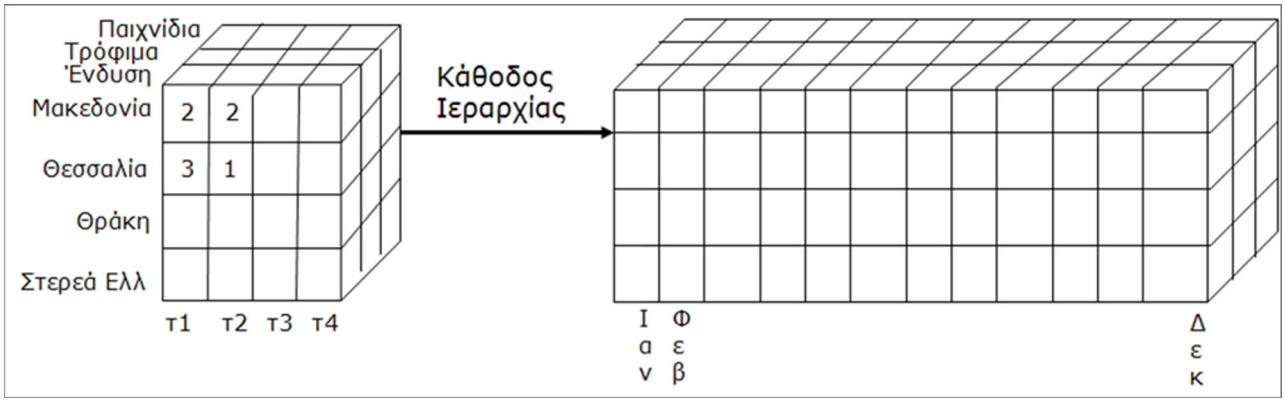
Η Αναλυτική Κάθοδος είναι η αντίστροφη πράξη από τη Συναθροιστική Άνοδο. Αποτελεί τη μετάβαση από δεδομένα υψηλής γενίκευσης σε δεδομένα αυξημένης λεπτομέρειας. Τα δεδομένα χωρίζονται σε μικρότερα σύνολα και γίνεται επιμερισμός των αρχικών συνολικών τιμών. Και η Αναλυτική Κάθοδος μπορεί να επιτευχθεί με δύο τρόπους:

- με κάθοδο στην ιεραρχία εννοιών μιας διάστασης,
- με αύξηση των διαστάσεων.

Με την κάθοδο στην ιεραρχία εννοιών μιας διάστασης, ο αντίστοιχος άξονας βαθμονομείται σύμφωνα με μέτρο ειδικότερο από το τρέχων. Αν για παράδειγμα, η τοποθεσία στο τρέχων κυβοειδές ορίζεται μέσω της περιοχής, με την πράξη της αναλυτικής καθόδου η τοποθεσία ορίζεται μέσω της πόλης. Τα δεδομένα που αντιστοιχούν σε μια περιοχή διασπώνται σύμφωνα με την πόλη, και το σύνολο της περιοχής επιμερίζεται. Στο παράδειγμα του Σχήματος 4.13, στο αρχικό κυβοειδές παρουσιάζονται στοιχεία πωλήσεων ανά τρίμηνο. Στη συνέχεια εκτελείται Αναλυτική Κάθοδος με κάθοδο στην ιεραρχία εννοιών του χρόνου, ο άξονας του χρόνου βαθμονομείται σε μήνες και παρουσιάζονται τα στοιχεία πωλήσεων ανά κατηγορία προϊόντος, περιοχή και μήνα. Σε μορφή ψευδογλώσσας, η εντολή που εκτελεί την αναλυτική κάθοδο του σχήματος 4.13 είναι

- Drill-down on Time (from quarter to month).

Ο δεύτερος τρόπος για Αναλυτική Κάθοδο είναι με την αύξηση των διαστάσεων. Σε ένα δισδιάστατο κυβοειδές, που παρουσιάζει στοιχεία πωλήσεων ανά κατηγορία προϊόντος και περιοχή, προστίθεται η διάσταση του χρόνου. Το νέο κυβοειδές παρουσιάζει στοιχεία πωλήσεων ανά κατηγορία προϊόντος, περιοχή και χρονική περίοδο, πχ τρίμηνο. Το σύνολο πωλήσεων για κάθε ζεύγος τιμών περιοχής και κατηγορίας προϊόντος επιμερίζεται σε περισσότερα κελιά, σύμφωνα με τις αντίστοιχες τριμηνιαίες πωλήσεις.



Σχήμα 4.13 *Αναλυτική Κάθοδος με κάθοδο στην ιεραρχία εννοιών του χρόνου*

4.7.3 Οριζόντιος Τεμαχισμός

Με την πράξη slice δημιουργούμε έναν νέο κύβο, επιλέγοντας μια ή περισσότερες τιμές σε μια μόνο διάσταση. Στο παράδειγμα του Σχήματος 4.14, από τον αρχικό τρισδιάστατο κύβο επιλέγουμε από τον άξονα της περιοχής την τιμή «Μακεδονία». Δημιουργείται τότε ένας νέος κύβος, που παρουσιάζει τις πωλήσεις της Μακεδονίας ανά κατηγορία προϊόντος και τρίμηνο. Θα μπορούσαμε να είχαμε επιλέξει δύο τιμές, πχ Μακεδονία και Θεσσαλία και τότε θα είχε δημιουργηθεί ένας νέος τρισδιάστατος κύβος, που θα προέκυπτε από τον αρχικό με αποκοπή των «φετών» που αντιστοιχούν στη Θράκη και τη Στερεά Ελλάδα. Σε μορφή ψευδογλώσσας, η αντίστοιχη εντολή του σχήματος 4.14 είναι

- Slice for Region = «Μακεδονία».

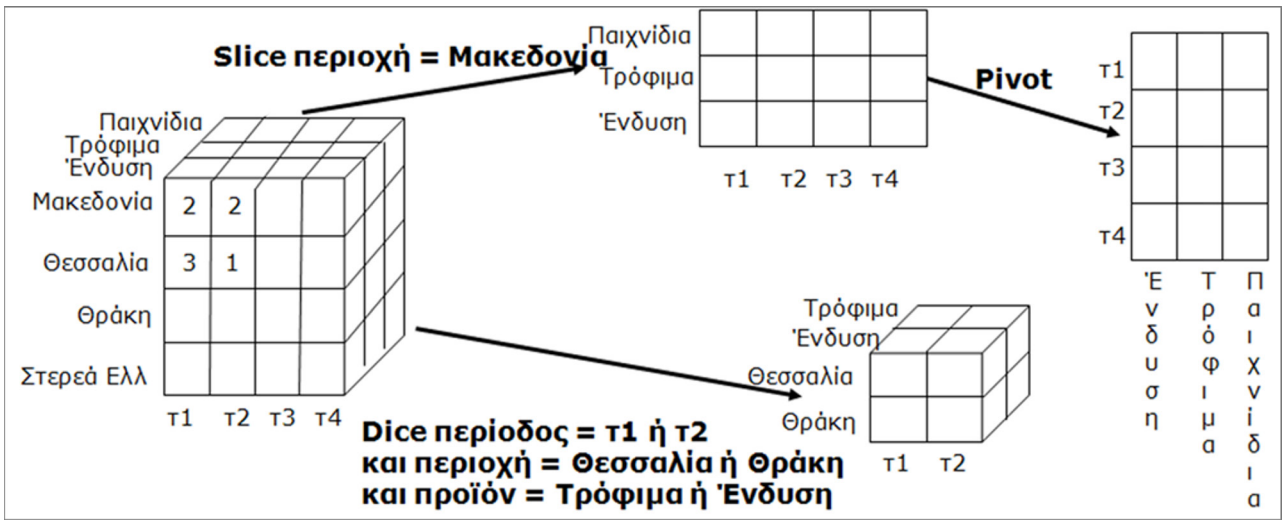
4.7.4 Κάθετος Τεμαχισμός

Με την πράξη dice δημιουργούμε έναν νέο κύβο, επιλέγοντας μια ή περισσότερες τιμές σε δύο ή περισσότερες διαστάσεις. Στο παράδειγμα του Σχήματος 4.14, από τον αρχικό κύβο επιλέγουμε το πρώτο και δεύτερο τρίμηνο για τον χρόνο, τη Θεσσαλία και τη Θράκη από τις περιοχές, καθώς και τα Τρόφιμα και την Ένδυση από την κατηγορία προϊόντων. Δημιουργείται ένας νέος κύβος με τα στοιχεία πωλήσεων αυτών των περιοχών, τριμήνων και προϊόντων. Σε μορφή ψευδογλώσσας η αντίστοιχη εντολή είναι η ακόλουθη:

- Dice for (Product-Category = “Τρόφιμα” or “Ένδυση”) and (Region = “Θεσσαλία” or “Θράκη”) and (Time = “τ1” or “τ2”).

4.7.5 Περιστροφή

Πρόκειται για πράξη αλλαγής της διάταξης των αξόνων, που προκαλεί περιστροφή του κύβου και προβολή του από διαφορετική οπτική γωνία. Με την πράξη pivot δεν απαιτείται κανένας νέος υπολογισμός, ούτε αυξάνονται ή μειώνονται τα δεδομένα. Στο παράδειγμα του Σχήματος 4.14, ο κύβος με τα στοιχεία πωλήσεων της Μακεδονίας περιστρέφεται και γίνεται αντιμετάθεση των αξόνων του χρόνου και της κατηγορίας προϊόντων.



Σχήμα 4.14 *Οι πράξεις Slice, Dice και Pivot*

4.8 Καθοδηγούμενη Διερεύνηση

Οι πολυδιάστατοι κύβοι και οι πράξεις OLAP είναι μια μέθοδος οργάνωσης και προβολής των δεδομένων, η οποία επιτρέπει στον χρήστη να συναθροίσει τα δεδομένα με πολλαπλούς τρόπους, να συγκρίνει τιμές και να αντλήσει συμπεράσματα. Ο χρήστης, με δική του πρωτοβουλία, υλοποιεί τους κύβους που θεωρεί σημαντικούς και αναζητά σε αυτούς ενδιαφέρουσα πληροφορία. Η επιλογή των κύβων γίνεται με βάση τη γνώση σχετικά με τα επιχειρηματικά ζητήματα, τη διαίσθηση, τη φαντασία και γενικώς υποκειμενικά στοιχεία του χρήστη. Ουσιαστικά, ο αναλυτής διατυπώνει εκ των προτέρων υποθέσεις, και στη συνέχεια αναζητά την επιβεβαίωση ή τη διάψευση των υποθέσεων μέσα από τα δεδομένα. Μια τέτοιου τύπου ανάλυση ονομάζεται διερεύνηση καθοδηγούμενη από υποθέσεις.

Η καθοδηγούμενη από υποθέσεις διερεύνηση έχει αδυναμίες. Η πρώτη αδυναμία αφορά τον όγκο της πληροφορίας που πρέπει να ελέγξει ο αναλυτής. Ένας πλήρης κύβος περιλαμβάνει πολλά κυβοειδή, όπως φαίνεται και στο Σχήμα 4.8. Κάθε ένα από αυτά είναι δυνατόν να έχει αρκετά μέτρα και πολλές τιμές. Είναι εξαιρετικά δύσκολο για τον αναλυτή να ελέγξει όλα τα κυβοειδή. Ακόμα και όταν απομονώσει ορισμένα κελιά με εφαρμογή των πράξεων OLAP, το πλήθος των δεδομένων μπορεί να είναι πολύ μεγάλο. Η δεύτερη αδυναμία αφορά την πιθανότητα να κρύβονται ενδιαφέρουσες λεπτομέρειες μέσα σε γενικευμένες πληροφορίες. Για παράδειγμα, μια μεγάλη πτώση των πωλήσεων για μια συγκεκριμένη κατηγορία προϊόντων μπορεί να ισοσταθμίζεται από μια μικρή αύξηση πωλήσεων σε άλλες κατηγορίες προϊόντων. Αν ο χρήστης μελετήσει τις πωλήσεις ανά γεωγραφική περιοχή και χρονική περίοδο, τότε θα του διαφύγει η πτώση πωλήσεων της συγκεκριμένης κατηγορίας.

Για την αντιμετώπιση τέτοιων προβλημάτων έχει προταθεί η **καθοδηγούμενη από την ανακάλυψη διερεύνηση**. Σύμφωνα με αυτήν την τεχνική, το λογισμικό σαρώνει αυτόματα τα κελιά του κύβου κατά μήκος όλων των διαστάσεων και, εφαρμόζοντας στατιστικά κριτήρια, προσπαθεί να εντοπίσει ακραίες και αποκλίνουσες τιμές. Τα κελιά αυτά, αφού εντοπιστούν, υποδεικνύονται στον χρήστη με κάποιον οπτικό τρόπο (πχ με κατάλληλο χρωματισμό), ώστε να επικεντρώσει σε αυτά την προσοχή του. Ένα κελί επισημαίνεται εάν αποτελεί εξαίρεση σε σχέση με άλλα κελιά, που αντιστοιχούν στο ίδιο επίπεδο γενίκευσης, ή εάν υπάρχει εξαίρεση σε πιο λεπτομερή κελιά, που συναθροίζονται στο επισημασμένο κελί. Στη δεύτερη αυτή περίπτωση, ο αναλυτής πρέπει να εκτελέσει πράξη drill down και να μελετήσει τα λεπτομερέστερα κελιά. Έχουν προταθεί ειδικές τεχνικές για τον αυτόματο εντοπισμό των κελιών – εξαιρέσεων (Sarawagi, Agrawal & Megiddo, 1998).

4.9 Πρατήρια Δεδομένων

Στη βιβλιογραφία για τις Αποθήκες Δεδομένων αναφέρεται συχνά ο όρος **Πρατήρια Δεδομένων** (ΠΔ) (Data Marts). Τα ΠΔ είναι μικρές και εξειδικευμένου σκοπού Αποθήκες Δεδομένων. Η βασική λογική, οι σχεδιαστικές αρχές και η χρήση των ΠΔ είναι ίδιες με αυτές των ΑΔ. Ωστόσο, υπάρχουν και αρκετές διαφορές:

- Η ΑΔ είναι ένα σύστημα, που αναφέρεται στη λειτουργία ολόκληρου του οργανισμού και χρησιμο-

ποιείται από πολλά τμήματα του. Το ΠΔ σχετίζεται με τη λειτουργία ενός συγκεκριμένου τμήματος του οργανισμού, πχ το τμήμα πωλήσεων, και χρησιμοποιείται μόνο από αυτό.

- Η ΑΔ καλύπτει πολλά και διαφορετικά αντικείμενα, όπως πωλήσεις, χρηματοοικονομικά, παραγωγή κλπ. Αντιθέτως, τα ΠΔ καλύπτουν συνήθως ένα συγκεκριμένο αντικείμενο, πχ. χρηματοοικονομικά.
- Οι ΑΔ αναπτύσσονται κεντρικά από τον οργανισμό με ευθύνη κάποιου αρμόδιου τμήματος, όπως είναι το τμήμα Πληροφορικής Τεχνολογίας. Τα ΠΔ αναπτύσσονται περιφερειακά από κάποιο επιμέρους τμήμα του οργανισμού, πχ το τμήμα μάρκετινγκ, χρησιμοποιούνται και συντηρούνται από αυτό και ανήκουν σε αυτό.
- Το μέγεθος των ΠΔ είναι σημαντικά μικρότερο από αυτό μιας ΑΔ. Τυπικά μια ΑΔ υπερβαίνει το 1 TB, ενώ ένα πρατήριο δεδομένων είναι συνήθως μικρότερο από 100 GB.
- Για τον σχεδιασμό μιας ΑΔ, η οποία καλύπτει πολλά αντικείμενα, το πιο κατάλληλο σχήμα είναι αυτό του Αστερισμού. Αντιθέτως, στα ΠΔ, που είναι μονοθεματικά, καταλληλότερο σχήμα συνήθως είναι του Αστέρα ή της Χιονονιφάδας.
- Στις ΑΔ ολοκληρώνονται δεδομένα από όλες τις σημαντικές πηγές. Ένα ΠΔ περιλαμβάνει δεδομένα από επιλεγμένες και συνήθως λίγες πηγές, που αφορούν ένα αντικείμενο.
- Συνήθως σε ένα ΠΔ αποθηκεύονται πιο συγκεντρωτικά δεδομένα.
- Η διαδικασία ανάπτυξης και υλοποίησης ενός ΠΔ είναι απλούστερη και ο απαραίτητος χρόνος κυμαίνεται σε λίγους μήνες. Αντιθέτως, ο χρόνος υλοποίησης μιας ΑΔ μπορεί άνετα να υπερβαίνει τον ένα χρόνο.
- Η υλικοτεχνική υποδομή που απαιτείται για ένα ΠΔ είναι σημαντικά χαμηλότερου κόστους.

Τα Πρατήρια Δεδομένων χωρίζονται σε δύο κατηγορίες, τα *εξαρτημένα ΠΔ* (dependent data marts) και τα *ανεξάρτητα ΠΔ* (independent data marts). Τα εξαρτημένα ΠΔ αντλούν τα δεδομένα τους από μια κεντρική Αποθήκη Δεδομένων η οποία προϋπάρχει. Τα ανεξάρτητα ΠΔ αντλούν δεδομένα από πηγαία συστήματα, όπως συστήματα παρακολούθησης συναλλαγών, αρχεία, εξωτερικές πηγές κλπ. με τον ίδιο τρόπο που τροφοδοτούνται και οι ΑΔ. Η διαφορά είναι σημαντική, γιατί στην περίπτωση των εξαρτημένων ΠΔ υπάρχουν έτοιμα ενοποιημένα και ποιοτικά δεδομένα, τα οποία απλώς μεταφέρονται στο ΠΔ. Στην περίπτωση όμως των ανεξάρτητων ΠΔ, πρέπει να εκτελεστούν από την αρχή όλες οι διαδικασίες εξαγωγής, μετασχηματισμού και φόρτωσης, οι οποίες είναι αρκετά περίπλοκες και χρονοβόρες, όπως εξηγείται στο σχετικό υποκεφάλαιο. Τα εξαρτημένα ΠΔ συνήθως κατασκευάζονται για να εξασφαλίσουν σε ένα τμήμα ταχύτερη πρόσβαση και περισσότερο έλεγχο στην πληροφορία. Τα ανεξάρτητα ΠΔ κατασκευάζονται για να καλύψουν ανάγκες, όταν δεν υπάρχει κεντρική ΑΔ και η κατασκευή της κρίνεται αργή ή ασύμφορη.

Τα ΠΔ παίζουν σημαντικό ρόλο και στην ανάπτυξη συστημάτων ΑΔ. Για την ανάπτυξη των ΑΔ υπάρχουν δύο προσεγγίσεις, που αντανακλούν τις διιστάμενες απόψεις δύο γκουρού του χώρου. Σύμφωνα με την προσέγγιση «top down»,η οποία προσιδιάζει στις απόψεις του Bill Inmon, κατασκευάζεται πρώτα η κεντρική ΑΔ, η οποία περιλαμβάνει όλα τα δεδομένα του οργανισμού. Στη συνέχεια μπορεί να κατασκευαστούν τα ΠΔ. Αντιθέτως, η προσέγγιση «bottom-up», η οποία προκρίνεται από τον Ralph Kimball, προβλέπει αρχικά την κατασκευή μικρών ΠΔ, που προσφέρουν ειδική πληροφόρηση για πιο εντοπισμένα αντικείμενα, και στη συνέχεια τον συνδυασμό των ΠΔ και την ολοκλήρωση τους σε μια ΑΔ. Κατά την άποψη του Kimball, «μια Αποθήκη Δεδομένων δεν είναι τίποτα παραπάνω από την ένωση όλων των Πρατηρίων Δεδομένων». Για την ιστορία, αλλά και για την αξία του ευφυολογήματος, παραθέτουμε την απάντηση του Inmon. «Μπορείς να πιάσεις όλα τα ψαράκια του ωκεανού και να τα ενώσεις, αλλά πάλι δεν κάνουν μια φάλαινα».

4.10 Εξαγωγή Μετασχηματισμός Φόρτωση

Όπως έχει ήδη αναφερθεί, οι Αποθήκες Δεδομένων δεν λειτουργούν με δεδομένα τα οποία παράγουν οι ίδιες πρωτογενώς, αλλά χρησιμοποιούν τα δεδομένα άλλων συστημάτων. Όπως απεικονίζεται και γραφικά στο Σχήμα 4.1, τα πηγαία δεδομένα πρέπει να συλλεχθούν, να ομογενοποιηθούν και να φορτωθούν στην ΑΔ. Οι εργασίες αυτές είναι γνωστές ως εργασίες **Εξαγωγής, Μετασχηματισμού και Φόρτωσης** (ΕΜΦ) (Extract, Transform, Load (ETL)), και είναι από τις σημαντικότερες στις ΑΔ. Η όλη διαδικασία δεν είναι καθόλου απλή και έχει να αντιμετωπίσει μια σειρά από προκλήσεις. Για τον λόγο αυτό, είναι η πιο χρονοβόρα διαδικασία κατά την ανάπτυξη και συντήρηση της ΑΔ, και σε αυτήν εμπλέκονται αναλυτές, σχεδιαστές βάσεων δεδομένων και κατασκευαστές λογισμικού. Η πληθώρα των προβλημάτων που πρέπει να αντιμετωπιστούν, και η σημασία τους για την ομαλή λειτουργία της ΑΔ, καθιστούν τις εργασίες ΕΜΦ ένα διακριτό αντικείμενο μελέτης. Πάροχοι λογισμικού, όπως η Oracle, η Microsoft, η IBM και άλλοι, διαθέτουν στην αγορά έτοιμο

λογισμικό ETL, ενώ οι ιδιοκτήτες των ΑΔ έχουν και την επιλογή να κατασκευάσουν τα δικά τους εργαλεία.

Σε γενικές γραμμές οι εργασίες ΕΜΦ περιλαμβάνουν τα παρακάτω στάδια:

- τον εντοπισμό και την εξαγωγή των πηγαίων δεδομένων,
- τη μεταφορά τους σε ειδικό χώρο, όπου θα γίνει η επεξεργασία τους,
- τον μετασχηματισμό και τον καθαρισμό των δεδομένων,
- τη φόρτωση τους στην Αποθήκη Δεδομένων.

Ο καθορισμός μιας πολιτικής ανανέωσης των δεδομένων της ΑΔ είναι αντικείμενο του διαχειριστή της. Στα πλαίσια αυτής της διαδικασίας πρέπει να οριστεί η σειρά εργασιών, καθώς και η ροή των δεδομένων από τα αρχικά συστήματα μέχρι τον τελικό τους προορισμό. Μια σημαντική παράμετρος είναι η διαθεσιμότητα υπολογιστικών και επικοινωνιακών πόρων και η εύρυθμη λειτουργία των υπολογιστικών συστημάτων και των δικτύων. Οι εργασίες ΕΜΦ είναι δυνατόν να προκαλέσουν μεγάλες καθυστερήσεις στη λειτουργία κυρίως των πηγαίων συστημάτων, τα οποία όμως είναι απαραίτητα για την καθημερινή λειτουργία του οργανισμού. Για τον λόγο αυτό, οι εργασίες ΕΜΦ εκτελούνται σε ώρες που τα συστήματα δεν είναι πολύ απασχολημένα, κυρίως τη νύχτα. Επίσης, η όλη διαδικασία πρέπει να ολοκληρωθεί μέσα σε συγκεκριμένο χρονικό διάστημα. Ο διαχειριστής επιμελείται αυτών των ζητημάτων.

Σε κάθε ένα από τα στάδια των εργασιών ΕΜΦ πρέπει να αντιμετωπιστούν ειδικότερα προβλήματα:

Το πρώτο ζητούμενο αφορά τον **εντοπισμό των πηγαίων δεδομένων που θα μεταφερθούν** και θα εισαχθούν στην ΑΔ. Δύο βασικές απαιτήσεις είναι η ελάχιστη δυνατή επιβάρυνση του πηγαίου συστήματος και η ελάχιστη δυνατή παρέμβαση στις ρυθμίσεις του. Η απλούστερη εκδοχή είναι να μεταφερθεί και να εισαχθεί το σύνολο των πηγαίων δεδομένων στην ΑΔ, όπως είχε συμβεί την πρώτη φορά φόρτωσης της ΑΔ. Ο όγκος όμως των δεδομένων είναι πολύ μεγάλος και η πλειοψηφία τους βρίσκεται ήδη στην ΑΔ από την προηγούμενη φόρτωση. Μια αποδοτικότερη προσέγγιση είναι να μεταφερθούν μόνο οι αλλαγές, δηλαδή τα νέα δεδομένα που εισήχθησαν και τα δεδομένα που τροποποιήθηκαν ή διαγράφηκαν. Υπάρχουν διάφορες τεχνικές γι' αυτήν την εργασία. Μια εκδοχή είναι να ληφθεί ένα στιγμιότυπο των δεδομένων και να συγκριθεί με το στιγμιότυπο της προηγούμενης φόρτωσης. Με τον τρόπο αυτό εντοπίζονται εύκολα οι αλλαγές και εισάγονται στην ΑΔ. Μια άλλη εκδοχή είναι να εντοπιστούν οι αλλαγές στα πηγαία συστήματα. Αυτό μπορεί να επιτευχθεί είτε με τη χρήση triggers στη βάση δεδομένων, είτε με έλεγχο στο αρχείο συμβάντων (log file) του πηγαίου συστήματος.

Σε ότι αφορά τη **μεταφορά των δεδομένων**, πρέπει να ληφθούν υπόψη ζητήματα ασφαλείας και εξοικονόμησης εύρους ζώνης. Ο όγκος των δεδομένων είναι πολύ μεγάλος και αυτό μπορεί να προκαλέσει υπερφόρτωση στο δίκτυο μεταφοράς. Για τον λόγο αυτό, τα δεδομένα συμπιέζονται. Επίσης, τα δεδομένα μπορεί να είναι διαβαθμισμένα ως εμπιστευτική πληροφορία, και να επιτρέπεται η πρόσβαση σε αυτά μόνο σε εξουσιοδοτημένα άτομα. Στην περίπτωση αυτή εφαρμόζονται τεχνικές κρυπτογράφησης.

Τα δεδομένα αφού εξαχθούν και μεταφερθούν δεν καταχωρούνται απευθείας στην ΑΔ, αλλά αποθηκεύονται σε ένα ενδιάμεσο σύστημα, που ονομάζεται Data Staging Area (DSA), το οποίο λειτουργεί ως μέσο λογικού και φυσικού διαχωρισμού μεταξύ των πηγαίων συστημάτων και της ΑΔ. Στο DSA τα δεδομένα υποβάλλονται στους απαραίτητους ελέγχους και τροποποιήσεις. Εκεί γίνεται η σύγκριση των στιγμιότυπων των δεδομένων και η αναγνώριση των νέων ή τροποποιημένων εγγραφών. Η παρεμβολή του DSA περιορίζει τις επιπτώσεις των εργασιών ΕΜΦ στην επίδοση τόσο των πηγαίων συστημάτων, όσο και του συστήματος προορισμού. Τυπικά, σε ένα DSA μπορεί να υπάρχουν αρχεία δεδομένων, σχεσιακοί πίνακες, πίνακες συμβάντων και διαστάσεων πολυδιάστατων μοντέλων, μεταδεδομένα κλπ.

Τα δεδομένα, αφού μεταφερθούν στο DSA, υπόκεινται σε **μετασχηματισμούς**. Ορισμένοι από τους μετασχηματισμούς επιβάλλονται λόγω διαφορών ανάμεσα στο *σχήμα* της ΑΔ και στο *σχήμα* των πηγαίων συστημάτων. Μια συνήθης τέτοια διαφορά είναι η χρήση διαφορετικών ονομάτων για το ίδιο αντικείμενο ή του ίδιου ονόματος για διαφορετικά αντικείμενα. Μία άλλη, είναι ο διαφορετικός τρόπος αναπαράστασης του ίδιου αντικειμένου στα δύο συστήματα. Ανάγκη για μετασχηματισμούς προκύπτει και λόγω διαφορών σε επίπεδο δεδομένων. Υπάρχει περίπτωση χρήσης διαφορετικών μονάδων μέτρησης (πχ μέτρα και γιάρδες), χρήσης διαφορετικής κωδικοποίησης για την ίδια πληροφορία (πχ το φύλο μπορεί να κωδικοποιείται ως «Άρρεν» και «Θήλυ» ή ως «Α» και «Θ» ή ως «1» και «0» κοκ), χρήσης διαφορετικού τύπου δεδομένων (πχ ακέραιος ή πραγματικός αριθμός). Διαφορές μπορεί να υπάρχουν και στο επίπεδο συναθροίσεων. Για παράδειγμα, το ένα σύστημα μπορεί να τηρεί συγκεντρωτικά στοιχεία πωλήσεων ανά ημέρα, ενώ το άλλο στοιχεία πωλήσεων ανά μήνα.

Μια συνηθισμένη εργασία που εκτελείται στα πλαίσια του μετασχηματισμού των δεδομένων είναι αυτή της **αποκανονικοποίησης**. Τα πηγαία δεδομένα τηρούνται συνήθως σε σχεσιακές βάσεις δεδομένων και είναι

κανονικοποιημένα, ώστε να ελαχιστοποιηθεί ο πλεονασμός. Στις ΑΔ, ο μεγάλος όγκος δεδομένων και η ανάγκη για μαζική επεξεργασία θα προκαλούσε μεγάλες καθυστερήσεις στο σύστημα, λόγω των πολλών συνενώσεων πινάκων (joins) που θα απαιτούνταν αν τα δεδομένα ήταν κανονικοποιημένα. Για τον λόγο αυτό, είναι αποδεκτό στις ΑΔ να υπάρχει ένας βαθμός αποκανονικοποίησης. Για παράδειγμα, σε μια σχεσιακή βάση στον πίνακα πωλήσεων, η τοποθεσία θα ορίζονταν με έναν κωδικό πόλης, και θα υπήρχε ένας άλλος πίνακας, όπου για κάθε πόλη θα αναφέρονταν ο κωδικός, το όνομα, η περιοχή, το ύψος του πληθυσμού, ο χαρακτηρισμός (πχ αστική, αγροτική) κλπ. Στην ΑΔ, στον πίνακα πωλήσεων σε κάθε γραμμή θα επαναλαμβάνονταν όλα τα στοιχεία της πόλης (περιοχή, πληθυσμός, χαρακτηρισμός κλπ.), ώστε να μειωθούν τα joins και να επιταχυνθεί το σύστημα.

Τα δεδομένα του πραγματικού κόσμου πάσχουν από προβλήματα όπως τα λάθη, οι ελλείψεις και η ύπαρξη αντικρουόμενης πληροφορίας. Τέτοια προβλήματα αντιμετωπίζονται με τον λεγόμενο **καθαρισμό** των δεδομένων. Στα πηγαία συστήματα υπάρχει περίπτωση ύπαρξης διπλοκαταχωρημένων εγγραφών ή εγγραφών με αντικρουόμενο περιεχόμενο, δεδομένων που παραβιάζουν λογικούς κανόνες, εσφαλμένες τιμές (πχ αρνητικές ποσότητες), ακραίες τιμές και εξαιρέσεις που δεν προσφέρουν χρήσιμη πληροφορία, ελλιπή δεδομένα και χαμένες τιμές, χρήση συνώνυμων τιμών σε κάποιο πεδίο (πχ η «Θεσσαλονίκη» μπορεί να καταχωρείται ολογράφως ή ως «Θεσ/νίκη») κλπ. Όλα αυτά τα προβλήματα πρέπει να αντιμετωπιστούν, ώστε τα τελικά δεδομένα που θα φορτωθούν στην ΑΔ να είναι γενικευμένα σε κατάλληλο επίπεδο, εύχρηστα, ομογενοποιημένα και σωστά. Η **ποιότητα των δεδομένων** είναι πολύ σημαντική, γιατί εσφαλμένα δεδομένα είναι δυνατόν να επηρεάσουν την ανάλυση. Προβληματικά δεδομένα μπορεί να οδηγήσουν σε εσφαλμένα συμπεράσματα και συνακόλουθα σε αποτυχημένες αποφάσεις. Οι ποιοτικές αποφάσεις βασίζονται σε ποιοτικά δεδομένα. Για την εξασφάλιση της ποιότητας των δεδομένων μπορεί να οριστούν ειδικές διαδικασίες ελέγχου, που θα κάνουν χρήση εξειδικευμένου λογισμικού και θα προβλέπουν την ύπαρξη προσωπικού επιφορτισμένου με αυτό το καθήκον.

Ένα συνηθισμένο και σοβαρό πρόβλημα καθαρισμού δεδομένων είναι η **ύπαρξη διαφορετικών κωδικών** για το ίδιο αντικείμενο, σε δύο ή περισσότερες πηγές. Για παράδειγμα, στη σχεσιακή βάση δεδομένων ΒΔ1, το προϊόν «MiracleTV» έχει τον κωδικό «200», ενώ στη σχεσιακή βάση δεδομένων ΒΔ2 έχει τον κωδικό «300». Επίσης, στη ΒΔ2, ο κωδικός «200» χρησιμοποιείται για το προϊόν «Super Blue Ray Player». Σημειωτέον ότι οι κωδικοί αυτοί είναι πρωτεύοντα κλειδιά και είναι κρισιμότατης σημασίας στις σχεσιακές βάσεις, αφού χρησιμοποιούνται για τη διασύνδεση των πινάκων. Εάν το πρόβλημα δεν γίνει αντιληπτό, και στην ΑΔ τηρηθεί η κωδικοποίηση της ΒΔ1, τότε οι πωλήσεις των «Super Blue Ray Player» που προέρχονται από τη ΒΔ2 θα εμφανίζονται ως πωλήσεις της «MiracleTV». Εάν το πρόβλημα γίνει αντιληπτό, μπορεί να αντιμετωπιστεί με τη χρήση *υποκατάστατων κλειδιών* (surrogate keys). Με τη μέθοδο αυτή ορίζονται ενιαίοι κωδικοί, οι οποίοι χρησιμοποιούνται στην ΑΔ και αντικαθιστούν τους κωδικούς που χρησιμοποιούνται στα πηγαία συστήματα. Οι κωδικοί αυτοί μπορεί να είναι ίδιοι με αυτούς που χρησιμοποιούνται σε ένα από τα πηγαία συστήματα ή να είναι τελείως διαφορετικοί. Δημιουργείται επίσης ένας πίνακας αντιστοιχίσεων, που ορίζει το πηγαίο σύστημα, τον κωδικό στο πηγαίο σύστημα και τον νέο ενιαίο κωδικό της ΑΔ όπως φαίνεται στον πίνακα 4.2

Πηγή	Κωδικός πηγαίου συστήματος	Νέος Κωδικός
ΒΔ1	200	1200
ΒΔ2	300	1200
ΒΔ2	200	1300

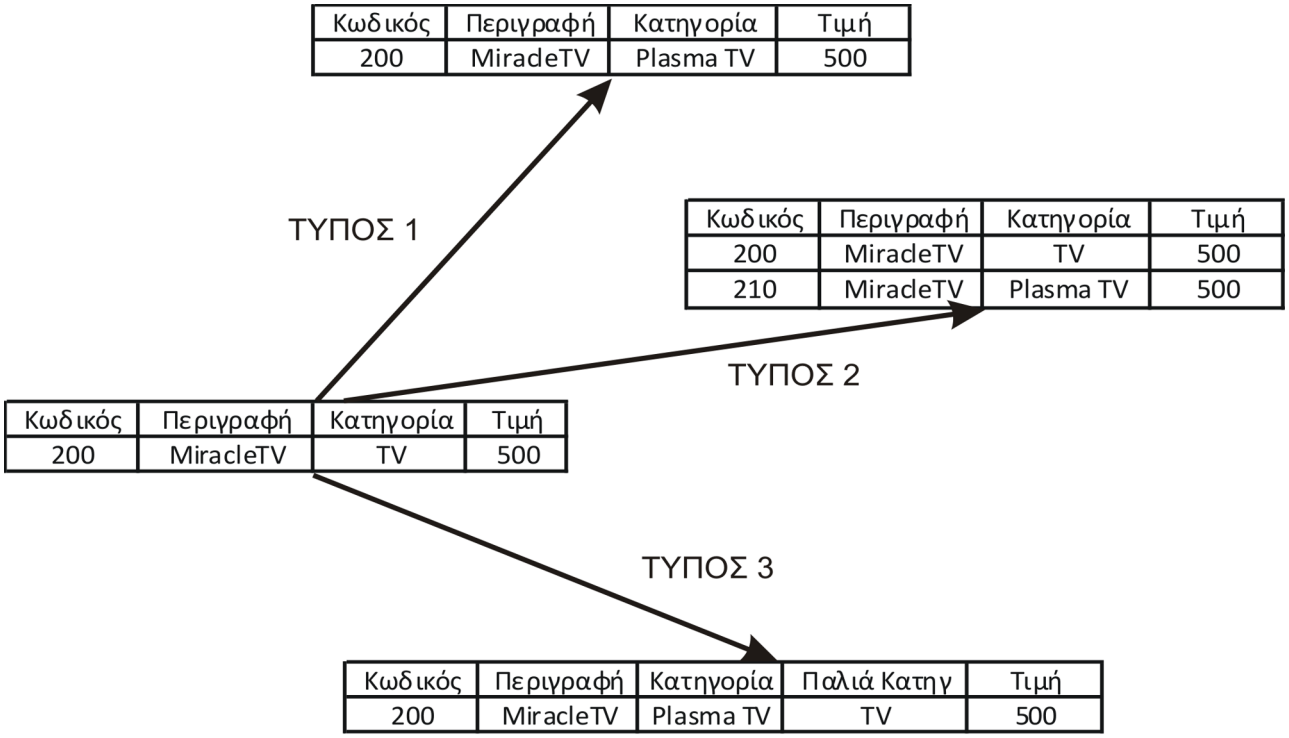
Πίνακας 4.2 *Πίνακας Αντιστοιχίσεων Κλειδιών*

Ένα άλλο πρόβλημα είναι η **διαχείριση των μεταβολών σε τιμές γραμμών των πινάκων των διαστάσεων**. Για παράδειγμα, η κατηγορία της «MiracleTV» μπορεί να αλλάξει από «TV» σε «Plasma TV». Υπάρχουν τρεις τύποι αντιμετώπισης αυτού του προβλήματος. Σύμφωνα με τον πρώτο τύπο, η καινούργια τιμή αντικαθιστά την παλιά, δηλαδή στον πίνακα της διάστασης στη γραμμή της «MiracleTV» στο πεδίο κατηγορία μπαίνει η τιμή «Plasma TV», διαγράφοντας την παλιά τιμή «TV». Ο δεύτερος τύπος διατηρεί ανέπαφη την παλιά γραμμή του προϊόντος και εισάγει μια νέα γραμμή, όμοια με την παλιά, με μόνες διαφορές ότι ορίζεται νέος κωδικός για το προϊόν, και ότι στην κατηγορία καταχωρείται η νέα τιμή «Plasma TV». Ο τρίτος τύπος προβλέπει την εισαγωγή στον πίνακα της διάστασης μιας νέας στήλης, στην οποία μπαίνει η παλιά τιμή της κατηγορίας («TV»), ενώ στην παλιά στήλη της κατηγορίας μπαίνει η νέα τιμή «Plasma TV». Οι τρεις τύποι παρουσιάζονται διαγραμματικά στο Σχήμα 4.15.

Στο τελικό στάδιο γίνεται η **φόρτωση** των δεδομένων στην ΑΔ. Τα ζητήματα που πρέπει να αντιμετωπιστούν αφορούν σε σημαντικό βαθμό τον χρόνο που απαιτείται για να ολοκληρωθεί η διαδικασία. Η τυπι-

κή διαδικασία εισαγωγής με χρήση εντολών SQL είναι πολύ αργή, γιατί εισάγει τις εγγραφές μια προς μια. Ευτυχώς, οι βάσεις δεδομένων προβλέπουν διαδικασίες μαζικής εισαγωγής εγγραφών, οι οποίες είναι πολύ ταχύτερες. Ένα άλλο πρόβλημα είναι η ενημέρωση των δεικτών (indexes), η οποία επιβραδύνει υπερβολικά το σύστημα. Ο χρήστης μπορεί να καταργήσει τους δείκτες, να φορτώσει τα δεδομένα, και στη συνέχεια να επαναδημιουργήσει τους δείκτες. Παρεμφερές ζήτημα υπάρχει και με το αρχείο συμβάντων (log file), όπως και με το σύστημα αναίρεσης αλλαγών (rollback segment). Ο χρήστης μπορεί να το απενεργοποιήσει, αλλά θα αντιμετωπίσει πρόβλημα εάν η φόρτωση αποτύχει, και το σύστημα πρέπει να επανέλθει στην προηγούμενη κατάσταση πραγμάτων. Η χρήση τεχνικών παράλληλης φόρτωσης επιταχύνει την όλη διαδικασία. Για παράδειγμα, είναι δυνατόν να φορτωθούν ταυτόχρονα οι πίνακες συμβάντων ή οι πίνακες διαστάσεων. Επίσης, χρήσιμο είναι οι συναθροίσεις να μην υπολογίζονται στην ΑΔ, αλλά να έχουν προϋπολογιστεί και να φορτώνονται στην ΑΔ.

Περισσότερες πληροφορίες για τις εργασίες Εξαγωγής Μεταφοράς και Φόρτωσης μπορεί να βρει ο αναγνώστης στο Adzic, Fiore and Sisto (2007), στο Simitisis, Vassiliadis, Skiadopoulos and Sellis (2007) και στο Vassiliadis and Simitisis(2009).



Σχήμα 4.15 *Μεταβολές σε τιμές διάστασης*

4.11 Μεταδεδομένα

Ένα πολύ σημαντικό τμήμα μιας Αποθήκης Δεδομένων είναι το Υποσύστημα Μεταδεδομένων (ΥΜ). Ο όρος **Μεταδεδομένα** σημαίνει δεδομένα που περιγράφουν άλλα δεδομένα. Το ΥΜ παρέχει πληροφορίες για τα δεδομένα που τηρούνται στην ΑΔ, αναφέρει εφαρμοζόμενες πολιτικές, επεξηγεί επιχειρηματικές έννοιες σχετικές με την πληροφόρηση που παρέχει η ΑΔ κλπ. Γενικώς, είναι μια πολύτιμη πηγή πληροφοριών και μπορεί να χρησιμεύσει ως οδηγός χρήσης και συντήρησης, αλλά και ως κανονιστικό πλαίσιο λειτουργίας της ΑΔ. Λόγω της ιδιαίτερης σημασίας του, πρέπει να τηρείται σχολαστικά και να περιέχει έγκυρα και σύγχρονα στοιχεία. Η πρόσβαση στο ΥΜ πρέπει να είναι εύκολη για κάθε χρήστη και διαβαθμισμένη ανάλογα με τον ρόλο του (διευθυντικό στέλεχος, αναλυτής, τεχνικός κλπ.).

Ειδικότερα το Υποσύστημα Μεταδεδομένων μπορεί να περιλαμβάνει τα παρακάτω:

- Τους ορισμούς επιχειρηματικών εννοιών, οι οποίοι σχετίζονται με την πληροφόρηση που παρέχει η ΑΔ, καθώς και οδηγίες για τη διεξαγωγή σχετικών αναλύσεων.
- Τη δομή της ΑΔ, δηλαδή το σχήμα της ΑΔ, τους πίνακες συμβάντων, τις διαστάσεις, και τις ιεραρχίες εννοιών, τις όψεις και τις συναθροίσεις.

- Τις πηγές των δεδομένων καθώς και τα σχήματα τους.
- Τον καθορισμό των διαδικασιών και κανόνων εξαγωγής, μετασχηματισμού και φόρτωσης των δεδομένων.
- Επεξηγήσεις σχετικά με τους μετασχηματισμούς που έχουν υποστεί τα δεδομένα.
- Την καταγραφή της γενεαλογίας των δεδομένων, την ηλικία τους, διαδοχικές εκδόσεις (versions) των δεδομένων, αναφορές που συνδέουν επιμέρους δεδομένα με την εκάστοτε πηγή τους κλπ.
- Τα Πιθανά Πρακτορεία Δεδομένων καθώς και τα περιεχόμενα τους.
- Τη φυσική περιγραφή υπολογιστικών συστημάτων και δικτύων.
- Στατιστικά στοιχεία χρήσης της ΑΔ όπως συχνότητα χρήσης, κατάλογο χρηστών, είδη αναλύσεων που διεξήχθησαν κλπ.
- Τους ιδιοκτήτες των δεδομένων.
- Προφίλ χρηστών και δικαιώματα πρόσβασης.

Το ΥΜ παρέχει πληροφορίες για όλο το φάσμα των εμπλεκόμενων συστημάτων, από τις πηγές των δεδομένων μέχρι το λογισμικό των τελικών αναλύσεων. Δεν είναι σπάνιο οι κατασκευαστές αυτών των συστημάτων να είναι διαφορετικοί, οπότε προκύπτει ανάγκη καθορισμού προτύπων. Τα πρότυπα επιτρέπουν σε διαφορετικά λογισμικά να αντλούν ή να καταχωρούν στοιχεία στο υποσύστημα μεταδεδομένων. Έχουν προταθεί δύο πρότυπα, το Open Information Model (OIM) από τον οργανισμό Metadata Coalition, και το Common Warehouse Metamodel (CWM) από τον οργανισμό OMG. Για λεπτομέρειες και συγκριτική παρουσίαση των δύο αυτών προτύπων, ο αναγνώστης μπορεί να ανατρέξει στο Vetterli, Vaduva and Staudt (2000).

4.12 Πεδία εφαρμογής και οφέλη των Αποθηκών Δεδομένων

Οι Αποθήκες Δεδομένων γνώρισαν μεγάλη διάδοση από τα τέλη της δεκαετίας του 1990 και μετέπειτα. Τα ιδιαίτερα χαρακτηριστικά τους, τα οποία επέτρεπαν την ταχεία πρόσβαση σε ποιοτική πληροφορία, δομημένη με βάση επιχειρηματικούς κανόνες και ζητήματα, καθιστούσαν τις ΑΔ ένα πολύτιμο εργαλείο για την υποστήριξη των ανώτατων και μεσαίων στελεχών, σε καθήκοντα σχετικά με την ανάλυση των δεδομένων και τη λήψη αποφάσεων. Η σύγχρονη αυτή τεχνολογία βρήκε σύντομα εφαρμογή σε μια σειρά από επιχειρηματικούς τομείς. Ενδεικτικά, και όχι περιοριστικά, αναφέρονται οι παρακάτω:

- Διοίκηση
  - Στρατηγικός αναπροσανατολισμός και επανακαθορισμός στρατηγικών στόχων (πχ η περίπτωση της First American Corporation).
  - Παρακολούθηση Κρίσιμων Δεικτών Επίδοσης.
- Πωλήσεις και διαφήμιση
  - Τμηματοποίηση αγοράς.
  - Ανάλυση πωλήσεων και πρόβλεψη.
  - Ανάλυση και σχεδιασμός ερευνών αγοράς.
- Χρηματοοικονομικά και Λογιστική
  - Σύνταξη προϋπολογισμού.
  - Ανάλυση χρηματοοικονομικής απόδοσης.
  - Κοστολόγηση βάσει δραστηριοτήτων (Activity Based Costing).
- Παραγωγή
  - Σχεδιασμός παραγωγής.
  - Ανάλυση ελαττωματικών προϊόντων.

Σήμερα όλοι οι μεγάλοι οργανισμοί διαθέτουν μια Αποθήκη Δεδομένων. Η ευρεία αποδοχή των ΑΔ από τον επιχειρηματικό κόσμο δεν είναι τυχαία. Οι ΑΔ προσφέρουν ταχύτατη και ακριβή πληροφόρηση, που αντανακλά περίπλοκες σχέσεις δεδομένων, και που προκύπτει από μαζικούς υπολογισμούς. Η χρήση της πληροφορίας αυτής μπορεί να αποφέρει στον οργανισμό πολλαπλά, απτά και μετρήσιμα, αλλά και όχι άμεσα μετρήσιμα οφέλη.

Ορισμένα από τα άμεσα μετρήσιμα οφέλη είναι τα ακόλουθα:

- Μείωση κόστους λήψης αποφάσεων με την επιτάχυνση της διαδικασίας.
- Μείωση διαφημιστικού κόστους με την εφαρμογή στοχευμένης διαφήμισης.
- Μείωση διαφημιστικού κόστους με την καλύτερη αποτίμηση και σχεδιασμό των διαφημιστικών

εκστρατειών.
- Αύξηση κερδών λόγω καλύτερης διαχείρισης των αποθεμάτων της αποθήκης και επαρκούς τροφοδότησης.
- Μείωση κόστους λόγω αποδοτικότερης διαχείρισης των αγορών.
- Μείωση του φόρτου εργασίας των πηγαίων συστημάτων.

Ωστόσο, τα σημαντικότερα οφέλη των Αποθηκών Δεδομένων δεν είναι απτά και άμεσα μετρήσιμα. Σε αυτά περιλαμβάνονται:

- Η βελτίωση της ποιότητας των αποφάσεων λόγω ταχείας πρόσβασης σε ποιοτική πληροφόρηση και συνακόλουθης μείωσης του ρίσκου.
- Η ταχύτερη και αποτελεσματικότερη ανταπόκριση στις νέες προκλήσεις των αγορών.
- Η αύξηση της παραγωγικότητας λόγω της συγκέντρωσης των δεδομένων και της ταχείας πρόσβασης σε αυτά.
- Η βελτίωση της σχέσης με τους πελάτες μέσω της αναγνώρισης καταναλωτικών τάσεων.
- Η καλύτερη κατανόηση των επιχειρηματικών διαδικασιών και κατ' επέκταση η αύξηση της δυνατότητας αναπροσαρμογής τους.
- Η αναβάθμιση της πληροφοριακής υποδομής της επιχείρησης και η αύξηση της δυνατότητας για παραγωγή και διανομή πληροφορίας. Πιθανώς ευκολότερη ανάπτυξη νέων εφαρμογών.
- Η ενθάρρυνση της δημιουργικότητας των στελεχών μέσω της άμεσης πρόσβασης στα δεδομένα και της δυνατότητας πρωτόβουλης ανάλυσης των δεδομένων χωρίς τη χρήση προκαθορισμένων φορμών.

Οι Watson and Haley (1998) αναφέρονται σε ορισμένα οφέλη που προσφέρουν οι ΑΔ στη σύγχρονη επιχείρηση.

Βιβλιογραφία / Αναφορές

Adzic, J., Fiore, V., & Sisto, L. (2007). Extraction, Transformation and Loading Process. In R. Wrembel & C. Koncilia (Eds.), *Data Warehouses and OLAP: Concepts, Architectures and Solutions* (pp. 88 - 110). Hersey, PA: Idea Group Inc.

Business Intelligence Tools & Data Warehousing Applications – Teradata. (n.d.). Retrieved 30 January, 2015, from http://www.teradata.com/products-and-services/applications/?LangType=1033&LangSelect=true

Coronel, C., & Morris, S. (2014). *Database Systems: Design, Implementation and Management*. Stamfort Place, CT: Cengage Learning.

Date, J. (2012). *Database Design and Relational Theory: Normal forms and All That Jazz*. North Sebastopol, CA: O'Reilly Media Inc.

Dupupet, C., & Grays, D. (2013). *Oracle Data Integrator 11g Cookbook*. Birmingham, UK: Packt Publishing Ltd.

Gupta, H., & Mumick, I. S. (2005). Selection of Views to Materialize in a Data Warehouse. *IEEE Transactions on Knowledge and Data Engineering*, *17*(1), 24-43. doi: 10.1109/TKDE.2005.16

Han, J., & Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Diego, CA: Morgan Kaufman.

Harinaryan, V., Rajaraman, A., & Ulman, J. (1996). Implementing Data Cubes Efficiently. *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data,* 205-216. New York, NY: ACM. doi: 10.1145/233269.233333

Hurtado, C. A., & Gutierrez, C. (2007). Handling Structural Heterogeneity in OLAP. In R. Wrembel & C. Koncilia (Eds.), *Data Warehouses and OLAP: Concepts, Architectures and Solutions* (pp 27-57). Hersey, PA: IRM Press. doi: 10.4018/987-1-59904-364-7.ch002

Inmon, W. H. (1996). *Building the Data Warehouse.* New York, NY: John Wiley & Sons Inc.

Kimball Group. (1995). *Is ER Modeling Hazardous to DSS? – Kimball Group*. Retrieved 21 September, 2015, from http://www.kimballgroup.com/1995/10/is-er-modeling-hazardous-to-dss/

Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*. Indianapolis, IN: Wiley Publications Inc.

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Indianapolis, IN: John Willey & Sons Inc.

Kimball Group. (2015). *Kimball Group | Dimensional Data Warehousing Experts.* Retrieved 30 January, 2015, from http://www.kimballgroup.com/.

Knight, B., Knight, D., Moss, J., Davis, M., & Rock, C. (2014). *Professional Microsoft SQL Server Integration Services*. Indianapolis, IN: John Wiley & Sons Inc.

Sarawagi, S., Agrawal, R., & Megiddo, N. (1998). Discovery-Driven Exploration of OLAP Data Cubes. *Lecture Notes in Computer Science*, *1377*, 168-182. doi: 10.1007/BFb0100984

Simitisis, A., Vassiliadis, P., Skiadopoulos, S., & Sellis, T. (2007). Data Warehouse Refreshment. In R. Wrembel & C. Koncilia (Eds.), *Data Warehouses and OLAP: Concepts, Architectures and Solutions* (pp. 111 - 134). Hersey, PA: Idea Group Inc.

Vassiliadis, P., & Simitisis, A. (2009). Extraction, Transformation and Loading. In L. Liu & M.T. Ozsu (Eds.), *Encyclopedia of Database Systems* (pp. 1095 – 1101). New York, NY: Springer.

Vetterli, T., Vaduva, A., & Staudt, M. (2000). Metadata standards for data warehousing: Open Information Model vs. Common Warehouse Metamodel. *ACM SIGMOD Records*, *3*(23), 68-75. doi: 10.1145/362084.362138

Watson, H. J., & Haley, B. J. (1998). Managerial Considerations. *Communications of the ACM*, *41*(9), 32-37. doi: 10.1145/285070.285077

Κριτήρια Αξιολόγησης

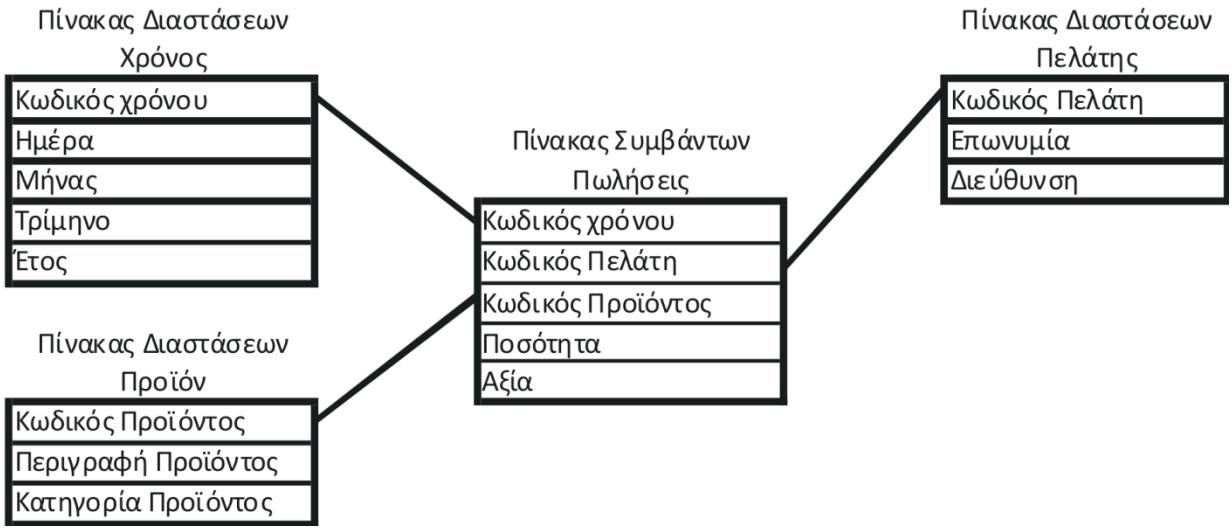
Άσκηση Υπολογισμών 4.1

Ένα ηλεκτρονικό κατάστημα τηρεί μια αποθήκη δεδομένων για να παρακολουθεί τα στοιχεία των πωλήσεων. Αναλυτικότερα, για κάθε πώληση καταγράφεται ο πελάτης, το προϊόν, η χρονική στιγμή, η ποσότητα και η αξία. Για κάθε πελάτη καταγράφεται ο κωδικός, η επωνυμία και η διεύθυνση, για κάθε προϊόν ο κωδικός, η περιγραφή και η κατηγορία προϊόντος και για τη χρονική στιγμή καταγράφεται η ημέρα, ο μήνας, το τρίμηνο και ο χρόνος.

- **Σχεδιάστε τη λογική δομή της Αποθήκης Δεδομένων χρησιμοποιώντας Σχήμα Αστέρα.**
- **Ορίστε τις πράξεις OLAP που απαιτούνται για να βρείτε το σύνολο των αγορών του κάθε πελάτη για το έτος 2015.**

Λύση

Βήμα 1. Σύμφωνα με την περιγραφή του προβλήματος, η Αποθήκη Δεδομένων περιέχει έναν Πίνακα Συμβάντων για τις πωλήσεις και τρεις Πίνακες Διαστάσεων για τα προϊόντα, τους πελάτες και τον χρόνο. Η λογική δομή της ΑΔ παρουσιάζεται στο Σχήμα 4.16



Σχήμα 4.16 *Λογική δομή ΑΔ Άσκησης 1*

Βήμα 2. Για να βρεθεί το σύνολο των αγορών του κάθε πελάτη για το έτος 2015 απαιτούνται οι παρακάτω πράξεις OLAP

- Roll-up on Χρόνος (from Κωδικός Χρόνου to Έτος).
- Slice on Χρόνος (with Έτος= “2015”).
- Roll-up on Προϊόν (from Κωδικός Προϊόντος to all).

Άσκηση Υπολογισμών 4.2

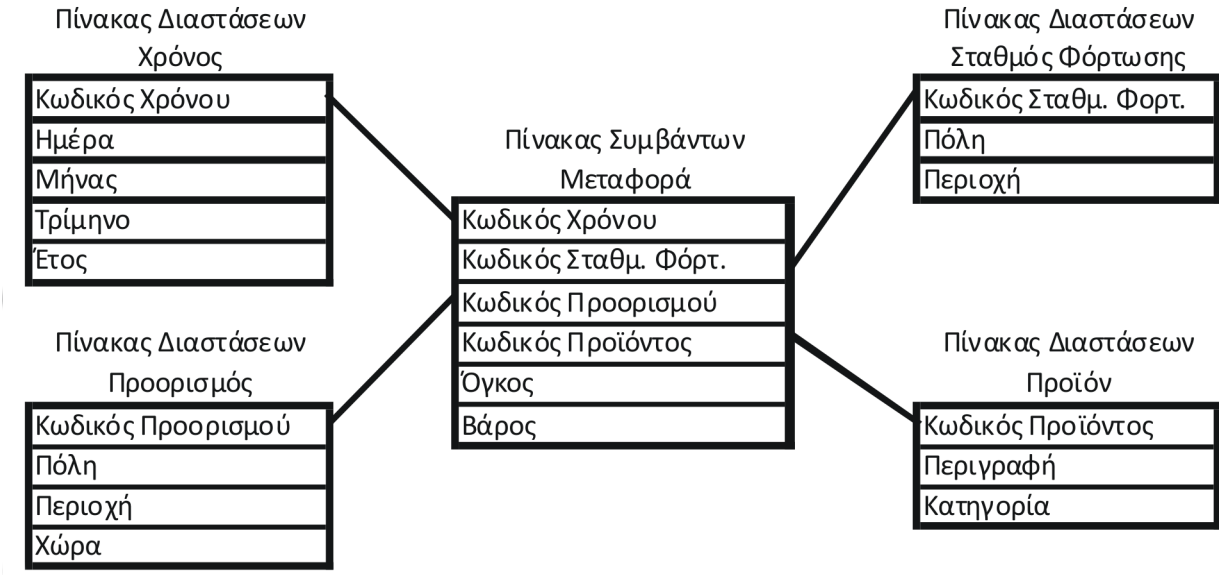
Μια ελληνική εταιρεία διεθνών μεταφορών διακινεί εμπορεύματα από την Ελλάδα προς πόλεις του εξωτερικού. Η εταιρεία διαθέτει σε διάφορες ελληνικές πόλεις υποκαταστήματα και σταθμούς φόρτωσης. Για την παρακολούθηση και διαχείριση της ροής των εμπορευμάτων, η εταιρεία διατηρεί μια Αποθήκη Δεδομένων. Για κάθε μεταφορά τηρείται πληροφορία σχετικά με το υποκατάστημα φόρτωσης, την πόλη προορισμού, τον χρόνο φόρτωσης, το εμπόρευμα, τον όγκο και το βάρος. Σχετικά με τον προορισμό καταγράφεται η πόλη, η περιοχή και η χώρα, ενώ σχετικά με το εμπόρευμα καταγράφεται η περιγραφή και η κατηγορία του.

- **Σχεδιάστε τη λογική δομή της Αποθήκης Δεδομένων χρησιμοποιώντας Σχήμα Αστέρα.**

- **Ορίστε τις πράξεις OLAP που απαιτούνται για να βρείτε τον όγκο των εμπορευμάτων που διακινήθηκαν από τη Μακεδονία προς κάθε ιταλική πόλη το έτος 2014.**

Λύση

Βήμα 1. Σύμφωνα με την περιγραφή του προβλήματος, η Αποθήκη Δεδομένων περιέχει έναν Πίνακα Συμβάντων για τις μεταφορές και τέσσερεις Πίνακες Διαστάσεων για τον χρόνο, τα προϊόντα, τον σταθμό φόρτωσης και τον προορισμό. Η λογική δομή της ΑΔ παρουσιάζεται στο Σχήμα 4.17.



Σχήμα 4.17 *Λογική δομή ΑΔ Άσκησης 2*

Βήμα 2. Για να βρεθεί ο όγκος των εμπορευμάτων που διακινήθηκαν από τη Μακεδονία προς κάθε ιταλική πόλη απαιτούνται οι παρακάτω πράξεις OLAP:

- Roll-up on Προϊόν (from Κωδικός Προϊόντος to all).
- Roll-up on Προορισμός (from Κωδικός Προορισμού to Χώρα).
- Roll-up on Σταθμός Φόρτωσης (from Κωδικός Σταθμ. Φορτ. to Περιοχή).
- Roll-up on Χρόνος (from Κωδικός Χρόνου to Έτος).
- Dice on Προορισμός, Σταθμός Φόρτωσης, Χρόνος (with Χώρα = «Ιταλία» and Περιοχή = «Μακεδονία» and Έτος = «2014»).
- Drill-down on Προορισμός (from Χώρα to Πόλη).

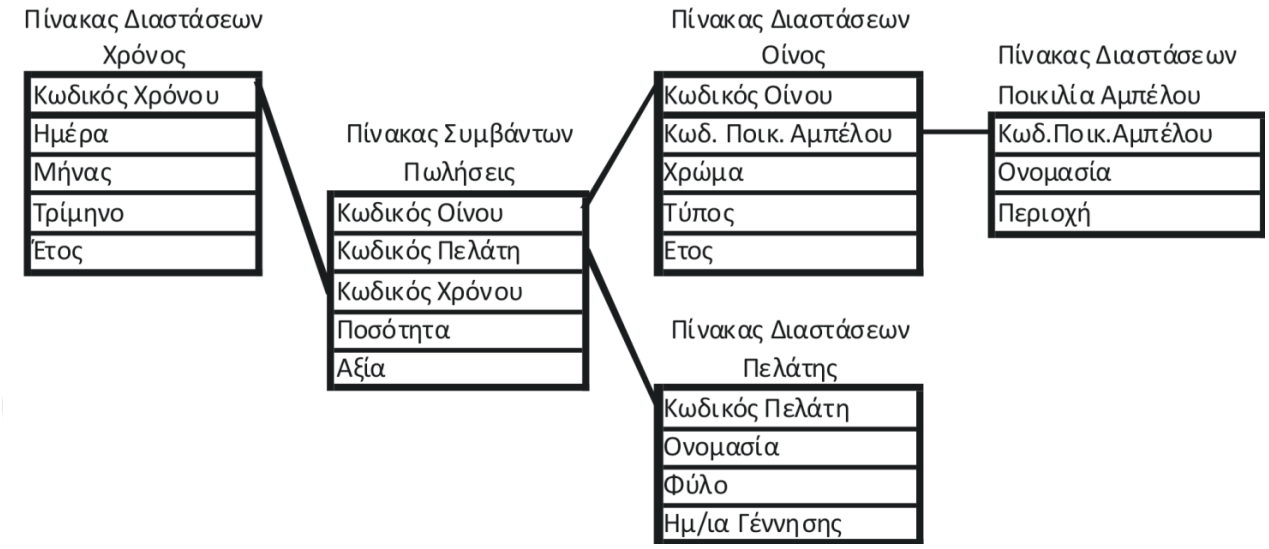
Άσκηση Υπολογισμών 4.3

Μια επιχείρηση εμπορίας οίνου τηρεί Αποθήκη Δεδομένων για να παρακολουθεί τα στοιχεία πωλήσεων. Για κάθε πώληση καταγράφεται η χρονική στιγμή, το κρασί, ο πελάτης, η ποσότητα και η αξία. Στοιχεία του κρασιού είναι ο τύπος, το χρώμα, η ποικιλία αμπέλου και η χρονιά, ενώ για την ποικιλία αμπέλου καταγράφεται η ονομασία της και η περιοχή προέλευσης. Για τον πελάτη τηρούνται στοιχεία σχετικά με το όνομα του, το φύλο και την ημερομηνία γέννησης.

- **Σχεδιάστε τη λογική δομή της Αποθήκης Δεδομένων χρησιμοποιώντας Σχήμα Χιονονιφάδας.**
- **Ορίστε τις πράξεις OLAP που απαιτούνται για να βρείτε τη συνολική ποσότητα κρασιού που αγοράστηκε από γυναίκες.**

Λύση

Βήμα 1. Η Αποθήκη Δεδομένων περιέχει έναν πίνακα συμβάντων για τις πωλήσεις και τρεις πίνακες διαστάσεων για τον χρόνο, τον οίνο και τον πελάτη. Υπάρχει ένας πρόσθετος πίνακας για τον οίνο, ο οποίος αναφέρεται στην ποικιλία αμπέλου. Η λογική δομή της ΑΔ παρουσιάζεται στο Σχήμα 4.18



Σχήμα 4.18 *Λογική δομή ΑΔ Άσκησης 3*

Βήμα 2. Για να βρεθεί η συνολική ποσότητα κρασιού που αγοράστηκε από γυναίκες απαιτούνται οι παρακάτω πράξεις OLAP

- Roll-up on Χρόνος (from Κωδικός Χρόνου to all).
- Roll-up on Οίνος (from Κωδικός Οίνου to all).
- Roll-up on Πελάτης (from Κωδικός Πελάτη to Φύλο).
- Slice on Πελάτης (with Φύλο=»Γυναίκα»).

5 Οπτική και Διερευνητική Ανάλυση Δεδομένων

Σύνοψη

Θέμα του παρόντος Κεφαλαίου είναι η οπτική αναπαράσταση των δεδομένων και η ανάλυση τους με χρήση γραφικών μέσων. Οι τεχνικές οπτικοποίησης διαθέτουν μια σειρά από πλεονεκτήματα, τα οποία τις καθιστούν χρήσιμο εργαλείο για τον εντοπισμό και αναγνώριση δομών και ιδιοτήτων σε ένα σύνολο δεδομένων. Αρχικά γίνεται μια εισαγωγή στην οπτικοποίηση των δεδομένων, παρέχονται μερικά ιστορικά στοιχεία για την εξέλιξη της, παρουσιάζονται οι τρέχουσες τάσεις της και προτείνονται σχεδιαστικές αρχές για τη δημιουργία αποτελεσματικών γραφικών, καθώς και οδηγίες για την επιλογή κατάλληλης τεχνικής οπτικοποίησης. Ακολουθεί μια εισαγωγή στη Διερευνητική Ανάλυση Δεδομένων, εξηγούνται οι αρχές και η μεθοδολογία της, οι οποίες αντιπαραβάλλονται με τις αντίστοιχες του Ελέγχου Υπόθεσης και επιπλέον συνοψίζονται τα πλεονεκτήματα της.
Κατά καιρούς έχουν προταθεί διάφορες τεχνικές για τη γραφική απεικόνιση των δεδομένων. Οι τεχνικές αυτές διαφέρουν ποικιλόμορφα μεταξύ τους. Στο παρόν Κεφάλαιο οι τεχνικές ταξινομούνται με βάση τον τύπο των δεδομένων που χειρίζονται, τη μέθοδο οπτικοποίησης που εφαρμόζουν και τον τρόπο αλληλεπίδρασης με τον χρήστη. Ως προς τη μέθοδο οπτικοποίησης, χωρίζονται σε Τυπικές δύο ή τριών διαστάσεων, Γεωμετρικού Μετασχηματισμού, Εικονογραφικές, τεχνικές Εικονοστοιχείων και τεχνικές Στοίβας ή Ιεραρχικές. Ακολούθως, παρουσιάζονται μερικές από τις πλέον γνωστές και εφαρμοσμένες τεχνικές. Συγκεκριμένα: τα Γραφήματα Γραμμής, τα Ραβδογράμματα, οι Πίτες, τα Διαγράμματα Διασποράς, ο Πίνακας Διαγραμμάτων Διασποράς, τα διαγράμματα Παράλληλων Συντεταγμένων, τα διαγράμματα HyperSlice, τα πρόσωπα του Chernoff, οι εικόνες Stick Figure, τα διαγράμματα Αστέρων, η τεχνική Shape Coding, τα διαγράμματα Επαναληπτικών Προτύπων, τα διαγράμματα Κυκλικών Τομέων, η τεχνική Dimensional Stacking, η τεχνική Worlds within Worlds και οι Δενδροχάρτες. Ως μελέτη περίπτωσης σχετικά με την εφαρμογή γραφικών μέσων για την εξαγωγή συμπερασμάτων παρατίθεται το πρόβλημα της ανίχνευσης απάτης. Παρουσιάζονται δύο συστήματα, ένα για τον έλεγχο των συναλλαγών στο Χρηματιστήριο της Νέας Υόρκης και ένα για την αντιμετώπιση εγκλημάτων νομιμοποίησης παράνομου χρήματος στις ΗΠΑ. Επίσης, παρουσιάζεται μια ελληνική εργασία, στα πλαίσια της οποίας αναπτύχθηκε σύστημα εντοπισμού περιπτώσεων απάτης, με συνεργασία υπαλλήλων του οργανισμού και πελατών. Τέλος, γίνεται αναφορά στα λεγόμενα Ταμπλό ή dashboards, τα οποία αποτελούν τη βασική πλατφόρμα οπτικοποίησης πληροφοριών στα συστήματα Επιχειρηματικής Ευφυΐας.

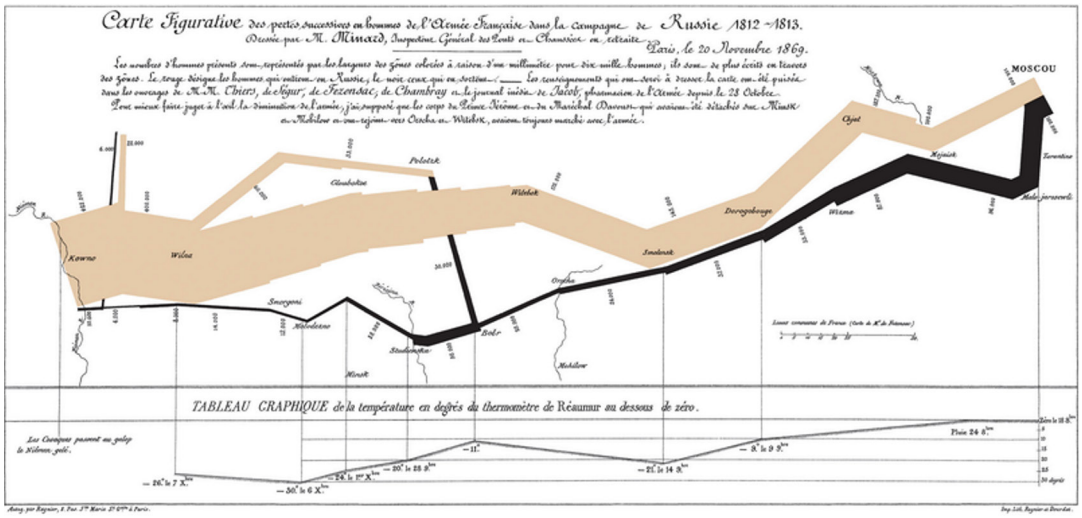
Προαπαιτούμενη γνώση

Η οπτικοποίηση των δεδομένων και η χρήση γραφικών μέσων για την ανάλυση τους είναι ένα ευρύ πεδίο. που μπορεί άνετα να αποτελέσει το θέμα ενός πλήρους συγγράμματος. Στο παρόν Κεφάλαιο επιχειρείται μια εισαγωγή στο αντικείμενο αυτό. Ο αναγνώστης μπορεί να εντρυφήσει στα θέματα που αναπτύσσονται, χωρίς να χρειάζεται εξειδικευμένες γνώσεις. Ωστόσο, η προηγούμενη ανάγνωση ορισμένων συγγραμμάτων μπορεί να συμβάλλει στη βαθύτερη κατανόηση του κεφαλαίου. Η εργασία του Friendly (2005) καταγράφει τα ορόσημα στην ιστορία της οπτικοποίησης των δεδομένων και συνοψίζει το χρονικό της εξέλιξης της. Το βιβλίο του Bertin (1983) αποτελεί σταθμό στην οπτικοποίηση των δεδομένων. Ο συγγραφέας εισάγει μια μεθοδολογία οργάνωσης των οπτικών στοιχείων των γραφικών, βασισμένη στα χαρακτηριστικά και στις σχέσεις των δεδομένων, η οποία στην πορεία καθιερώθηκε. Επίσης, πολύ σημαντικό είναι το σύγγραμμα του Tukey (1977), στο οποίο παρουσιάζεται για πρώτη φορά και συστηματοποιείται η Διερευνητική Ανάλυση των Δεδομένων. Στο ίδιο βιβλίο ο συγγραφέας προτείνει μια σειρά νέων γραφημάτων. Η εργασία του Shneiderman (1996) εισηγείται μια μεθοδολογία για την οπτικοποίηση των δεδομένων.
Πέραν των τεχνικών που παρουσιάζονται στο παρόν κεφάλαιο έχουν προταθεί και πολλές άλλες. Μερικές από τις κυριότερες, μαζί με το σύγγραμμα στο οποίο ο αναγνώστης μπορεί να βρει σχετικές λεπτομέρειες, είναι οι ακόλουθες: Color Icons (Levkowitz, 1991), Δένδρα των Kleiner and Hartigan (Kleiner & Hartigan, 1981), Τεχνικές Σπείρας / Τεχνικές Αξόνων (Keim & Kriegel, 1994), Pixel Bar Charts (Keim, Hao, Ladish, Hsu & Dayal, 2001), Attribute Blocks (Miller, 2007), Andrews Curves (Andrews, 1972),Radial Coordinate Visualization (Hoffman, Grinstein, Marx, Grosse & Stanley, 1997), RINGS (Teoh & Ma, 2002), Cone Trees (Robertson, Mackinlay & Card, 1991), Infocube (Rekimoto & Green, 1993), Tile Bars (Hearst, 1995).

5.1 Οπτικοποίηση

Ο όρος Οπτικοποίηση Δεδομένων αναφέρεται σε ένα σύνολο τεχνικών, που αντικείμενο έχουν την οπτική αναπαράσταση των δεδομένων με τη χρήση γραφικών μέσων. Με τη βοήθεια της οπτικοποίησης μπορεί να απεικονιστούν σε γραφήματα ιδιότητες των δεδομένων, σχέσεις συνάφειας, συγκρίσεις τιμών, γεωγραφική διασπορά συμβάντων, ανοδικές και καθοδικές τάσεις, επιμερισμός συνόλων σε υποσύνολα και πολλές άλλες πληροφορίες. Τα γραφικά είναι ένα μέσο για τον εντοπισμό και αναγνώριση δομών και ιδιοτήτων σε ένα σύνολο δεδομένων (Card, Mackinlay & Shneiderman, 1999; Rober 2000). Η πληροφόρηση αυτή παρέχεται με τρόπο κατανοητό με «μια ματιά». Ο ανθρώπινος εγκέφαλος κατανοεί καλύτερα και γρηγορότερα μια πληροφορία, όταν αυτή αποτυπώνεται σε μια εικόνα, παρά όταν περιγράφεται με μορφή αναλυτικού κειμένου. Επιπλέον, η γραφική απεικόνιση της πληροφορίας είναι καλαίσθητη και σαφώς πιο ευχάριστη από την ανάγνωση κειμένου. Αυτές οι ιδιότητες της οπτικοποίησης την έχουν καταστήσει χρήσιμο εργαλείο για την ανάλυση των δεδομένων και την εξαγωγή συμπερασμάτων. Επίσης, η οπτικοποίηση αποτελεί εξαιρετικό μέσο για την παρουσίαση και τη μετάδοση της πληροφορίας.

Η χρήση γραφικών για την οπτική αναπαράσταση των δεδομένων έχει μακρά ιστορία και έχει εφαρμοστεί εδώ και αιώνες. Πίνακες με αστρονομικές πληροφορίες, οι οποίοι χρονολογούνται τον 2ο αιώνα π.Χ., έχουν βρεθεί στην Αίγυπτο. Οι γεωγραφικοί χάρτες είναι μια άλλη, πολύ παλιά μέθοδος απεικόνισης. Σύμφωνα με τον Friendly (2005), η απαρχή της οπτικής αναπαράστασης τοποθετείται χρονολογικά τον 17ο αιώνα, με την εμφάνιση των θεωριών πιθανοτήτων και μέτρησης σφάλματος. Την εποχή εκείνη, ο Φλαμανδός αστρονόμος Michael Florent Van Langren εφάρμοσε πρώτος τρόπους οπτικής αναπαράστασης στατιστικών δεδομένων. Τον 18ο αιώνα προτάθηκαν οι θεματικοί χάρτες, που δεν περιορίζονται στα γεωγραφικά στοιχεία, αλλά παρέχουν πρόσθετες ιατρικές, οικονομικές ή άλλες πληροφορίες. Ο Playfair (1786) επινόησε μια σειρά από γραφικά εργαλεία, που χρησιμοποιούνται ευρύτατα μέχρι και σήμερα, όπως γραφήματα γραμμής, ραβδογράμματα, πίτες και πολλά άλλα. Ο Friendly (2005) χαρακτηρίζει το πρώτο μισό του 19ου αιώνα ως το ορόσημο δημιουργίας των σύγχρονων γραφικών και αναφέρεται στο έργο του C. Minard. Στο Σχήμα 5.1 παρουσιάζεται ο χάρτης του Minard, που αποτυπώνει τη συρρίκνωση του στρατού του Μ. Ναπολέοντα κατά την εκστρατεία του στη Ρωσία. Ο οριζόντιος άξονας υποδεικνύει την τοποθεσία. Τα σημεία ορίζουν τις ακραίες θερμοκρασίες που προκάλεσαν έξαρση κρυοπαγημάτων. Ο διάσημος στατιστικολόγος Tufte (1983, p.40) χαρακτήρισε αυτό το γράφημα ως «Ισως το καλύτερο στατιστικό γράφημα που σχεδιάστηκε ποτέ». Χρυσή εποχή των στατιστικών γραφικών αποτέλεσε το δεύτερο μισό του 19ου αιώνα, όταν η επιστήμη της στατιστικής γνώρισε αλματώδη εξέλιξη με τις εργασίες των Gauss και Laplace. Επίσης, το δεύτερο μισό του 20ου αιώνα η οπτικοποίηση των δεδομένων γνωρίζει νέα άνθηση με τη συμβολή του Bertin (1983), που συνδέει στοιχεία των γραφικών με τα χαρακτηριστικά και τις σχέσεις των δεδομένων, καθώς και με τις εργασίες του Tukey (1977), ο οποίος εισήγαγε τη Διερευνητική Ανάλυση των Δεδομένων (Exploratory Data Analysis). Φυσικά, η έλευση της πληροφορικής την ίδια εποχή έδωσε νέα τεράστια ώθηση στην οπτικοποίηση των δεδομένων.



Σχήμα 5.1 *Ο χάρτης του Charles Minard (1869) (Αναπαραγωγή από Wikimedia Commons)*

Στη σημερινή εποχή, η ευρύτατη εφαρμογή της πληροφορικής, τα επιτεύγματα στην ανάπτυξη του υλικού και του λογισμικού και η μαζική παραγωγή δεδομένων προσφέρουν πρωτόγνωρες δυνατότητες στην οπτι-

κοποίηση των δεδομένων. Τα σύγχρονα γραφικά είναι διαδραστικά και επιτρέπουν στον χρήστη να επιλέξει διαφορετικά επίπεδα λεπτομέρειας ή γενίκευσης. Με την ταυτόχρονη εμφάνιση πολλών αλληλοσχετιζόμενων γραφημάτων, επιτυγχάνεται η συγκριτική αντιπαράθεση υποσυνόλων των δεδομένων ή διαφορετικών χαρακτηριστικών τους. Η γεωχωρική οπτικοποίηση με τη χρήση χαρτών αποτυπώνει τη χωρική διασπορά συμβάντων. Η μεγαλύτερη ίσως πρόκληση της εποχής είναι η ραγδαία διόγκωση των δεδομένων. Ο όγκος των δεδομένων που παράγεται και διακινείται μετριέται σε δεκάδες exabytes καθημερινά. Έχουν προταθεί γραφικές μέθοδοι για τη σάρωση μεγάλου όγκου δεδομένων και τον εντοπισμό ακραίων τιμών και τάσεων. Η άνθηση του Διαδικτύου τροφοδότησε τη μαζική παραγωγή νέου τύπου δεδομένων, όπως είναι τα δίκτυα οντοτήτων με περίπλοκες διασυνδέσεις. Παράδειγμα τέτοιου δικτύου είναι οι «φιλίες» των μέσων κοινωνικής δικτύωσης. Η οπτικοποίηση του δικτύου διευκολύνει τη μελέτη τους.

Η απεικόνιση δεδομένων με γραφικό τρόπο δεν είναι πάντα μια εύκολη εργασία και δεν υπάρχει μια μαγική συνταγή που να εξασφαλίζει ένα ποιοτικό αποτέλεσμα. Σε μεγάλο βαθμό το αποτέλεσμα εξαρτάται από τη δημιουργικότητα και φαντασία του σχεδιαστή. Ωστόσο, έχουν προταθεί κάποιες σχεδιαστικές αρχές για τη δημιουργία αποτελεσματικών γραφικών. Σύμφωνα με τον Tufte (1983), ο σχεδιαστής πρέπει να ακολουθεί τις παρακάτω υποδείξεις:

- Να δείχνει τα δεδομένα.
- Να μην διαταράσσει το νόημα των δεδομένων.
- Να παρουσιάζει πολλά δεδομένα σε περιορισμένο χώρο.
- Να κάνει συνεκτικά μεγάλα σύνολα δεδομένων.
- Να ενθαρρύνει την επαγωγή συμπερασμάτων, πχ με σύγκριση τιμών.
- Να δίνει διαφορετικές οπτικές γωνίες των δεδομένων, από συνοπτικές έως αναλυτικές.

Για ένα σύνολο δεδομένων και για μια εργασία ανάλυσης, η επιλογή της κατάλληλης τεχνικής οπτικοποίησης πρέπει να λαμβάνει υπόψη της μια σειρά από παράγοντες. Ο Mazza (2009) συνοψίζει μερικούς τέτοιους παράγοντες που προτάθηκαν από τους Spence (2001) και Card et al. (1999):

- Το πρόβλημα. Αφορά το τι πρέπει να βρεθεί η να παρουσιαστεί.
- Η φύση των δεδομένων. Τα δεδομένα μπορεί να είναι αριθμητικά, ονομαστικά, κείμενο κλπ.
- Το πλήθος των διαστάσεων. Η εμπειρία του ανθρώπου προέρχεται από τρισδιάστατο χώρο και γι' αυτό η απεικόνιση χώρου με περισσότερες διαστάσεις αποτελεί πρόκληση.
- Η δομή των δεδομένων. Τα δεδομένα μπορεί να είναι γραμμικά, γεωγραφικά, χρονικά, ιεραρχικά ή να έχουν δικτυακή δομή.
- Ο τύπος της αλληλεπίδρασης, όπως μεγέθυνση δεδομένων, επιλογή δεδομένων κλπ.

Οι σύγχρονες τεχνικές οπτικοποίησης της πληροφορίας είναι πολύτιμα εργαλεία για την ανάλυση των δεδομένων και την εξαγωγή συμπερασμάτων, καθώς προσφέρουν μια σειρά από πλεονεκτήματα. Ειδικότερα οι τεχνικές οπτικοποίησης:

- Απεικονίζουν ιδιότητες των δεδομένων με μια άμεσα κατανοητή εικόνα.
- Παρέχουν συμπυκνωμένη πληροφόρηση με μια ματιά.
- Αποκαλύπτουν τάσεις δεδομένων, εξαιρέσεις και ακραίες τιμές, συστάδες δεδομένων και κενά.
- Είναι ικανές να χειρίζονται μεγάλους όγκους δεδομένων.
- Αναπαριστούν την πληροφορία με αντικειμενικό τρόπο. Αντιθέτως, η λεκτική περιγραφή μπορεί να αντανακλά ή να υποκρύπτει υποκειμενικές αντιλήψεις.
- Είναι διαδραστικές και επιτρέπουν στον χρήστη τη διεξαγωγή διαφορετικών αναλύσεων.
- Αποκαλύπτουν κρυμμένη πληροφορία, που θα χρησιμοποιηθεί για την εξαγωγή συμπερασμάτων.
- Τα αποτελέσματα τους, τα οποία αποκαλύπτουν ιδιότητες των δεδομένων, μπορούν να χρησιμοποιηθούν για τον προσανατολισμό της περαιτέρω ανάλυσης με άλλα μέσα.

Βασικότερο μειονέκτημα των τεχνικών οπτικοποίησης είναι ότι οι νέοι σύνθετοι τρόποι μπορεί να μην είναι κατανοητοί από εξειδικευμένους χρήστες. Η δυσχέρεια στην κατανόηση τους μπορεί να επιφέρει σύγχυση. Επίσης, υπάρχει ο κίνδυνος της εσφαλμένης ερμηνείας της οπτικής πληροφορίας.

5.2 Διερευνητική Ανάλυση Δεδομένων

Ο όρος Διερευνητική Ανάλυση Δεδομένων (ΔΑΔ) (Exploratory Data Analysis (EDA)) προτάθηκε από τον Αμερικανό στατιστικολόγο Tukey (1977). Η συνεισφορά του ήταν καθοριστική και άλλαξε το τοπίο της περιγραφικής Στατιστικής. Ο ίδιος κατέγραψε παλιότερους, αλλά και πρότεινε νέους τρόπους απεικόνισης των δεδομένων. Στόχος είναι η αναζήτηση και εύρεση απεικονίσεων και ποσοτήτων, οι οποίες προσφέρουν κατανόηση και γνώση. Η φιλοσοφία της ΔΑΔ την διαφοροποιεί από άλλες προσεγγίσεις.

Ο Shneiderman (2002) αντιπαραβάλλει τη ΔΑΔ με τον Έλεγχο Υπόθεσης. Ο έλεγχος υπόθεσης, βασισμένος στην προβληματική του Fisher, προκρίνει την αρχική διατύπωση μιας υπόθεσης και την ακόλουθη πειραματική επιβεβαίωση ή απόρριψη της. Η προσέγγιση αυτή έχει το πλεονέκτημα ότι η υπόθεση εδράζεται σε θεωρητικό υπόβαθρο και καθοδηγείται από αυτό. Επίσης, το πείραμα διεξάγεται με επιλογή μεταβλητών, οπότε επιτυγχάνεται μείωση του χώρου του προβλήματος. Η φειδωλή επιλογή δεδομένων, ο καθορισμός των συνθηκών και η ακριβής μέτρηση επιτρέπουν την επανάληψη του πειράματος με την αποκόμιση των ίδιων αποτελεσμάτων. Τα αποτελέσματα γενικεύονται και επαληθεύουν την υπόθεση. Η προσέγγιση αυτή επικρίνεται, με το επιχείρημα ότι οι ελεγχόμενες εργαστηριακές συνθήκες απέχουν πολύ από την πραγματικότητα, η οποία είναι πολύ πιο σύνθετη. Επίσης, επικρίνεται με το επιχείρημα ότι η επιλογή ανεξάρτητων μεταβλητών μπορεί να αποκλείσει σημαντικές μεταβλητές που επηρεάζουν το αποτέλεσμα. Επιπλέον, το γεγονός ότι ο ερευνητής διατυπώνει εκ των προτέρων την υπόθεση του, τον ωθεί να βρει τρόπους να την επιβεβαιώσει.

Η μεθοδολογία της ΔΑΔ έχει διαφορετική αφετηρία. Ο ερευνητής συγκεντρώνει μεγάλους όγκους δεδομένων, τα επεξεργάζεται και αναζητά ενδιαφέροντα πρότυπα. Δεν απαιτείται καμία εκ των προτέρων διατύπωση υπόθεσης. Τα πρότυπα μπορεί να αποκαλύπτουν σύνθετες σχέσεις μεταξύ των δεδομένων, οι οποίες μέχρι τώρα ήταν άγνωστες. Η άντληση συμπερασμάτων απευθείας από τα δεδομένα, χωρίς την προηγούμενη διατύπωση υποθέσεων, αυξάνει τον βαθμό αντικειμενικότητας. Η σύνθετη ανάλυση των δεδομένων που απαιτείται καθίσταται εφικτή χάρη στις νέες δυνατότητες που προσφέρει η σύγχρονη πληροφορική. Ο Keim (2002) συνοψίζει μερικά από τα πλεονεκτήματα της ΔΑΔ:

- Αξιοποιεί και ενσωματώνει στην αναλυτική διαδικασία την ανθρώπινη δημιουργικότητα και αντίληψη.
- Είναι ιδιαίτερα χρήσιμη όταν δεν είναι γνωστές οι ιδιότητες των δεδομένων και όταν οι σκοποί της διερεύνησης είναι ασαφείς.
- Ο αναλυτής μπορεί άμεσα να αλλάξει τους στόχους της διερεύνησης.
- Η ανακάλυψη ιδιοτήτων των δεδομένων με γραφικό τρόπο μπορεί να αποτελέσει προστάδιο για τη διατύπωση υποθέσεων.
- Μπορεί να χειριστεί μη ομογενή και θορυβώδη δεδομένα.
- Δεν απαιτεί τη γνώση σύνθετων μαθηματικών ή στατιστικών αλγορίθμων.
- Μπορεί να αποδώσει αποτελέσματα όταν οι στατιστικοί αλγόριθμοι αποτύχουν.
- Προσφέρει μεγαλύτερο βαθμό εμπιστοσύνης στα αποτελέσματα.

Η κριτική που ασκήθηκε στη ΔΑΔ επικεντρώνεται στο γεγονός ότι η εύρεση σχέσεων μεταξύ των δεδομένων δεν συνεπάγεται αυτόματα την αποκάλυψη της σχέσης αίτιου και αιτιατού. Αν και στα πλαίσια της λογικής της ΔΑΔ δεν διατυπώνεται κάποια εκ των προτέρων υπόθεση, η χρήση γραφικών μέσων για την επαλήθευση μιας υπόθεσης δεν μπορεί να αποκλειστεί. Ο Mazza (2009) επισημαίνει ότι γραφικά μέσα μπορούν να χρησιμοποιηθούν για την επιβεβαίωση ή απόρριψη μιας υπόθεσης.

Ο Shneiderman (2002) αντιπαραβάλλει και τη χρήση στατιστικών μεθόδων με την οπτικοποίηση των δεδομένων. Η εφαρμογή στατιστικών μεθόδων εξασφαλίζει μια σαφή, μαθηματικά θεμελιωμένη ανάλυση των δεδομένων, καθώς και μια αντικειμενικότητα στη μεθοδολογία. Η ισχύς των στατιστικών μεθόδων όμως περιορίζεται από τη διατύπωση παραδοχών και υποθέσεων. Για παράδειγμα, σε μια στατιστική ανάλυση μπορεί ο ερευνητής να υποθέσει ότι η σχέση μεταξύ των δεδομένων είναι γραμμική και να εφαρμόσει μια γραμμική μέθοδο. Εάν η πραγματική σχέση μεταξύ των δεδομένων δεν είναι γραμμική, αλλά π.χ. εκθετική, η μέθοδος πιθανώς δεν θα μπορέσει να αναγνωρίσει τη σχέση. Η ανάπτυξη της πληροφορικής κατέστησε εφικτή την ταχύτατη εκτέλεση υπολογισμών και κατά συνέπεια την εφαρμογή πολύ πιο σύνθετων στατιστικών τεχνικών. Ταυτόχρονα όμως άνοιξε και νέους ορίζοντες στην οπτικοποίηση των δεδομένων. Οι νέες μέθοδοι οπτικοποίησης επιτρέπουν τη διαδραστική απεικόνιση σύνθετων εξαρτήσεων των δεδομένων και την εξαγωγή συμπερασμάτων. Επιπλέον, οι οπτικές μέθοδοι τροφοδοτούν όχι μόνον τη λογική, αλλά και τη διαίσθηση, καθώς και το ένστικτο του ερευνητή, και μπορεί να τον οδηγήσουν σε αναλύσεις που θα του αποφέρουν αποκαλυπτικά συμπεράσματα.

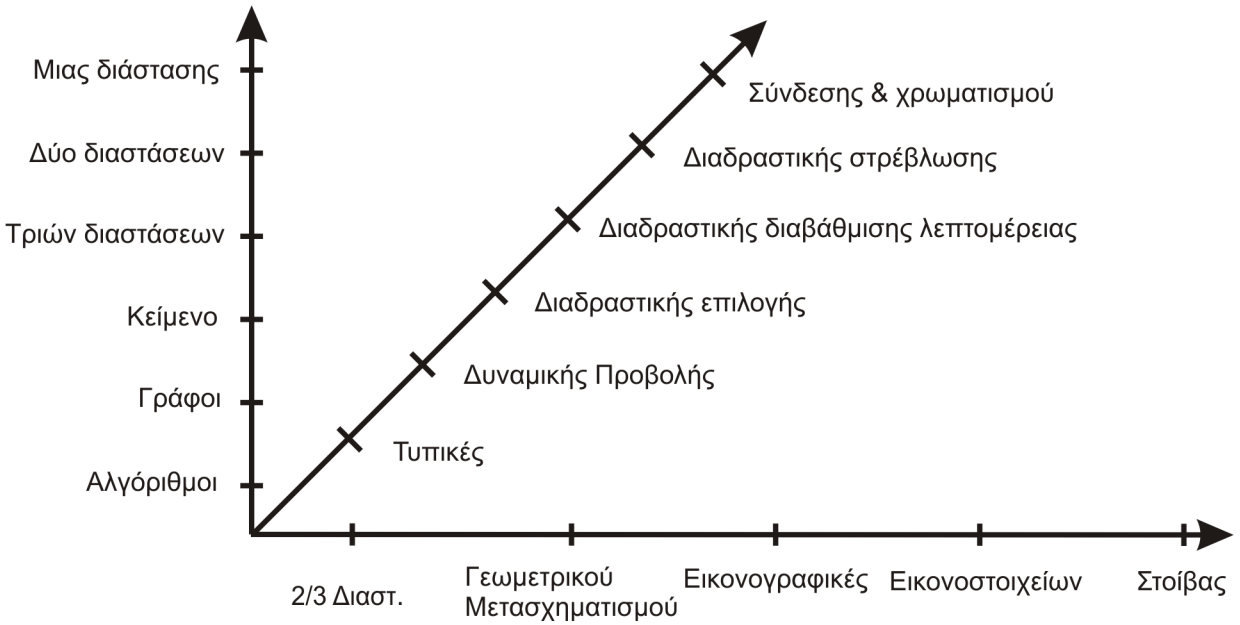
5.3 Ταξινόμηση μεθόδων οπτικοποίησης δεδομένων

Στη μακρόχρονη διαδρομή της απόπειρας του ανθρώπου να απεικονίσει με γραφικό τρόπο τα δεδομένα, έχουν προταθεί διάφορες τεχνικές. Οι τεχνικές αυτές διαφέρουν ποικιλότροπα μεταξύ τους και το πλήθος τους είναι πολύ μεγάλο. Σε μια προσπάθεια συστηματοποίησης του πεδίου, ερευνητές επιχείρησαν να εντάξουν τις τεχνικές αυτές σε κατηγορίες. Κατά καιρούς προτάθηκαν διάφοροι τρόποι ταξινόμησης. Ο δημοφιλέστερος ίσως τρόπος ταξινόμησης είναι αυτός που προτάθηκε από τους Keim and Kriegel (1996) και επαναδιατυπώθηκε από τον Keim(2002).

Σύμφωνα με τον Keim(2002), οι τεχνικές οπτικής αναπαράστασης μπορούν να ταξινομηθούν με βάση τρία κριτήρια:

- τον τύπο των δεδομένων,
- την τεχνική οπτικοποίησης,
- την τεχνική αλληλεπίδρασης και στρέβλωσης.

Κάθε τεχνική οπτικοποίησης μπορεί να συνδυαστεί με κάθε τεχνική αλληλεπίδρασης για κάθε τύπο δεδομένων. Ένα σύστημα μπορεί να χρησιμοποιεί συνδυασμούς τεχνικών οπτικοποίησης και αλληλεπίδρασης για διάφορους τύπους δεοδμένων. Η συνδιαστική αυτή δυνατότητα αποδίδεται διαγραμματικά στο Σχήμα 5.2, όπου οι τρείς άξονες είναι ορθογώνιοι και αντιστοιχούν στα τρία κριτήρια ταξινόμησης



Σχήμα 5.2 *Ταξινόμηση τεχνικών οπτικοποίησης δεδομένων*

Ως προς τον τύπο τους, τα δεδομένα που θα οπτικοποιηθούν μπορεί να είναι:

- Μονοδιάστατα. Τα δεδομένα έχουν μια διάσταση, πχ χρονικά δεδομένα.
- Δυσδιάστατα. Τα δεδομένα έχουν δύο διαστάσεις, πχ γεωγραφικά σημεία με γεωγραφικό μήκος και πλάτος.
- Πολυδιάστατα. Τα δεδομένα έχουν πολλές διαστάσεις, πχ πίνακες σχεσιακών βάσεων δεδομένων με πολλές στήλες.
- Κείμενο και Υπερκείμενο. Τα δεδομένα είναι αδόμητα και δεν μπορούν να εκφραστούν σε σχέση με διαστάσεις.
- Ιεραρχίες και Γράφοι. Τα αντικείμενα συνδέονται μεταξύ τους με σχέσεις. Μπορούν να αναπαρασταθούν με ένα γράφο, όπου τα αντικείμενα είναι οι κόμβοι και οι σχέσεις είναι οι ακμές που τους συνδέουν.
- Αλγόριθμοι και λογισμικό. Αφορά δεδομένα ροής πληροφοριών σε ένα πρόγραμμα.

Οι τεχνικές οπτικοποίησης υπάγονται στις παρακάτω κατηγορίες:

- Τυπικές δύο ή τριών διαστάσεων (Standard 2D/3D displays). Σχετικά απλές τεχνικές, που συνήθως εφαρμόζονται στα πρώτα στάδια της ανάλυσης. Δεν είναι κατάλληλες για την οπτικοποίηση σύνθετων δομών.

- Γεωμετρικού Μετασχηματισμού (Geometrically Tranformed). Πολυδιάστατα δεδομένα μετασχηματίζονται και προβάλλονται με γεωμετρικό τρόπο, ώστε να αποκαλυφθούν πιθανές σχέσεις τους. Περίπτωση τεχνικής γεωμετρικού μετασχηματισμού είναι τα διαγράμματα παράλληλων συντεταγμένων.
- Εικονικογραφικές (Iconic Displays). Κάθε παρατήρηση (αντικείμενο) αντιστοιχίζεται σε μια εικόνα και κάθε τιμή της παρατήρησης αντιστοιχίζεται με ένα χαρακτηριστικό της εικόνας, πχ σχήμα, μέγεθος, χρώμα κλπ. Στα πρόσωπα του Chernoff, κάθε παρατήρηση αντιστοιχεί σε ένα πρόσωπο, και τα χαρακτηριστικά του προσώπου (μύτη, αυτιά κλπ.) εκφράζουν τις τιμές των μεταβλητών. Ο ερευνητής συγκρίνει τις εικόνες για να βρει ομοιότητες και διαφορές. Τεχνικές κατάλληλες για μέτριο πλήθος δεδομένων και σχετικά μικρό αριθμό μεταβλητών.
- Εικονοστοιχείων (Dense Pixel Displays). Κάθε τιμή των δεδομένων αντιστοιχίζεται σε ένα pixel, το οποίο χρωματίζεται ανάλογα με την τιμή. Τα εικονοστοιχεία μιας διάστασης τοποθετούνται σε γειτονικές περιοχές. Τεχνικές κατάλληλες για απεικόνιση περίπου 1.000.000 τιμών, με αδυναμίες όμως στον εντοπισμό σύνθετων δομών δεδομένων.
- Στοίβας (ή ιεραρχικές) (Stacked Displays). Η παρουσίαση των δεδομένων γίνεται στη βάση μιας ιεράρχησης. Οι τύποι των ιεραρχήσεων ποικίλουν. Παράδειγμα τέτοιας τεχνικής είναι η Dimensional Stacking, όπου ο χώρος απεικόνισης χωρίζεται σε τμήματα ανάλογα με δύο διαστάσεις και εντός αυτών των τμημάτων γίνεται απεικόνιση των δεδομένων ανάλογα με δύο άλλες διαστάσεις. Άλλο παράδειγμα, με τελείως διαφορετικό τρόπο ιεράρχησης, είναι τα Δενδρογράμματα, που αποτυπώνουν τη διαδικασία διαδοχικής συγχώνευσης συστάδων και χρησιμοποιούνται στην Ανάλυση Συστάδων (Κεφάλαιο 11).

Τέλος, έχουν προταθεί ως τεχνικές οπτικοποίησης μέθοδοι μείωσης διαστάσεων, όπως η μέθοδος Principal Components Analysis και οι Αυτοοργανούμενοι Χάρτες (Self Organizing Maps)

Αναφορικά με τον τρόπο αλληλεπίδρασης και στρέβλωσης οι τεχνικές κατηγοριοποιούνται ως:

- Δυναμικής προβολής (Dynamic Projections). Συνίσταται στη μεταβολή του τρόπου προβολής των δεδομένων.
- Διαδραστικής επιλογής (Interactive Filtering). Επιτρέπει την τμηματοποίηση των δεδομένων και την επικέντρωση σε ένα υποσύνολο. Το υποσύνολο των δεδομένων μπορεί να προκύψει είτε με την εκτέλεση κάποιου ερωτήματος είτε με την άμεση επιλογή από τον χρήστη.
- Διαδραστικής Διαβάθμισης Λεπτομέρειας (Interactive Zooming). Είναι η δυνατότητα προβολής σε διαφορετικό βαθμό λεπτομέρειας. Τα αντικείμενα μπορεί να μεγεθυνθούν ή μπορεί να προβληθεί διαφορετικού τύπου πληροφορία, όπως πχ κείμενο.
- Διαδραστικής στρέβλωσης (Interactive Distortion). Συνίσταται στην προβολή του συνόλου των δεδομένων με χαμηλό βαθμό λεπτομέρειας, με ταυτόχρονη προβολή τμήματος των δεδομένων με υψηλό βαθμό λεπτομέρειας.
- Διαδραστικής Σύνδεσης και Χρωματισμού (Interactive Linking and Brushing). Είναι ο συνδυασμός διαφορετικών τεχνικών οπτικοποίησης. Για παράδειγμα, σε ένα σύνολο διαγραμμάτων διασποράς μπορεί να χρωματιστούν και να συνδεθούν ορισμένα σημεία σε όλα τα διαγράμματα.

Περισσότερες λεπτομέρειες σχετικά με την ταξινόμηση των μεθόδων οπτικοποίησης μπορεί να αναζητήσει ο αναγνώστης στο έργο του Keim (2002).

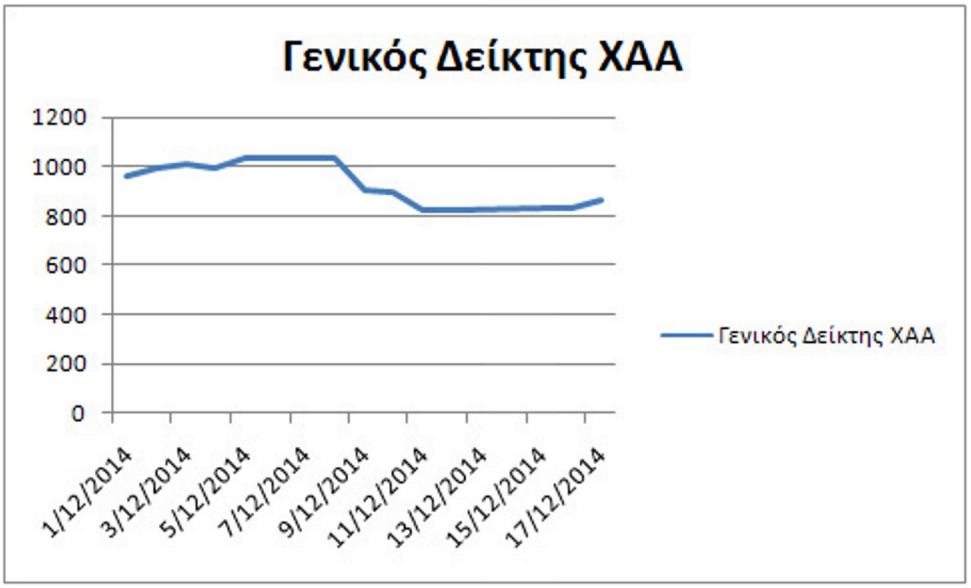
5.4 Τεχνικές Απεικόνισης Δεδομένων

Υπάρχει ένα μεγάλο πλήθος τεχνικών για την απεικόνιση των δεδομένων. Θα ακολουθήσει μια σύντομη παρουσίαση μερικών από τις βασικότερες τεχνικές.

5.4.1 Τυπικές

5.4.1.1 Γραφήματα Γραμμής.

Πρόκειται για μια από τις απλούστερες μορφές γραφικών και αποτυπώνει τη σχέση μιας μεταβλητής με μίαν άλλη. Συνήθως χρησιμοποιούνται για την απεικόνιση της μεταβολής μιας ποσότητας με την πάροδο του χρόνου. Παράδειγμα γραφήματος γραμμής βρίσκεται στο Σχήμα 5.3 και δείχνει τη διακύμανση του Γενικού Δείκτη ΧΑΑ από 1/12/2014 έως 17/12/2014.



Σχήμα 5.3 *Γενικός Δείκτης ΧΑΑ για το διάστημα 1/12/2014-17/12/2014*

5.4.1.2 Ραβδογράμματα

Επίσης ένας πολύ συνηθισμένος τύπος γραφήματος. Χρησιμοποιείται συνήθως για τη σύγκριση της ποσότητας διαφορετικών ομάδων ή κατηγοριών. Παράδειγμα, με συγκριτικά στοιχεία πωλήσεων ανά κατηγορία προϊόντος, παρουσιάζεται στο Σχήμα 5.4.

Σχήμα 5.4 *Συγκριτικά στοιχεία πωλήσεων ανά κατηγορία προϊόντος.*

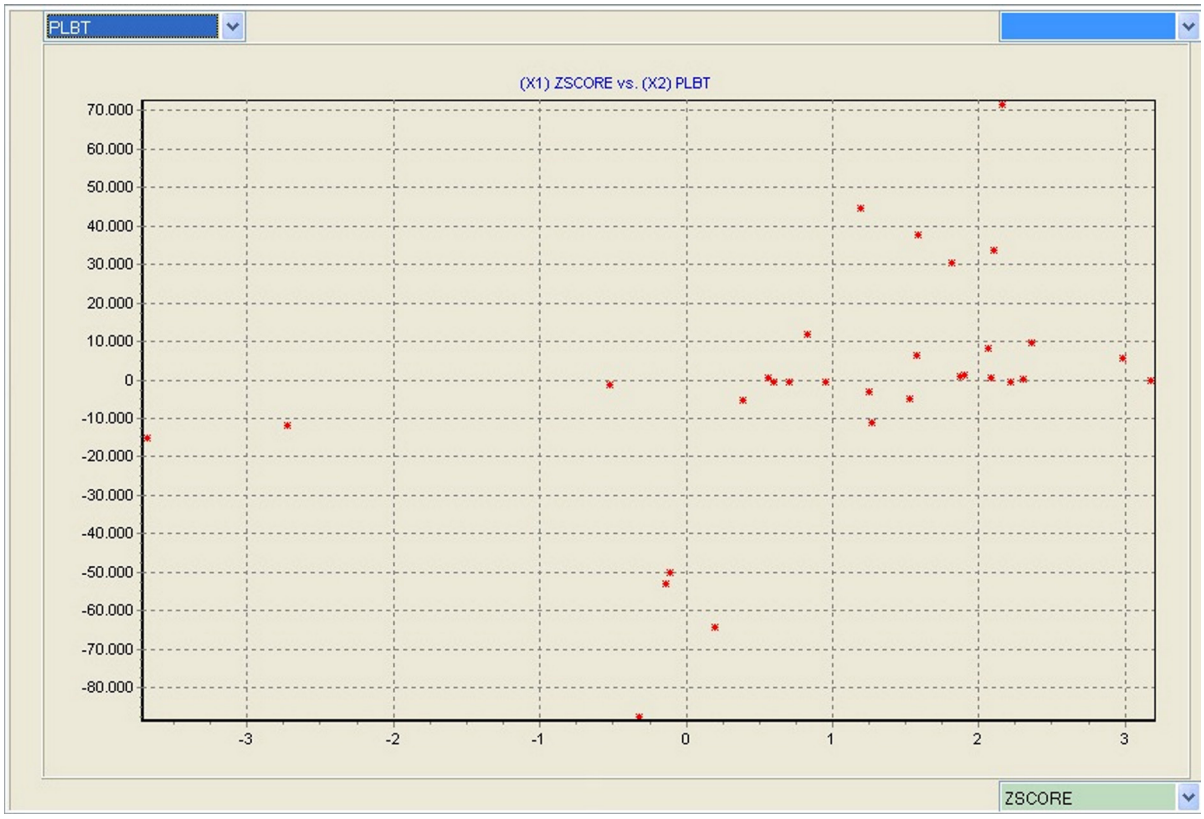
5.4.1.3 Γραφήματα Πίτας

Τα διαγράμματα πίτας χρησιμοποιούνται για τη σύγκριση τμημάτων ενός συνόλου. Έχουν δεχτεί κριτική, γιατί το ανθρώπινο μάτι δυσκολεύεται να συγκρίνει επιφάνειες και γωνίες, καθώς και φέτες που έχουν παρόμοιο μέγεθος αλλά δεν γειτονεύουν. Ο χρήστης μπορεί να επιλέξει εναλλακτικά τα ραβδογράμματα. Στο Σχήμα 5.4 εμφανίζονται τα δεδομένα του Σχήματος 5.5 με μορφή πίτας.

Σχήμα 5.5 *Στοιχεία πωλήσεων ανά κατηγορία προϊόντος (μορφή Γραφήματος Πίτας)*

5.4.1.4 Διαγράμματα Διασποράς

Τα Διαγράμματα Διασποράς είναι ίσως τα συνηθέστερα διαγράμματα, τα οποία θα συναντήσει ο χρήστης σε λογισμικά Εξόρυξης Δεδομένων. Είναι δισδιάστατες απεικονίσεις, που δείχνουν τη σχέση μεταξύ δύο μεταβλητών. Οι δύο μεταβλητές αντιστοιχούν στους δύο άξονες. Κάθε παρατήρηση τοποθετείται ως σημείο στο επίπεδο, ανάλογα με τις τιμές της στις δύο μεταβλητές. Τα διαγράμματα διασποράς δίνουν αίσθηση της εξάρτησης των μεταβλητών, το πόσο διασκορπισμένα είναι τα δεδομένα και το εάν υπάρχουν πρότυπα στην κατανομή των δεδομένων. Το Σχήμα 5.6 δείχνει ένα διάγραμμα διασποράς ανάμεσα στα Κέρδη προ φόρων (Profit Loss Before Taxation (PLBT)) και στον αριθμοδείκτη ZScore του Altman, ο οποίος εκφράζει την οικονομική ευρωστία της επιχείρησης. Παρατηρούμε ότι οι επιχειρήσεις με ζημίες (PLBT < 0) τείνουν να έχουν αρνητικές ή μικρές θετικές τιμές ZScore, ενώ οι επιχειρήσεις που δεν έχουν ζημίες ή έχουν κέρδη τείνουν να έχουν θετικές τιμές Z Score. Το διάγραμμα κατασκευάστηκε με το ανοικτό λογισμικό εξόρυξης δεδομένων Tanagra.



Σχήμα 5.6 *Διάγραμμα Διασποράς μεταξύ κερδών (PLBT) και Z Score*

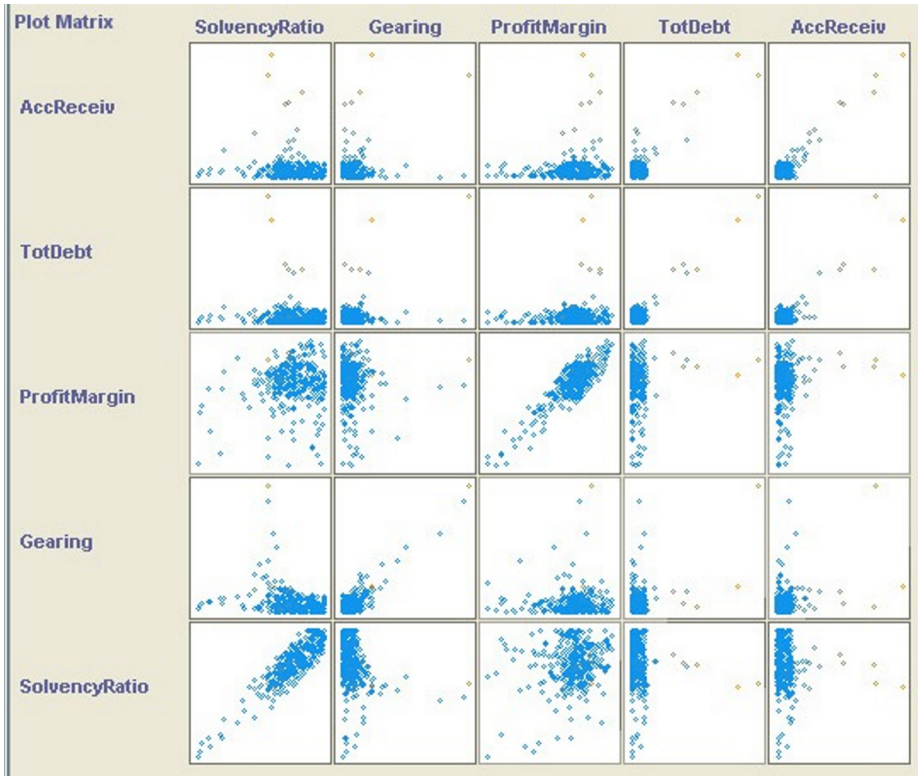
5.4.2 Γεωμετρικού Μετασχηματισμού

5.4.2.1 Πίνακας Διαγραμμάτων Διασποράς (Scatter plot Matrix)

Στα Διαγράμματα Διασποράς, εάν οι μεταβλητές είναι περισσότερες από δύο, τότε μπορεί να κατασκευαστεί ένας πίνακας διαγραμμάτων, όπου συνδυάζονται με όλους τους δυνατούς τρόπους οι μεταβλητές ανά ζεύγη. Διαγωνίως τα διαγράμματα αντιστοιχούν στον συνδυασμό κάθε μεταβλητής με τον εαυτό της, οπότε δεν υπάρχει ουσιαστική πληροφορία. Επίσης, η διαγώνιος χωρίζει τον πίνακα σε δύο τμήματα με επαναλαμβανόμενη πληροφορία, εφόσον το διάγραμμα των μεταβλητών X-Y και των μεταβλητών Y-X είναι το ίδιο μετά από περιστροφή. Στο Σχήμα 5.7 παρουσιάζεται ο Πίνακας Διαγραμμάτων Διασποράς πέντε οικονομικών αριθμοδεικτών. Το διάγραμμα κατασκευάστηκε με το ανοικτό λογισμικό εξόρυξης δεδομένων WEKA

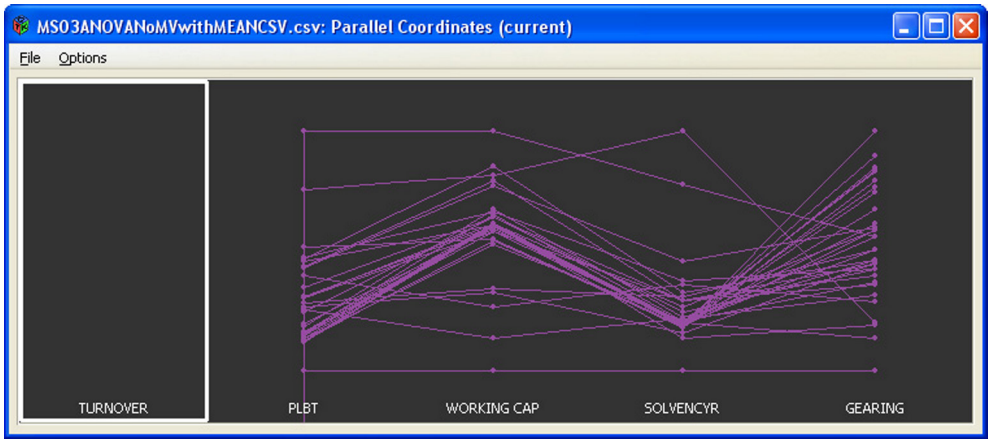
5.4.2.2 Διαγράμματα Παράλληλων Συντεταγμένων

Τα διαγράμματα Παράλληλων Συντεταγμένων προτάθηκαν από τον Inselberg (1985) και τους Inselberg and Dimsdale (1990). Στα διαγράμματα αυτού του τύπου, κάθε διάσταση (μεταβλητή) των δεδομένων αντιστοιχίζεται με έναν άξονα. Οι άξονες τοποθετούνται παράλληλα σε ίσες αποστάσεις. Κάθε αντικείμενο συμβολίζεται με μια τεθλασμένη γραμμή. Για κάθε αντικείμενο και για κάθε μεταβλητή τοποθετείται ως σημείο η αντίστοιχη τιμή στον άξονα της διάστασης, τα σημεία ενώνονται με γραμμές και προκύπτει η τεθλασμένη γραμμή του αντικειμένου. Θεωρητικά μπορεί να αναπαρασταθεί απεριόριστος αριθμός διαστάσεων, όμως για απεικόνιση σε οθόνη υπολογιστή οι διαστάσεις δεν μπορούν να υπερβαίνουν έναν σχετικά μικρό αριθμό, πχ δέκα. Επίσης, εάν τα αντικείμενα είναι πολλά προκύπτουν πολλές γραμμές, που δεν ξεχωρίζουν μεταξύ τους. Οι άξονες μπορεί να βαθμονομηθούν με βάση τις αρχικές τιμές ή μπορεί να γίνει μετασχηματισμός των τιμών και αναγωγή τους σε κάποια κοινή περιοχή τιμών.



Σχήμα 5.7 *Πίνακας Διαγραμμάτων Διασποράς*

Τα διαγράμματα Παράλληλων Συντεταγμένων μπορούν να αποκαλύψουν ενδιαφέροντα πρότυπα μέσω της μελέτης των τεθλασμένων γραμμών. Ακραίες τιμές και εξαιρέσεις είναι εύκολο να εντοπιστούν. Σύγκριση των θέσεων των σημείων σε διαφορετικούς άξονες και παράλληλα ευθύγραμμα τμήματα μπορεί να καταδείξουν γραμμικές συσχετίσεις μεταξύ των αντίστοιχων μεταβλητών. Ένας περιορισμός είναι ότι για να αναδειχθεί αυτή η πληροφορία θα πρέπει οι άξονες να είναι γειτονικοί. Για τον λόγο αυτό, θα πρέπει να διαταχθούν οι άξονες με όλους τους δυνατούς συνδυασμούς ώστε να αποκαλυφθούν οι σχέσεις. Σε περίπτωση που οι διαστάσεις είναι πολλές, κάτι τέτοιο είναι ανέφικτο. Μια λύση είναι να αξιολογηθεί η σημαντικότητα των μεταβλητών σύμφωνα με κάποιο κριτήριο, και να διαταχθούν οι άξονες σε φθίνουσα σειρά σημαντικότητας. Στο Σχήμα 5.8 παρουσιάζεται διάγραμμα Παράλληλων Συντεταγμένων για 30 επιχειρήσεις και για τέσσερεις αριθμοδείκτες. Το διάγραμμα δημιουργήθηκε με το λογισμικό GGobi.



Σχήμα 5.8 *Διάγραμμα Παράλληλων Συντεταγμένων για τέσσερεις αριθμοδείκτες*

5.4.2.3 HyperSlice

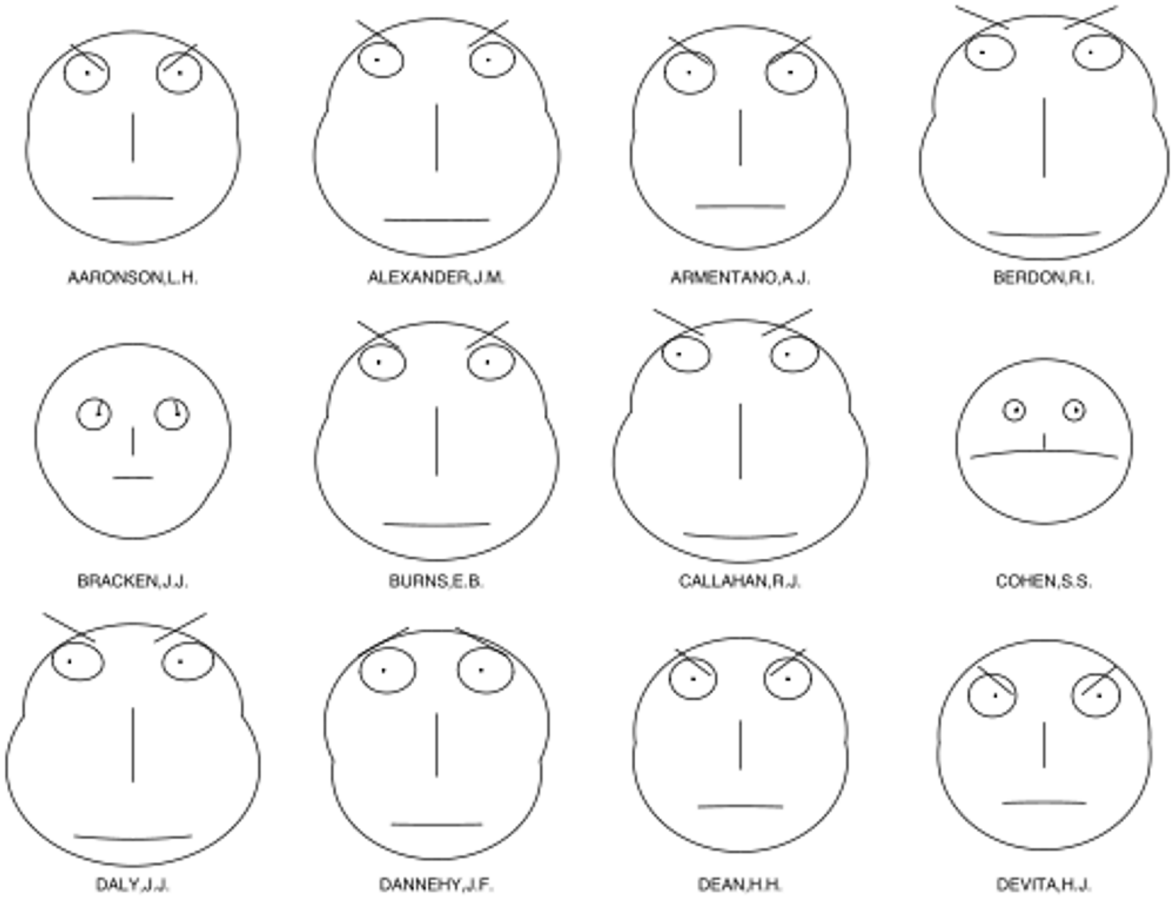
Η τεχνική HyperSlice προτάθηκε από τους Van Wijk and Van Liere (1994) και μπορεί να θεωρηθεί εξέλιξη του πίνακα των Διαγραμμάτων Διασποράς. Βασική ιδέα είναι η αναπαράσταση μιας συνάρτησης πολλών διαστάσεων ως ένας πίνακας δισδιάστατων ορθογώνιων «φετών». Κάθε φορά καθορίζεται ένα σημείο ενδιαφέροντος και μια περιοχή γύρω από αυτό. Ο χρήστης επικεντρώνει σε ένα σημείο ενός χώρου N διαστάσεων, το οποίο καλείται τρέχων σημείο. Για κάθε διάσταση ορίζεται ένα εύρος τιμών w_i, όπου $i = 1,...,N$. Για κάθε διάσταση i, η περιοχή τιμών ενδιαφέροντος είναι $R_i = [c_i-w_i/2, c_i+w_i/2]$. Μια δισδιάστατη φέτα S_{kl} με $k<>l$ είναι μια οπτική αναπαράσταση της συνάρτησης $f(x)$, όπου το $x_k \epsilon R_k$, και το $x_l \epsilon R_l$ μεταβάλλονται παίρνοντας τιμές από όλο το πεδίο ορισμού τους, ενώ τα υπόλοιπα x_i παραμένουν ίσα με c_i. Ο οριζόντιος και κατακόρυφος άξονας του διαγράμματος αντιστοιχούν στα x_k και x_l.
Ένα διάγραμμα HyperSlice είναι ένας πίνακας NXN από διαγράμματα $<I,j>$ όπου $1<=i,j<=N$. Στον πίνακα διαγωνίως βρίσκονται μονοδιάστατα διαγράμματα G_k, όπου μεταβάλλονται οι τιμές της μιας διάστασης, ενώ οι άλλες παραμένουν σταθερές. Εκτός διαγωνίου βρίσκονται διαγράμματα φετών S_{kl}. Το τρέχων σημείο βρίσκεται πάντα στο κέντρο των διαγραμμάτων. Εκατέρωθεν της διαγωνίου βρίσκονται ίδια διαγράμματα περιστρεμμένα κατά 90 μοίρες. Παραδείγματα διαγραμμάτων HyperSlice υπάρχουν στο Van Wijk and Van Liere (1994).

5.4.3 Εικονογραφικές

5.4.3.1 Πρόσωπα Chernoff

Η τεχνική αυτή προτάθηκε το 1973 από τον Chernoff (1973). Βασική ιδέα είναι να αναπαρασταθούν τα δεδομένα με σκίτσα ανθρώπινων προσώπων. Κάθε διάσταση (μεταβλητή) των δεδομένων αντιστοιχίζεται με ένα χαρακτηριστικό του προσώπου (μάτια, μύτη, στόμα κλπ.). Το σχήμα και το μέγεθος του χαρακτηριστικού εξαρτάται από τις τιμές της μεταβλητής. Οι τιμές δεν εφαρμόζονται αυτούσιες, αλλά υπόκεινται σε μετασχηματισμό, ώστε να σχεδιαστούν σχετικά «φυσιολογικά» χαρακτηριστικά. Για κάθε παρατήρηση κατασκευάζεται ένα πρόσωπο, που στα χαρακτηριστικά του αποτυπώνει τις τιμές των μεταβλητών της παρατήρησης. Χρησιμοποιώντας 18 χαρακτηριστικά του προσώπου μπορούμε να κωδικοποιήσουμε αντίστοιχο πλήθος μεταβλητών. Η αντιστοίχιση των μεταβλητών σε χαρακτηριστικά προσώπου μπορεί να επιλεγεί από τον χρήστη ή να είναι τυχαία. Ωστόσο, χρήσιμο είναι η εικόνα να σηματοδοτεί κάποιο νόημα. Για παράδειγμα, η επιτυχία ή η αποτυχία μπορεί να κωδικοποιηθεί με το σχήμα του στόματος (χαμογελαστό ή σκυθρωπό).

Τα πρόσωπα του Chernoff βοηθούν στη γρήγορη αναγνώριση ιδιοτήτων των δεδομένων. Ο ανθρώπινος εγκέφαλος είναι εκπαιδευμένος να αναγνωρίζει χαρακτηριστικά προσώπων και να διαφοροποιεί με τον τρόπο αυτό τα πρόσωπα. Έτσι, με ταχεία οπτική παρατήρηση μπορούν εύκολα να αναγνωριστούν συστάδες ομοειδών αντικειμένων, αντικείμενα - εξαιρέσεις με ακραίες τιμές, καθώς και χρονικά μεταβαλλόμενες τάσεις. Μειονέκτημα της μεθόδου είναι ότι δεν μπορούν να αναπαρασταθούν δεδομένα με πολλές διαστάσεις ή με πολλές παρατηρήσεις. Επίσης, ο ανθρώπινος εγκέφαλος δεν αξιολογεί όλα τα χαρακτηριστικά του προσώπου με την ίδια βαρύτητα και τείνει να ομαδοποιεί πρόσωπα με βάση ορισμένα χαρακτηριστικά, ενώ υποβαθμίζει κάποια άλλα. Επιπλέον, η μέθοδος δεν παρουσιάζει τις πραγματικές τιμές. Στο Σχήμα 5.9 παρουσιάζεται παράδειγμα προσώπων Chernoff.



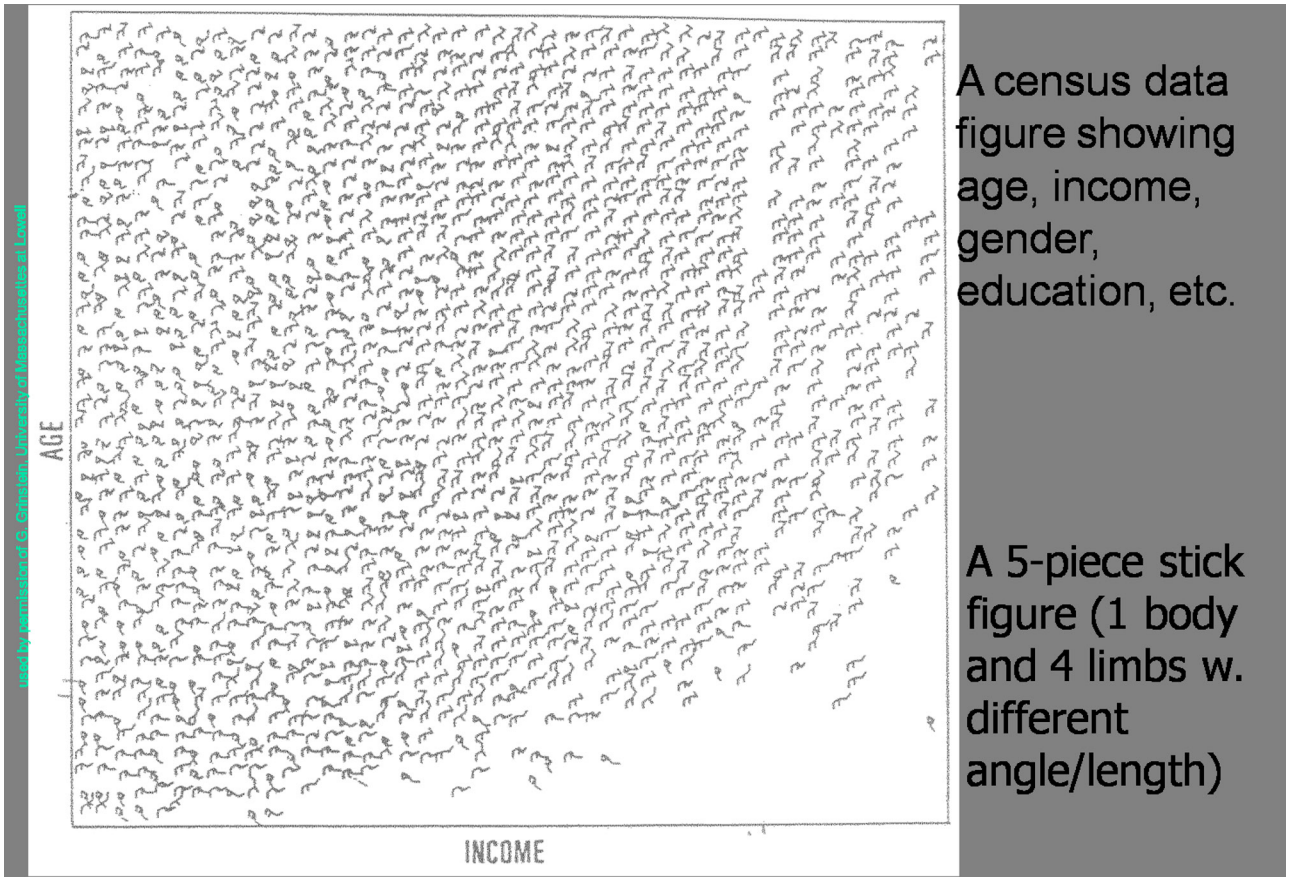
Σχήμα 5.9 *Πρόσωπα Chernoff (Αναπαραγωγή από Wikimedia Commons. Ιδιοκτήτης: Avenue)*

5.4.3.2 Εικόνες Stick Figure

Η τεχνική προτάθηκε από τους Pickett and Grinstein (1988) και επαφίεται στις αντιληπτικές ικανότητες του χρήστη. Τα δεδομένα απεικονίζονται σε μια μοναδική αναπαράσταση. Αντίθετα, γραφήματα όπως ο πίνακας των Διαγραμμάτων Διασποράς, διασπούν την απεικόνιση σε υποχώρους και απευθύνονται κυρίως στις γνωστικές ικανότητες του χρήστη. Η αναπαράσταση των δεδομένων γίνεται με χρήση μικρών πολυγωνικών γραμμών, που κωδικοποιούν την πληροφορία σχετικά με τις τιμές των δεδομένων στο σχήμα των γραμμών. Ειδικότερα, χρησιμοποιείται ένα ευθύγραμμο τμήμα για κάθε μεταβλητή. Στην αρχική εργασία των Pickett and Grinstein οπτικοποιήθηκαν δεδομένα με πέντε μεταβλητές. Μια μεταβλητή αντιστοιχίζεται με ένα ευθύγραμμο τμήμα, που αποτελεί το κύριο σώμα του γραφήματος. Οι υπόλοιπες μεταβλητές αντιστοιχίζονται με τα άλλα τέσσερα ευθύγραμμα τμήματα που αποτελούν τους βραχίονες και προσαρτώνται στο σώμα. Οι θέσεις των βραχιόνων, η σύνδεση τους δηλαδή στο σώμα ή μεταξύ τους, αντιστοιχούν σε δομές των δεδομένων. Οι γωνίες των βραχιόνων ελέγχονται από τις τιμές των μεταβλητών. Η κλίση του σώματος μπορεί να εξαρτάται από τις τιμές της πρώτης μεταβλητής. Οι τιμές μπορεί επίσης να κωδικοποιηθούν με χρήση του μήκους, του χρώματος και της λαμπρότητας των γραμμών. Για κάθε παρατήρηση δημιουργείται ένα γράφημα. Στη συνέχεια, τα γραφήματα τοποθετούνται μαζικά σε μια επιφάνεια. Η τοποθέτηση αυτών των σχημάτων δίνει στην επιφάνεια μια αίσθηση «υφής». Ο χρήστης αναγνωρίζει τμήματα της επιφάνειας με βάση την υφή τους και με τον τρόπο αυτό εντοπίζει δομές δεδομένων.

Ένας περιορισμός της μεθόδου σχετίζεται με το πλήθος των διαστάσεων που μπορεί να απεικονιστεί. Συνολικά μπορούν να κωδικοποιηθούν μέχρι και επτά μεταβλητές. Ένας άλλος περιορισμός είναι ότι απαιτείται εμπειρία από τον χρήστη στην κατανόηση των προτύπων υφής, αν και ο τρόπος και ο βαθμός κατανόησης παραμένει ως ερευνητικό ζητούμενο. Έχουν προταθεί επεκτάσεις της μεθόδου, οι οποίες κάνουν χρήση και του ήχου. Στο Σχήμα 5.10 παρουσιάζονται πέντε γραφήματα stick figure, τα οποία καθορίζονται από τον τρόπο που οι βραχίονες ενώνονται με τον κορμό. Στο Σχήμα 5.11 παρουσιάζεται επιφάνεια που δημιουργήθηκε με stick figures και αποτυπώνει μετεωρολογικά δεδομένα.

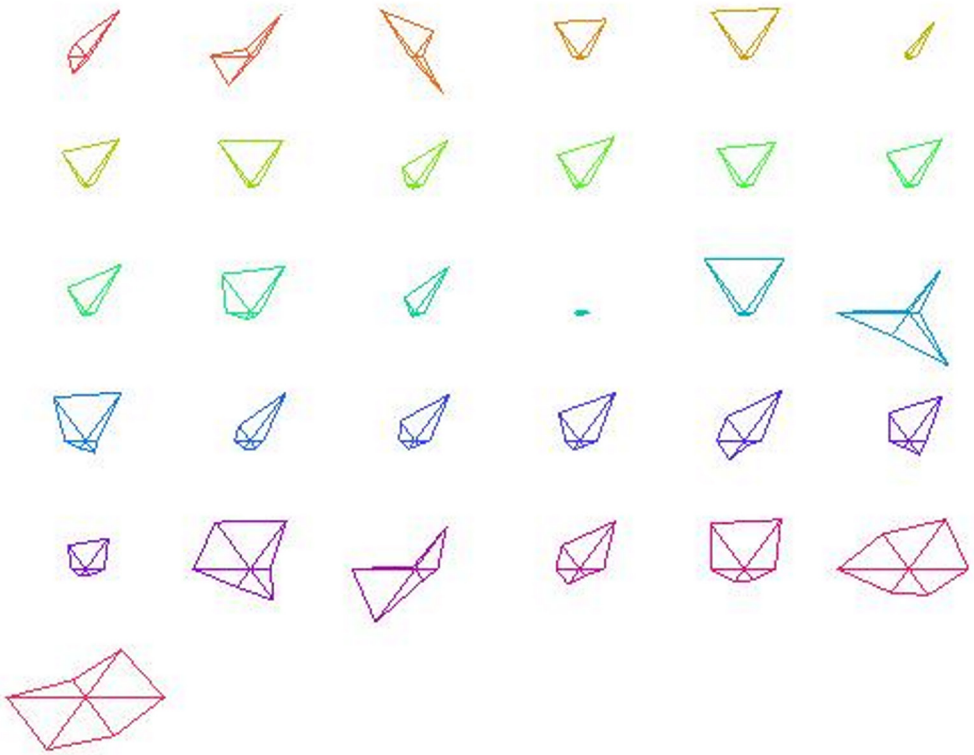
Σχήμα 5.10 *Stick figures*



Σχήμα 5.11. *Επιφάνεια με stick figures (Αναπαραγωγή από Slidewiki. Ιδιοκτήτης: sidraaslam)*

5.4.3.3 Διαγράμματα Αστέρων

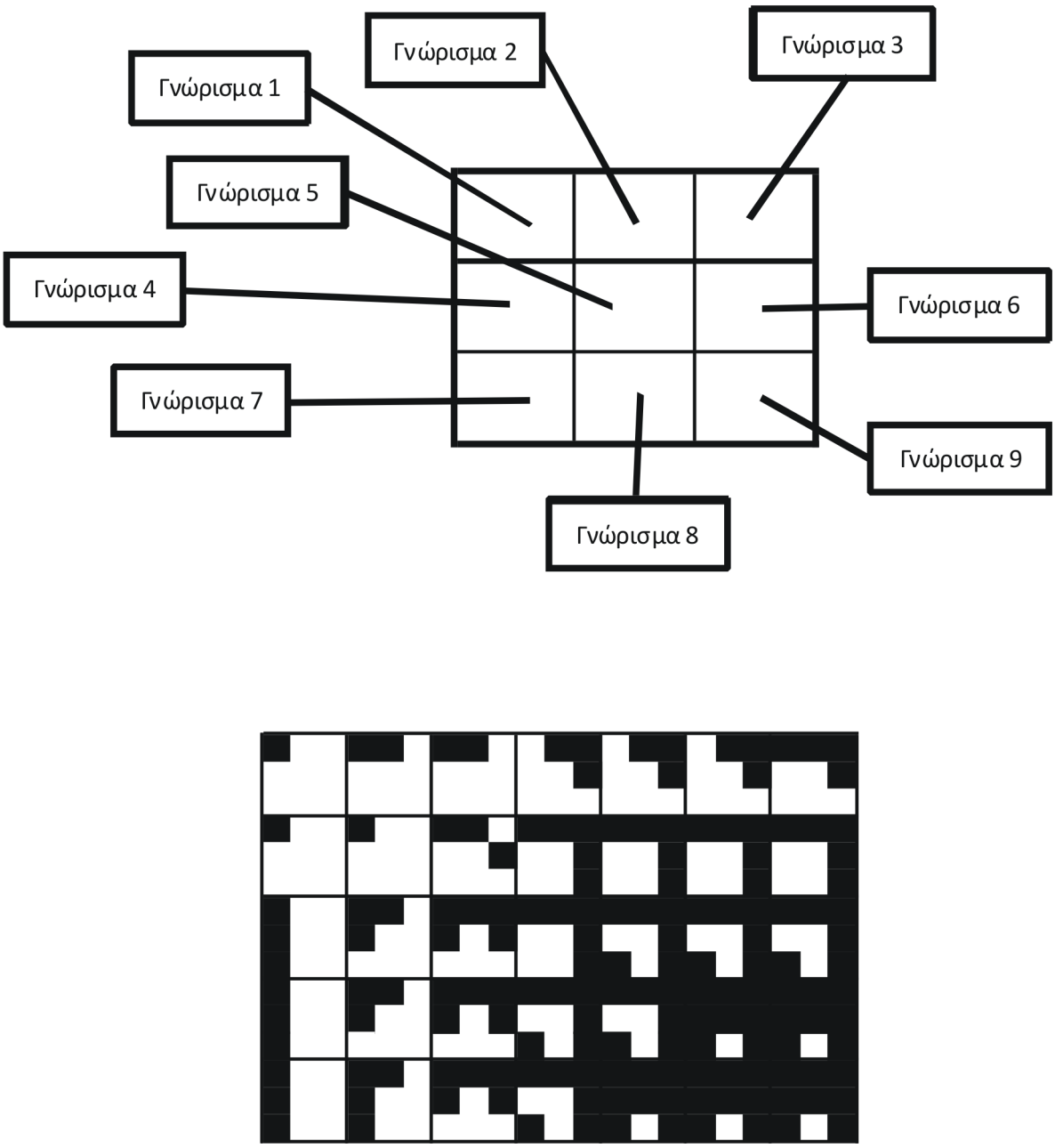
Τα Διαγράμματα Αστέρων είναι μια πολύ διαδεδομένη μέθοδος για την οπτικοποίηση δεδομένων με τη χρήση εικόνων. Η μέθοδος προτάθηκε από τους Chambers, Clevelant, Kleiner and Tuckey. (1983). Για την οπτικοποίηση χρησιμοποιείται το σχήμα του αστεριού, το οποίο έχει τόσες ακτίνες όσες είναι οι διαστάσεις των δεδομένων. Οι ακτίνες είναι τοποθετημένες σε ίσες αποστάσεις. Κατασκευάζεται ένα αστέρι για την κάθε παρατήρηση. Το μήκος της κάθε ακτίνας εξαρτάται από την τιμή της μεταβλητής για την κάθε παρατήρηση. Οι άκρες του αστεριού ενώνονται ώστε να δημιουργηθεί ένα πολύγωνο. Η μέθοδος λειτουργεί ικανοποιητικά για δεδομένα με λίγες διαστάσεις και σχετικά μικρό πλήθος παρατηρήσεων. Σημαντική είναι η σειρά με την οποία τοποθετούνται οι ακτίνες. Το Σχήμα 5.12 παρουσιάζει Διάγραμμα Αστέρων που δημιουργήθηκε με το λογισμικό XMVD Tool.



Σχήμα 5.12 *Διάγραμμα Αστέρα*

5.4.3.4 Τεχνική Shape Coding

Η τεχνική εισήχθηκε από τον Beddow (1990) για την οπτικοποίηση πολυδιάστατων δεδομένων και τον εντοπισμό σημαντικών προτύπων. Κάθε παρατήρηση κωδικοποιείται ως ένα μικρό παραλληλόγραμμο χωρισμένο σε τόσα τμήματα όσες είναι και οι διαστάσεις. Για κάθε παρατήρηση και για κάθε μεταβλητή, η αντίστοιχη τιμή χαρακτηρίζεται ως μεγάλη, μεσαία ή μικρή, σύμφωνα με κάποιο μέτρο. Ο Beddow προτείνει ως μέτρο τον συνδυασμό της μέσης τιμής με την τυπική απόκλιση. Ακολούθως, το τμήμα που αντιστοιχεί στη συγκεκριμένη διάσταση χρωματίζεται ως λευκό, γκρίζο ή μαύρο ανάλογα με τον χαρακτηρισμό της τιμής. Όλες οι παρατηρήσεις αποτυπώνονται με αυτόν τον τρόπο σε έναν πίνακα και ο ερευνητής μπορεί να αναζητήσει πρότυπα. Η μέθοδος λειτουργεί ικανοποιητικά για σχετικά μικρό αριθμό διαστάσεων αλλά για αρκετά μεγάλο αριθμό παρατηρήσεων. Στο Σχήμα 5.13 παρουσιάζεται γράφημα shape coding Κάθε παρατήρηση έχει εννέα γνωρίσματα. Το διάγραμμα απεικονίζει 35 παρατηρήσεις (5X7).



Σχήμα 5.13 *Shape Coding*

5.4.4 Τεχνικές Εικονοστοιχείων

5.4.4.1 Επαναληπτικών Προτύπων (Recursive Pattern)

Οι τεχνικές Εικονοστοιχείων οπτικοποιούν τα δεδομένα, χρωματίζοντας κατάλληλα τα εικονοστοιχεία της οθόνης. Κάθε pixel αντιστοιχεί σε μια τιμή των δεδομένων, δηλαδή στην τιμή ενός γνωρίσματος μίας παρατήρησης. Το χρώμα του pixel καθορίζεται από το μέγεθος της τιμής. Η τεχνική του Επαναληπτικού Προτύπου προτάθηκε από τους Keim, Kriegel and Ankerst (1995). Σύμφωνα με αυτήν την τεχνική, τα εικονοστοιχεία οργανώνονται σε επαναλαμβανόμενα πρότυπα. Οι παράμετροι του προτύπου ορίζονται από τον χρήστη, έτσι ώστε να διατάξει τις τιμές σύμφωνα με μια λογική δομή, πχ να οργανώσει τις τιμές σε έτη, μήνες κλπ.
Κάθε γνώρισμα των δεδομένων οπτικοποιείται σε ξεχωριστό υποπλαίσιο. Εντός του υποπλαισίου σχεδιάζονται τα επαναλαμβανόμενα πρότυπα. Το βασικό πρότυπο έχει ύψος *h1* και πλάτος *w1*, τα οποία καθορίζονται από τον χρήστη. Τα εικονοστοιχεία του βασικού προτύπου διατάσσονται στην πρώτη γραμμή με σειρά από τα αριστερά προς τα δεξιά, στη δεύτερη γραμμή από τα δεξιά προς τα αριστερά, στην τρίτη γραμμή από τα

αριστερά και προς τα δεξιά, και η διαδικασία συνεχίζεται με αυτήν την ακολουθία εναλλαγής κατεύθυνσης. Τα βασικά πρότυπα διατάσσονται και αυτά με μια σειρά σε μια περιοχή ύψους *h2* και πλάτους *w2*. Η περιοχή αυτή μπορεί να θεωρηθεί άλλο πρότυπο, και τα πρότυπα αυτά θα διαταχθούν με τη δική τους σειρά. Το μέγεθος του συνολικού πλαισίου που απαρτίζεται από τα υποπλαίσια καθορίζεται από τον ερευνητή, περιορίζεται όμως από το μέγεθος της οθόνης. Η μέθοδος των Επαναληπτικών Προτύπων είναι ιδιαίτερα αποτελεσματική για την οπτικοποίηση χρονοσειρών. Στην εργασία του Ankerst (2001) υπάρχουν εντυπωσιακά παραδείγματα οπτικοποίησης δεδομένων με τη μέθοδο των επαναληπτικών προτύπων.

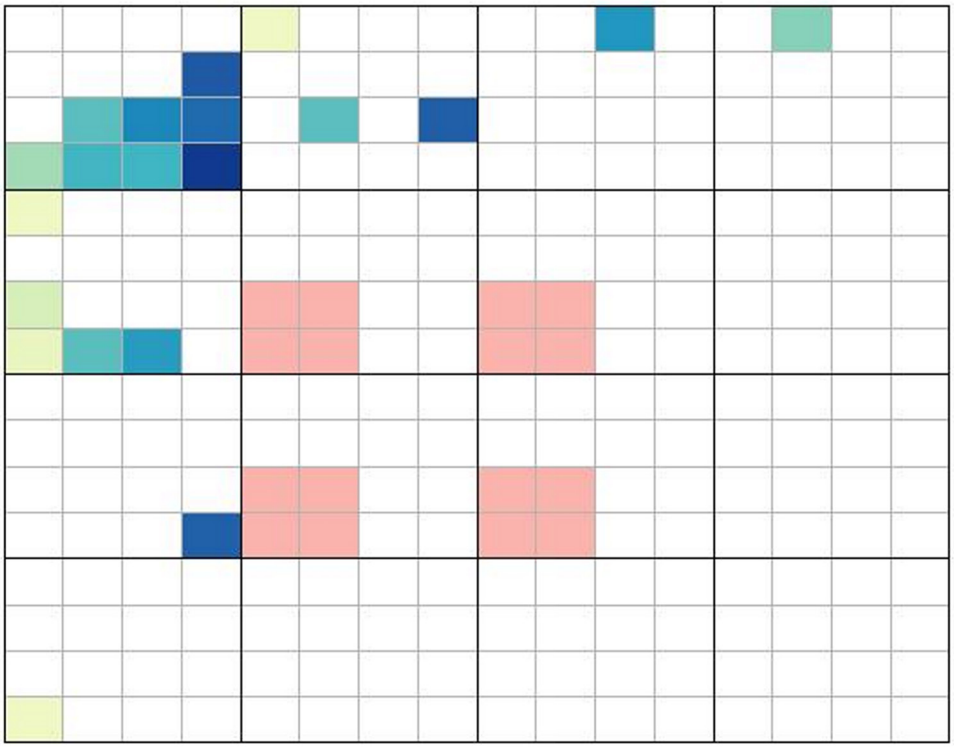
5.4.4.2 Κυκλικών Τομέων

Τεχνική των κυκλικών τομέων προτάθηκε από τους Ankerst, Keim and Kriegel, (1996) ως κατάλληλη για την οπτικοποίηση δεδομένων με πολλές διαστάσεις. Το σύνολο των δεδομένων απεικονίζεται ως ένας κύκλος, χωρισμένος σε ισομερή τμήματα, όπου κάθε τμήμα αντιστοιχεί σε μια διάσταση. Εντός των τμημάτων κωδικοποιούνται οι τιμές των παρατηρήσεων, χρησιμοποιώντας ένα εικονοστοιχείο για κάθε τιμή. Το χρώμα του εικονοστοιχείου εξαρτάται από το μέγεθος της τιμής. Οι τιμές διατάσσονται εντός των κυκλικών τμημάτων από το κέντρο προς την περιφέρεια πάνω σε γραμμές, που είναι κάθετες στις ακτίνες και ακολουθώντας εναλλασσόμενη κατεύθυνση. Η κατεύθυνση αλλάζει κάθε φορά που η γραμμή συναντά την ακτίνα, η οποία ορίζει τα όρια του κυκλικού τμήματος. Βασικό πλεονέκτημα της μεθόδου είναι ότι παρέχει ένα σημείο αναφοράς, το κέντρο του κύκλου, και ότι οι τιμές των παρατηρήσεων που βρίσκονται κοντά στο σημείο αναφοράς είναι συγκεντρωμένες στο κέντρο του κύκλου. Για παράδειγμα, αν γίνεται οπτικοποίηση ιστορικών δεδομένων, τότε τα παλαιότερα δεδομένα συγκεντρώνονται στο κέντρο του κύκλου και τα νεότερα στην περιφέρεια. Επίσης, η αναδιάταξη των κυκλικών τομέων και η κατάλληλη τοποθέτηση τους μπορεί να διευκολύνει τις συγκρίσεις. Παράδειγμα οπτικοποίησης δεδομένων με χρήση της τεχνικής των Κυκλικών Τομέων υπάρχει στο Ankerst (2001)

5.4.5 Τεχνικές Στοίβας

5.4.5.1 Dimensional Stacking

Η τεχνική αυτή προτάθηκε από τους LeBlanc, Ward and Wittels (1990), για τη δισδιάστατη απεικόνιση δεδομένων με N διαστάσεις και στηρίζεται στην ένθεση συστήματος συντεταγμένων μέσα σε άλλο σύστημα συντεταγμένων. Αρχικά επιλέγονται δύο διαστάσεις και δημιουργείται το πρώτο σύστημα συντεταγμένων. Οι μεταβλητές αυτές θεωρείται ότι κινούνται αργότερα. Οι εξωτερικές μεταβλητές χωρίζουν την επιφάνεια σε τμήματα ανάλογα με το πλήθος των τιμών τους. Οι τιμές των μεταβλητών είναι διακριτές. Για συνεχείς μεταβλητές μπορεί να εφαρμοστεί διακριτοποίηση. Σε κάθε ένα από τα τμήματα που ορίζουν οι εξωτερικές μεταβλητές, τοποθετείται σύστημα συντεταγμένων για τις αμέσως πιο γρήγορες μεταβλητές, οι οποίες με τη σειρά τους δημιουργούν νέα υποτμήματα βάσει των τιμών τους. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να ενσωματωθούν όλες οι διαστάσεις. Εάν το πλήθος των διαστάσεων είναι μονό, τότε στο τελευταίο επίπεδο δημιουργείται μια πρόσθετη διάσταση. Για κάθε διάσταση ορίζεται ένα επίπεδο και μια διεύθυνση (οριζόντια ή κάθετη), ώστε να αντιστοιχηθεί με άξονα συντεταγμένων. Το τελικό γράφημα εξαρτάται από την επιλογή επιπέδου για τις μεταβλητές. Ο καθορισμός του επιπέδου πρέπει να γίνει με βάση τη σημαντικότητα των μεταβλητών, με τις σημαντικότερες μεταβλητές να θεωρούνται «αργότερες» και να απαρτίζουν το εξωτερικό σύστημα συντεταγμένων, και τις λιγότερο σημαντικές μεταβλητές να θεωρούνται γρηγορότερες και να δημιουργούν ένθετα συστήματα συντεταγμένων. Στο Σχήμα 5.14 παρουσιάζεται διάγραμμα Dimensional Stacking που αναπαριστά οικονομικά στοιχεία επιχειρήσεων με τέσσερεις διαστάσεις, το κεφάλαιο κίνησης, τα κέρδη προς σύνολο ενεργητικού (εξωτερικές συντεταγμένες, αργές μεταβλητές), το σύνολο ενεργητικού και τον δείκτη πιστοληπτικής ικανότητας Quiskore (εσωτερικές συντεταγμένες, γρήγορες μεταβλητές). Το διάγραμμα κατασκευάστηκε με το λογισμικό XMDV Tool.

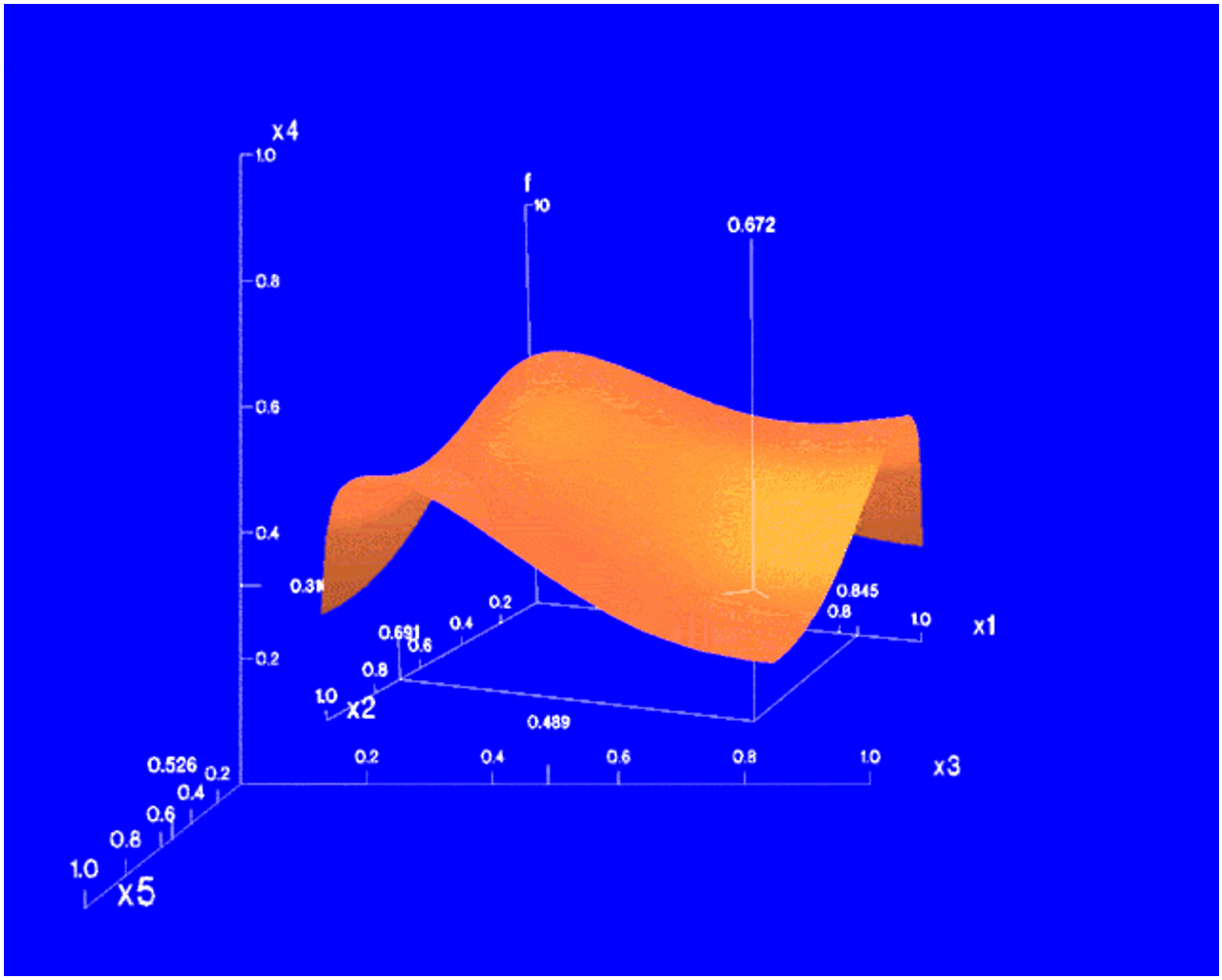


Σχήμα 5.14 *Dimensional Stacking.*

5.4.5.2 Worlds within Worlds

Η τεχνική Worlds within Worlds αναπτύχθηκε από τους Feiner and Beshers (1990) και στοχεύει στην αναπαράσταση δεδομένων Ν διαστάσεων σε χώρους 3 διαστάσεων. Αυτό επιτυγχάνεται διατηρώντας σταθερές τις τιμές από μία ή περισσότερες ανεξάρτητες μεταβλητές. Αν υποθέσουμε ότι τα δεδομένα είναι τετραδιάστατα, τότε επιλέγεται μια σταθερή τιμή για τη μια διάσταση. Αυτό είναι ισοδύναμο με την αποκοπή μιας πολύ λεπτής φέτας δεδομένων κάθετα στον άξονα της επιλεγμένης διάστασης. Η φέτα αντιστοιχεί στη σταθερή τιμή της μεταβλητής. Ο τρισδιάστατος χώρος που απομένει μπορεί να θεωρηθεί ως χώρος μιας συνάρτησης *f(x1,x2)* με δύο μεταβλητές, όπου οι δύο ανεξάρτητες μεταβλητές αντιστοιχούν στις δύο διαστάσεις και η τιμή της συνάρτησης αντιστοιχεί στην τρίτη διάσταση. Στον χώρο αυτό σχεδιάζεται η διακύμανση της συνάρτησης *f(x1,x2)*. Εάν οι διαστάσεις είναι περισσότερες από τέσσερεις, τότε επιλέγονται σταθερές τιμές και για τις επιπλέον μεταβλητές.

Οι διαστάσεις, οι οποίες απενεργοποιήθηκαν, μπορούν να επανέλθουν με τον παρακάτω τρόπο. Οι απενεργοποιημένες διαστάσεις συγκροτούν έναν εξωτερικό χώρο, που ενσωματώνει τον εσωτερικό χώρο των τριών διαστάσεων. Για παράδειγμα, αν οριστεί ένας εξωτερικός χώρος τριών διαστάσεων, μέσα στον οποίο τοποθετηθεί ένα εσωτερικός χώρος τριών άλλων διαστάσεων, επιτυγχάνεται απεικόνιση σε σχέση με χώρο έξι διαστάσεων. Κάθε διάγραμμα του εσωτερικού χώρου αναφέρεται σε ένα σημείο του εξωτερικού χώρου. Ο χρήστης επιλέγει ένα ή περισσότερα σημεία του εξωτερικού χώρου και δημιουργεί ένα ή αντίστοιχα περισσότερα διαγράμματα του εσωτερικού χώρου. Με τον τρόπο αυτό, ο χρήστης ελέγχει τη συμπεριφορά των δεδομένων για διάφορα σημεία του εξωτερικού χώρου. Στο Σχήμα 5.15 παρουσιάζεται ένα παράδειγμα. Ο εξωτερικός χώρος συντίθεται από τις διαστάσεις *X3*, *X4* και *X5*, ενώ ο εσωτερικός χώρος από τις διαστάσεις *X1*, *X2* και *F*. Το διάγραμμα δείχνει τη διακύμανση της μεταβλητής *F* ως προς τις μεταβλητές *X1* και *X2* για συγκεκριμένες τιμές των μεταβλητών *X3*, *X4* και *X5*. Στο ίδιο γράφημα θα μπορούσαν να προστεθούν επιπλέον διαγράμματα του εσωτερικού χώρου για επιπλέον τριάδες τιμών των μεταβλητών *X3*, *X4* και *X5*. Μειονεκτήματα είναι ότι η απεικόνιση πολλών διαστάσεων είναι περίπλοκη και ότι παρουσιάζονται πάντα φέτες δεδομένων και όχι το σύνολο τους. Επίσης, λεπτές τρισδιάστατες φέτες ενός πολυδιάστατου χώρου μπορεί να μην αποκαλύπτουν πρότυπα δεδομένων, για την εμφάνιση των οποίων απαιτούνται και οι επιπλέον διαστάσεις.



Σχήμα 5.15 *Worlds within Worlds (Αναπαραγωγή Από Slidewiki. Ιδιοκτήτης: sidraaslam)*

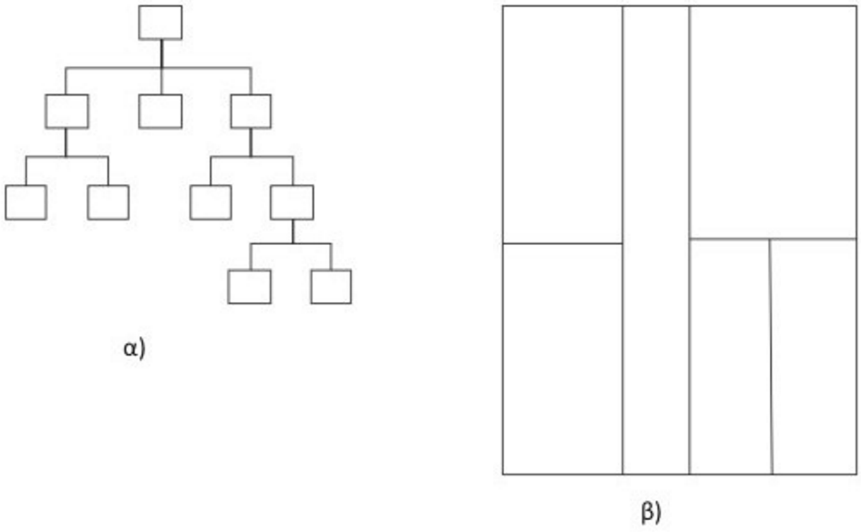
5.4.5.3 Δενδροχάρτες

Μεγάλο πλήθος των πληροφοριών του πραγματικού κόσμου έχουν ιεραρχική δομή. Επιχειρήσεις και οργανισμοί, οι διευθύνσεις του Διαδικτύου, οι κατάλογοι των βιβλιοθηκών είναι μερικά μόνον παραδείγματα. Για την οπτική αναπαράσταση δεδομένων με ιεραρχική δομή, οι Johnson and Shneiderman (1991) πρότειναν τη μέθοδο των Δενδροχαρτών (Treemaps). Ο παλαιότερος και τυπικός τρόπος αναπαράστασης δεδομένων αυτού του τύπου είναι με τη χρήση διαγραμμάτων δένδρου. Τα διαγράμματα δένδρου αποτελούνται από κόμβους συνδεδεμένους με τέτοιο τρόπο, ώστε να αποτυπώνουν την ιεραρχική δομή. Στο Σχήμα 5.16α παρουσιάζεται ένα διάγραμμα δένδρου. Τα διαγράμματα δένδρου παρουσιάζουν σημαντικά μειονεκτήματα. Σπαταλούν τον χώρο παρουσίασης και τα περισσότερα pixels της οθόνης χρησιμοποιούνται ως φόντο. Μεγάλα δένδρα δεν μπορούν να αναπαρασταθούν σε περιορισμένο χώρο. Επίσης, παρέχουν πληροφορία μόνο σχετικά με τη δομή και όχι σχετικά με το περιεχόμενο.

Οι Δενδροχάρτες αξιοποιούν ολόκληρο τον χώρο παρουσίασης, ενώ ταυτόχρονα παρέχεται πληροφορία σχετικά με το περιεχόμενο. Τα δεδομένα αναπαριστώνται με ένα ορθογώνιο παραλληλόγραμμο, το οποίο επιμερίζεται σε τόσα τμήματα όσοι είναι οι κόμβοι του πρώτου επιπέδου. Το κάθε τμήμα επιμερίζεται με τη σειρά του σε τόσα υποτμήματα, όσοι είναι οι κόμβοι που ανήκουν σ' αυτό. Η διαδικασία επαναλαμβάνεται μέχρι να εξαντληθούν όλα τα επίπεδα και οι κόμβοι. Για τον τρόπο επιμερισμού των παραλληλογράμμων, οι Johnson and Shneiderman (1991) πρότειναν τη μέθοδο slice and dice, σύμφωνα με την οποία το αρχικό παραλληλόγραμμο επιμερίζεται σε κάθετα τμήματα, τα τμήματα πρώτου επιπέδου επιμερίζονται σε οριζόντια υποτμήματα, τα τμήματα δεύτερου επιπέδου επιμερίζονται σε κάθετα υποτμήματα και η διαδικασία συνεχίζεται με την εναλλαγή της κατεύθυνσης επιμερισμού από κάθετη σε οριζόντια για κάθε μετάβαση σε χαμηλότερο επίπεδο. Το μέγεθος κάθε κόμβου εξαρτάται από το λεγόμενο «βάρος», το οποίο είναι μια ιδιότητα των δεδομένων. Αν για παράδειγμα, το δένδρο αναπαριστά τη δομή των φακέλων σε ένα σκληρό δίσκο, το βάρος

θα μπορούσε να είναι το συνολικό μέγεθος των αρχείων που περιέχονται στον κάθε φάκελο. Με τον τρόπο αυτό, οι δενδροχάρτες παρέχουν πληροφορία όχι μόνο για την ιεραρχική δομή, αλλά και για το περιεχόμενο. Άλλα οπτικά χαρακτηριστικά των παραλληλογράμμων, όπως το χρώμα, η υφή, έντονα πλαίσια κλπ. μπορούν να κωδικοποιούν επιπλέον πληροφορία για το περιεχόμενο.

Οι δενδροχάρτες είναι μια πολλά υποσχόμενη τεχνική για την αναπαράσταση δεδομένων με ιεραρχική δομή. Λόγω του ενδιαφέροντος που παρουσιάζουν, έχουν προταθεί κατά καιρούς διάφορες μετατροπές και βελτιώσεις. Τέτοιες παραλλαγές είναι οι τεχνικές Squarified Treemaps, Ordered Treemaps, Strip Treemap Algorithm, Cushion Treemaps, Cascaded Treemaps και Voronoi Treemaps. Παράδειγμα Δενδροχάρτη, το οποίο σχεδιάστηκε σύμφωνα με τον αρχικό αλγόριθμο των Johnson and Shneiderman και αντιστοιχεί στο διάγραμμα δένδρου του Σχήματος 5.16α παρουσιάζεται στο Σχήμα 5.16β.



Σχήμα 5.16 *Διάγραμμα Δένδρου και Δενδροχάρτης*

Ο κατάλογος των τεχνικών για την οπτικοποίηση των δεδομένων είναι πολύ μακρύς. Ενδεικτικά αναφέρουμε τις τεχνικές Landscape, Projection Views, Circular Parallel Coordinates, Radial Coordinate Visualization, Tile Bars, Cone Tree, Infocube και ο κατάλογος μπορεί να συνεχιστεί. Είναι προφανές ότι εξαντλητική κάλυψη όλων των τεχνικών δεν μπορεί να πραγματοποιηθεί στα πλαίσια ενός κεφαλαίου. Ο ενδιαφερόμενος αναγνώστης μπορεί να αναζητήσει περισσότερες πληροφορίες σε κάποιο από τα πολλά σχετικά βιβλία, όπως το Spence (2015).

5.5 Μελέτη περίπτωσης. Αναγνώριση απάτης με γραφικά μέσα

Έχει αναφερθεί ήδη ότι οι τεχνικές οπτικοποίησης των δεδομένων έχουν εφαρμοστεί για την αναγνώριση δομών και ιδιοτήτων των δεδομένων και την αποκόμιση χρήσιμης πληροφορίας. Η χρήση γραφικών μέσων για την ανάλυση επιχειρηματικών δεδομένων έχει ευρύτατο πεδίο εφαρμογής. Στο σημείο αυτό θα παρουσιάσουμε μια περίπτωση του εντοπισμού της επιχειρηματικής απάτης με χρήση τεχνικών οπτικοποίησης, ως ένα παράδειγμα εφαρμογής αυτών των μεθόδων για επιχειρηματικούς σκοπούς.

Το πρόβλημα της απάτης σε οργανισμούς δεν είναι καθόλου ευκαταφρόνητο. Σύμφωνα με την έκθεση της ACFE (Association of Certified Fraud Examiners) (2014), οι οργανισμοί παγκοσμίως υποφέρουν απώλειες της τάξης του 5% επί των εσόδων τους εξαιτίας της απάτης. Με βάση το Παγκόσμιο Ακαθάριστο Προϊόν, το ποσό αυτό αντιστοιχεί σε 3,7 τρισεκατομμύρια δολάρια, οπότε είναι προφανές ότι η αντιμετώπιση του φαινομένου είναι πολύ υψηλής σημασίας. Η απάτη μπορεί να πάρει πολλές μορφές. Μπορεί να διαπραχθεί από άτομα που ανήκουν στον οργανισμό ή από άτομα που δεν σχετίζονται με τον οργανισμό. Στην πρώτη περίπτωση χαρακτηρίζεται «εσωτερική απάτη» (internal fraud), ενώ στη δεύτερη περίπτωση χαρακτηρίζεται «εξωτερική απάτη» (external fraud). Διάφοροι τύποι πρακτικών εξαπάτησης, τους οποίους αντιμετωπίζουν οι επιχειρήσεις, είναι η απάτη από εργαζομένους, η απάτη από καταναλωτές, η απάτη από προμηθευτές, τα εγκλήματα με χρήση υπολογιστών, η ιατρική και ασφαλιστική απάτη και τέλος η απάτη με παραποίηση των χρηματοοικονομικών καταστάσεων. Η έκθεση της ACFE (2014), η οποία επικεντρώνει στην εξαπάτηση από εργαζομένους, επισημαίνει ότι το μέσο ύψος ζημίας ανά επιχείρηση είναι 130.000 δολάρια, ενώ ο μέσος χρόνος εντοπισμού

της απάτης είναι 18 μήνες. Επίσης, τονίζει ότι οι οργανισμοί που εφαρμόζουν μηχανισμούς ελέγχου απάτης μειώνουν και το ύψος της ζημίας και τον χρόνο εντοπισμού.

Για τον εντοπισμό των περιστατικών απάτης έχουν εφαρμοστεί διάφορες μέθοδοι, που περιλαμβάνουν στατιστικές αναλύσεις, τεχνικές εξόρυξης δεδομένων αλλά και τεχνικές οπτικοποίησης. Ορισμένες περιπτώσεις εφαρμογής τεχνικών οπτικοποίησης για τον εντοπισμό της απάτης είναι οι ακόλουθες. Το σύστημα ADS (Advance Detection System) της National Association of Securities Dealers παρακολουθεί τις προσφορές και συναλλαγές των μετοχών του δείκτη NASDAG του Χρηματιστήριου της Νέας Υόρκης και επιχειρεί να εντοπίσει πρότυπα συναλλαγών που χρίζουν διερεύνησης. Σημειωτέον ότι υπάρχουν περίπου 2.000.000 προσφορές ή συναλλαγές ημερησίως. Η λογική του συστήματος εδράζεται στην ταύτιση πρότυπων κανόνων και πρότυπων χρονοσειρών. Εφαρμόζονται μέθοδοι εξόρυξης δεδομένων, όπως οι Κανόνες Συσχέτισης και τα Δένδρα Αποφάσεων. Επίσης, γίνεται χρήση τεχνικών οπτικοποίησης. Οπτικοποιούνται οι τιμές προσφοράς και ζήτησης έναντι του χρόνου, η ροή των συναλλαγών με βάση τον χρόνο εκτέλεσης τους, οι σχέσεις των συναλλασσομένων, το ύψος των τιμών με χρήση διαγραμμάτων τοπίου και οι προσφορές και πρακτικές συναλλαγής σε ορισμένη χρονική περίοδο.

Το Financial Crimes Enforcement Network είναι ένας φορέας του Υπουργείου Οικονομικών των ΗΠΑ, που αποστολή έχει την καταπολέμηση πρακτικών νομιμοποίησης παράνομου χρήματος. Για την εξυπηρέτηση του σκοπού αυτού, ανέπτυξαν το πληροφοριακό σύστημα FAIS, το οποίο παρακολουθεί και συνδέει μεγάλες συναλλαγές που διεξάγονται με μετρητά, ώστε σε περίπτωση που θεωρηθούν ύποπτες να ακολουθήσει περαιτέρω έρευνα ή και δίωξη. Το σύστημα εφαρμόζει μεθόδους Τεχνητής Νοημοσύνης, όπως rule-base reasoning, αλλά και τεχνικές οπτικοποίησης. Ένα σημαντικό χαρακτηριστικό του είναι ότι εστιάζει σε υποκείμενα, όπως πρόσωπα ή οργανισμούς. Τεχνικές οπτικοποίησης εφαρμόζονται για την ανάλυση διασυνδέσεων μεταξύ των υποκειμένων. Τα δεδομένα μπορούν να οπτικοποιηθούν με γραφήματα link-and-edge ή με γραφήματα wagon-wheel. Τα γραφήματα είναι διαδραστικά και επιτρέπουν την αύξηση της λεπτομέρειας με ανάλυση μικρότερων συνόλων δεδομένων καθώς και τη συγχώνευση υποκειμένων. Στην εργασία των Senator et al. (1995) παρουσιάζεται το σύστημα FAIS και περιέχονται σχήματα οπτικοποίησης δεδομένων για αναγνώριση απάτης.

Μια ενδιαφέρουσα ελληνική εργασία σχετικά με εντοπισμό απάτης με χρήση τεχνικών οπτικοποίησης, είναι η Argyriou, Sotiraki and Symvonis (2013). Η ερευνητική ομάδα ανέπτυξε σύστημα για την ανίχνευση περιπτώσεων εργαζομένων ενός οργανισμού, οι οποίοι συνεργάζονται με πελάτες με σκοπό τη διάπραξη απάτης. Τα δεδομένα προέρχονται από αρχεία καταγραφής συμβάντων και συνίστανται σε μια τετράδα, που περιλαμβάνει μια χρονική στιγμή, έναν εργαζόμενο, έναν πελάτη και μια ενέργεια που πραγματοποίησε ο εργαζόμενος. Καθορίζονται μια σειρά από παράγοντες κινδύνου, όπως το πλήθος των συμβάντων που χρονικά βρίσκονται κοντά στην ημερομηνία πληρωμής του λογαριασμού του πελάτη, η περιοδικότητα των συμβάντων, συμβάντα που λαμβάνουν χώρα εκτός ωρών εργασίας κλπ. Για κάθε παράγοντα ορίζονται τρία επίπεδα κινδύνου, χαμηλό, μέσο και υψηλό. Ο ελεγκτής ιεραρχεί τη σημαντικότητα των παραγόντων. Η βασική μέθοδος απεικόνισης είναι ένα διάγραμμα τύπου σπιράλ, όπου κάθε περιέλιξη του σπιράλ αντιστοιχεί σε χρονικό διάστημα ενός μηνός. Στο σπιράλ τοποθετούνται κόμβοι, οι οποίοι συμβολίζουν ένα γεγονός που σχετίζεται με έναν πελάτη. Κόμβοι διαφορετικού χρώματος αντιστοιχούν σε διαφορετικούς πελάτες, ενώ κόμβοι διαφορετικού σχήματος αντιστοιχούν σε διαφορετικά πληροφοριακά συστήματα, από τα οποία προέρχεται η πληροφορία. Γεγονότα, τα οποία αφορούν τον ίδιο πελάτη και βρίσκονται πάνω στην ίδια ακτίνα ή σε κοντινές ακτίνες, υποδηλώνουν περιοδικότητα δραστηριότητας και μπορεί να σηματοδοτούν πρακτικές εξαπάτησης. Το σύστημα παράγει βίντεο, όπου κάθε πλαίσιο αντιστοιχεί σε έναν πελάτη, και οι πελάτες υψηλού κινδύνου τοποθετούνται στην αρχή του βίντεο.

5.6 Ταμπλό

Στα Συστήματα Επιχειρηματικής Ευφυΐας (ΣΕΕ) η διεπαφή του χρήστη έχει βαρύνουσα σημασία, ιδιαίτερα μάλιστα ο τρόπος παρουσίασης των πληροφοριών. Οι χρήστες των συστημάτων αυτών, κατά κανόνα υψηλόβαθμα στελέχη επιχειρήσεων, συνήθως δεν έχουν υψηλού επιπέδου γνώσεις πληροφορικής, είναι όμως άριστοι γνώστες των επιχειρηματικών ζητημάτων και απαιτούν η πληροφόρηση που δέχονται από το σύστημα να είναι όχι μόνο ακριβής και επίκαιρη, αλλά και προσαρμοσμένη στον δικό τους τρόπο λειτουργίας και σκέψης. Επίσης, ο τρόπος παρουσίασης των πληροφοριών πρέπει να είναι καλαίσθητος και ευχάριστος.

Ο συνηθέστερος ίσως τρόπος παρουσίασης των πληροφοριών στα ΣΕΕ είναι τα λεγόμενα «ταμπλό» ή «dashboards». Το ταμπλό είναι ένα παράθυρο διεπαφής, το οποίο παρουσιάζει με γραφικό τρόπο διάφορους επιλεγμένους δείκτες της επιχείρησης. Συνήθως οι δείκτες αυτοί είναι κάποιοι από τους Κύριους Δείκτες Επίδοσης (ΚΔΕ) (Key Performance Indicators (KPI)). Οι ΚΔΕ είναι προσεκτικά επιλεγμένοι δείκτες, που

αναφέρονται σε παράγοντες ζωτικής σημασίας για τον οργανισμό. Περισσότερες λεπτομέρειες για τους ΚΔΕ υπάρχουν στο πρώτο Κεφάλαιο του βιβλίου. Κατά κανόνα τα ταμπλό δεν περιέχουν πολλές πληροφορίες. Επιλέγονται οι πιο σημαντικές πληροφορίες, οι οποίες αφορούν ένα συγκεκριμένο επιχειρηματικό ζήτημα. Για παράδειγμα, ένα ταμπλό που θα αναφέρεται στην κερδοφορία της επιχείρησης μπορεί να περιλαμβάνει δείκτες όπως το σύνολο πωλήσεων και εξόδων, τα κέρδη προ φόρων και τόκων, κέρδη προς σύνολο ενεργητικού κλπ. Τα θεματικά προσανατολισμένα ταμπλό εξυπηρετούν τον μάνατζερ, καθώς παρουσιάζουν τις κυριότερες πληροφορίες σχετικά με αυτό το θέμα και διευκολύνουν τη λήψη αποφάσεων.
Τα δεδομένα του ταμπλό προέρχονται συνήθως από κάποια Αποθήκη Δεδομένων. Σπανιότερη πηγή δεδομένων είναι κάποια σχεσιακή βάση δεδομένων. Τα δεδομένα μπορούν να αναφέρονται στην τρέχουσα κατάσταση πραγμάτων του οργανισμού ή/και σε προηγούμενα, ιστορικά στοιχεία. Σε κάθε περίπτωση πάντως, αποτυπώνουν ένα στιγμιότυπο της επιχείρησης μια συγκεκριμένη χρονική στιγμή.
Τα ταμπλό έχουν καθιερωθεί ως ένας πολύ δημοφιλής τρόπος παρουσίασης πληροφοριών στα συστήματα Επιχειρηματικής Ευφυΐας, γιατί διαθέτουν μια σειρά από πλεονεκτήματα, ορισμένα εκ των οποίων είναι τα ακόλουθα:

- Αναπαριστούν με οπτικό και συνοπτικό τρόπο σημαντικούς δείκτες επίδοσης.
- Συγκεντρώνουν σε περιορισμένης έκτασης παράθυρα πολλές κρίσιμες πληροφορίες, οι οποίες αναφέρονται σε στοχευμένα επιχειρηματικά ζητήματα.
- Επιτρέπουν τον εντοπισμό αρνητικών συμβάντων και τάσεων με «μια ματιά». Ένας κόκκινος σηματοδότης δίπλα σε έναν δείκτη μπορεί να σημαίνει την παραβίαση των επιθυμητών ορίων του δείκτη.
- Μπορούν να χρησιμοποιηθούν ως μέσο διασποράς πληροφόρησης και επικοινωνίας μεταξύ των στελεχών της επιχείρησης.
- Εξοικονομούν χρόνο, σε σχέση με τις πολλαπλές εκθέσεις που παρουσιάζουν ένα ζήτημα.
- Σε αντίθεση με τις στατικές εκθέσεις, είναι πιο διαδραστικά και επιτρέπουν στον χρήστη την προβολή διαφορετικών δεδομένων και την πλοήγηση σε πληροφορίες.
- Καθιστούν εφικτή την άμεση οπτική αντιπαραβολή επιδόσεων και προκαθορισμένων στόχων.
- Είναι ευχάριστα.
- Προκαλούν το ενδιαφέρον των χρηστών.
- Βοηθούν στη λήψη καλύτερων επιχειρηματικών αποφάσεων.

Τα σύγχρονα λογισμικά Ε.Ε. επιτρέπουν στον χρήστη να σχεδιάσει τα δικά του ταμπλό. Μία εργαλειοθήκη περιέχει διάφορους τρόπους γραφικής απεικόνισης δεδομένων, όπως ραβδογράμματα, πίτες, συγκεντρωτικούς πίνακες κλπ. Πέρα από τα συνηθισμένα γραφήματα, σε ένα ταμπλό επιχειρηματικής ευφυΐας μπορεί κανείς να συναντήσει και άλλους πιο ευφάνταστους τρόπους απεικόνισης, όπως ταχύμετρα ή άλλα όργανα αυτοκινήτου και σηματοδότες κυκλοφορίας. Ένα ταμπλό που μοιάζει με πίνακα ελέγχου ενός οχήματος δημιουργεί στον χρήστη την αίσθηση ότι «οδηγεί» την επιχείρηση. Από αυτήν τη μεγάλη γκάμα γραφημάτων, ο χρήστης επιλέγει το κατάλληλο γράφημα για τα δεδομένα του και το τοποθετεί στο ταμπλό, συνήθως με την τεχνική drag and drop. Στη συνέχεια, συνδέει το γράφημα με τα δεδομένα. Μπορεί επίσης να ορίσει το μέγεθος και να διαλέξει τα χρώματα. Με τον τρόπο αυτό, ο χρήστης κατασκευάζει ένα ταμπλό που να ανταποκρίνεται στις δικές του εξατομικευμένες ανάγκες, στις απαιτήσεις της εργασίας του αλλά και τις αισθητικές του προτιμήσεις.

Τα σύγχρονα ταμπλό έχουν ισχυρά στοιχεία διαδραστικότητας. Καταρχήν δίνεται η δυνατότητα στον χρήστη να σχεδιάσει τα δικά του ταμπλό. Πέραν τούτου όμως, ένα ταμπλό μπορεί να περιέχει ενεργά στοιχεία, τα οποία επιτρέπουν στον χρήστη να επιλέξει στήλες δεδομένων, να εφαρμόσει φίλτρα στα δεδομένα, να εκτελέσει ερωτήματα βάσεων δεδομένων, να εκτελέσει πράξεις OLAP, όπως πχ drill down μέχρι ακόμα και το επίπεδο των συναλλαγών, να τροποποιήσει στοιχεία, να πλοηγηθεί σε πληροφορίες και να εκτελέσει αναλύσεις. Άλλες λειτουργικότητες των ταμπλό δίνουν τη δυνατότητα στον χρήστη να συνεργαστεί με άλλα στελέχη για την ανάλυση των στοιχείων ή να στείλει μηνύματα εγρήγορσης (alerts) στα κατάλληλα πρόσωπα μέσω του Διαδικτύου ή κινητών συσκευών. Ορισμένα λογισμικά δίνουν τη δυνατότητα εξαγωγής του ταμπλό ή κάποιων στοιχείων του ταμπλό, σε άλλες πλατφόρμες, όπως πχ φύλλα δεδομένων (spreadsheets) ή προγράμματα παρουσιάσεων.

Βιβλιογραφία / Αναφορές

Andrews, D. F. (1972). Plots of High Dimensional Data. *Biometrics*, *28*(1), 125-136. doi: 10.2307/2528964

Ankerst, M. (2001). *Visual Data Mining with Pixel-oriented Visualization Techniques.* Retrieved from CiteSeerX website: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.2511&rep=rep1&type=pdf

Ankerst, M., Keim D. A., & Kriegel, H. P. (1996). Circle Segments: A technique for visually exploring large Multidimensional data Sets. *Proceedings Visualization '96, Hot topic session*. doi: 10.1.1.68.1811

Argyriou, E., Sotiraki, A., & Symvonis, A. (2013). Occupational Fraud Detection through Visualization. *Proceedings of the 2013 IEEE International Conference on Intelligence and Security Informatics,* 4-6. Seattle: IEEE. doi: 10.1109/ISI.2013.6578773

Association of Certified Fraud Examiners. (2014). *Report to the Nations on Occupational Fraud and Abuse.*

Beddow, J. (1990). Shape Coding of Multidimensional Data on a Microcomputer Display. *Proceedings of the First IEEE Conference on Visualization,* 238-246. San Francisco: IEEE. doi: 10.1109/VISUAL.1990.146387

Bertin, J. (1983). *Semiology of Graphics*. Madison, WI: University of Wisconsin Press.

Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization.* San Francisco, CA: Morgan Kaufmann.

Chambers, J. M., Clevelant, W. S., Kleiner B., & Tuckey, P. (1983). *Graphical Methods for Data Analysis*. Wadsworth Publishing Co Inc.

Chernoff, H. (1973). The Use of Faces to Represent Points in k-dimensional Space Graphically. *Journal of American Statistical Association*, *68*(342*)*, 361-368. doi: 10.2307/2284077

Feiner, S., & Beshers, C. (1990). Visualizing n-Dimensional Virtual Worlds with n-Vision. *ACM SIGGRAPH Computer Graphics*, *24*(2), 37-38. doi: 10.1145/91394.91412

Friendly, M. (2005). Milestones in the History of Data Visualization: A Case Study of Statistical Historiography. In C. Weihs & W. Gaul (Eds.), *Classification – The Ubiquitous Challenge Studies in Classification, Data Analysis and Knowledge Organization* (pp. 34-52). Berlin – Heidelberg, Germany: Springer-Verlag. doi: 10.1007/3-540-28084-7_4

Hearst, M. (1995). TileBars: Visualization of Term Distribution Information in Full Text Information Access. *Proceedings of the 95' SIGCHI Conference on Human Factors in Computing Systems,* 59-66. Denver: ACM. doi: 10.1145/223904.223912

Hoffman, P., Grinstein, G., Marx, K., Grosse, I., & Stanley, E. (1997). DNA Visual and Analytic Data Mining. *Proceedings of the 1997 IEEE International Conference on Visualization,* 437-441. Phoenix: IEEE. doi: 10.1109/VISUAL.1997.663916

Inselberg, A. (1985). The Plane with Parallel Coordinates. *The Visual Computer*, *1*(2), 69-91. doi: 10.1007/BF01898350

Inselberg, A., & Dimsdale, B. (1990). Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry. *Proceedings of the First IEEE Conference on Visualization,* 361-378. San Francisco: IEEE. doi: 10.1109/VISUAL.1990.146402

Johnson, B., & Shneiderman, B. (1991). Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. *Proceedings of the 2nd IEEE Conference on Visualization*, 275-282. Los Alamos: IEEE. doi: 10.1109/VISUAL.1991.175815

Keim, D. A. (2002). Information Visualization and Visual Data Mining. *IEEE Transaction on Visualization and Computer Graphics*, 7(1), 100-107. doi: 10.1109/2945.981847

Keim, D. A., & Kriegel, H. P. (1994). VisDB: Database Exploration using Multidimensional Visualization. *Computer Graphics and Applications*, 14(5), 40-49. doi:10.1109/38.310723

Keim D. A., & Kriegel, H. P. (1996). Visualization Techniques for Mining Large Databases: A Comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 923-928. doi: 10.1109/69.553159

Keim, D. A., Kriegel, H. P., & Ankerst, M. (1995). Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data. *Proceedings of the '95 IEEE Conference on Visualization*, 279-286. Washington: IEEE. doi: 10.1109/VISUAL.1995.485140

Keim, D., Hao, M. C., Ladish, J., Hsu, M., & Dayal, U. (2001). Pixel Bar Charts: A new Technique for Visualizing Large Multi-Attribute Data Sets without Aggregation. *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, 113-120. San Diego, CA: IEEE. doi: 10.1109/

INFVIS.2001.963288
Kleiner, B., & Hartigan, J. A. (1981). Representing Point in Many Dimensions by Trees and Castles. *Journal of American Statistical Association, 76*(374), 260-269. doi: 10.2307/2287820
LeBlanc, J., Ward, M. O., & Wittels, N. (1990). Exploring N-Dimensional Databases. *Proceedings of the First IEEE Conference on Visualization*, 230-237. San Francisco: IEEE. doi: 10.1109/VISUAL.1990.146386
Levkowitz, H. (1991). Color Icons: Merging Color and Texture Perception for Integrated Visualization of Multiple Parameters. *Proceedings of the 2nd IEEE Conference on Visualization*, 164-170. Los Alamos: IEEE. doi: 10.1109/VISUAL.1991.175795
Mazza, R. (2009). *Introduction to Information Visualization.* London, UK: Springer Publishing Co.
Miller, J. R. (2007). Attribute Blocks: A Tool for Visualizing Multiple Continuously-Defined Attributes. *IEEE Computer Graphics and Applications*, 27(3), 57-69. doi: 10.1109/MCG.2007.54
Pickett, R. M., & Grinstein, G. G. (1988). Iconographic Displays for Visualizing Multidimensional Data. *Proceedings of the 1988 IEEE International Conference on Systems, Man and Cybernetics*, 514-519. doi: 10.1109/ICSMC.1988.754351
Playfair, W. (1786). *The Commercial and Political Atlas*.
Rekimoto, J., & Green, M. (1993). The Information Cube: Using Transparency in 3d Information Visualization. *Proceedings of the 3rd Annual Workshop on Information Technologies & Systems (WITS '93)*, 125-132.
Rober, N. (2000). *Multidimensional Analysis and Visualization Software for Dynamic SPECT* (M.Sc.). Otto-von-Guericke-Universitaet, Magdeburg.
Robertson, G. G., Mackinlay, J. D., & Card, S. K. (1991). Cone Trees: Animated 3D Visualizations of Hierarchical Information. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 189-194. New Orleans, LA: ACM. doi: 10.1145/108844.108883
Senator, T. E., Goldberg, H. G., Wotton, J., Cottini, M. A., Umar Khan, W. F., Klinger, C. D., Llamas, W. M., Marrone, M. P., & Wong, R. W. H. (1995). The Financial Crimes Enforcement Network AI System (FAIS). Identifying Potential Money Laundering from Reports of Large Cash Transactions. *AI Magazine*, 16(4), 21-39. doi: 10.1609/aimag.v16i4.1169
Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the '96 IEEE Symposium on Visual Languages*, 336-343. Boulder, CO: IEEE. doi: 10.1109/VL.1996.545307
Shneiderman, B. (2002). Inventing Discovery Tools: Combining Information Visualization with Data Mining. J*ournal of Information Visualization*, 1(1), 5-12. doi: 10.1057/palgrave/ivs/9500006
Spence, R. (2001). *Information Visualization*. Harlow, UK: Addison-Wesley.
Spence, R. (2015). *Information Visualization*: An Introduction. New York, NY: Springer.
Teoh, S. T., & Ma, K. L. (2002). RINGS: A Technique for Visualizing Large Hierarchies. In GD *'02 Revised Papers of the 10th International Symposium on Graph Drawing* (pp. 268-275). London, UK: Springer-Verlag.
Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley Publishing Co.
Van Wijk, J. J., & Van Liere, R. D. (1994). Visualization of Multi-dimensional Functions Using HyperSlice. *CWI Quarterly*, 7(2), 147-158.

6 Εξόρυξη Γνώσης από Δεδομένα

Σύνοψη

Τα τελευταία χρόνια παρατηρείται μια ταχύτατη αύξηση του όγκου των αποθηκευμένων δεδομένων. Η τάση αυτή οφείλεται στην αύξηση των δυνατοτήτων του λογισμικού και του υλικού, στη μείωση του κόστους, στη διείσδυση της πληροφορικής σε κάθε δραστηριότητα της σύγχρονης κοινωνίας, στην ευρύτατη εξάπλωση του Διαδικτύου και σε μεγάλο βαθμό στο Web 2.0, το οποίο καθιστά τους χρήστες πρωταγωνιστικούς παράγοντες παραγωγής πληροφοριών. Η ανάλυση όλων αυτών των δεδομένων και η ανάκτηση χρήσιμης πληροφορίας χωρίς τη χρήση εξειδικευμένων τεχνικών είναι αδύνατη. Η Εξόρυξη Δεδομένων (ΕΔ), αντλώντας μεθοδολογίες από τη Μηχανική Μάθηση, τις Βάσεις Δεδομένων, τη Στατιστική και άλλους κλάδους, έχει στόχο την ανακάλυψη γνώσης μέσα από μεγάλους όγκους δεδομένων. Αν και σε επίπεδο ορολογίας η Εξόρυξη Δεδομένων χρησιμοποιείται αρκετές φορές ως συνώνυμο της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων, επί της ουσίας αναγνωρίζεται ότι η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων είναι μια ευρύτερη διαδικασία και ότι η καθεαυτό Εξόρυξη Δεδομένων είναι ένα στάδιο της, κατά το οποίο ανακαλύπτονται πρότυπα δεδομένων. Το σύνολο των σταδίων της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων είναι α) η συλλογή, ολοκλήρωση και ο καθαρισμός των δεδομένων, β) η επιλογή των δεδομένων και ο μετασχηματισμός τους, γ) η εξόρυξη των δεδομένων και δ)η αξιολόγηση των προτύπων και η ανακάλυψη της γνώσης. Υπάρχουν διάφορες εργασίες Εξόρυξης Δεδομένων και χωρίζονται σε εργασίες επιβλεπόμενης μάθησης και εργασίες μη επιβλεπόμενης μάθησης. Οι κυριότερες εργασίες Εξόρυξης Δεδομένων είναι οι ακόλουθες: Η Κατηγοριοποίηση στοχεύει στην εκτίμηση των τιμών ενός γνωρίσματος-στόχου με ονομαστικές τιμές, το οποίο ορίζει την κατηγορία των αντικειμένων. Η Παλινδρόμηση μοιάζει με την κατηγοριοποίηση, αλλά το γνώρισμα-στόχος έχει αριθμητικές τιμές. Η Ανάλυση Συστάδων επιμερίζει ένα σύνολο αντικειμένων σε ομάδες, βάσει ομοιότητας και χωρίς την ύπαρξη προκαθορισμένων κατηγοριών. Η Ανάλυση Κανόνων Συσχέτισης ανακαλύπτει σχέσεις μεταξύ τιμών των γνωρισμάτων, οι οποίες εμφανίζονται συχνά μαζί. Η Ανάλυση Εξαιρέσεων εντοπίζει και αναλύει περιπτώσεις, οι οποίες αποκλίνουν από το κανονικό ή συνηθισμένο. Η Ανάλυση Χρονοσειρών αναλύει μεγέθη τα οποία παρουσιάζουν χρονική εξέλιξη.
Για τη σύγχρονη επιχείρηση, η γνώση αποτελεί πολύτιμο κεφάλαιο και η Εξόρυξη Δεδομένων είναι το εργαλείο για την ανάκτηση της. Ένας σύγχρονος τρόπος θεώρησης της διάρθρωσης της επιχείρησης είναι μέσω των επιχειρηματικών διαδικασιών. Η Εξόρυξη Δεδομένων έχει βρει πεδίο εφαρμογής σε πλήθος επιχειρηματικών διαδικασιών. Στη διαφήμιση και στις πωλήσεις χρησιμοποιείται για την αναγνώριση της καταναλωτικής συμπεριφοράς των πελατών και την τμηματοποίηση της αγοράς. Καθίσταται έτσι εφικτός ο σχεδιασμός προϊόντων για κατηγορίες πελατών, η στοχευμένη διαφήμιση, και υποβοηθούνται οι διασταυρούμενες πωλήσεις. Επίσης, η γνώση του προφίλ του καταναλωτή διευκολύνει το εισερχόμενο μάρκετινγκ και το καθοδηγούμενο από γεγονότα μάρκετινγκ. Στις τράπεζες, η Εξόρυξη Δεδομένων χρησιμοποιείται για πωλήσεις και διαφήμιση, για τη διαχείριση του ρίσκου, για την αντιμετώπιση της απάτης και την αντιμετώπιση του ξεπλύματος χρήματος. Μεγάλες εφαρμογές έχει η Εξόρυξη Δεδομένων στους οργανισμούς Τηλεπικοινωνιών. Σε αυτούς τους οργανισμούς η ΕΔ εφαρμόζεται για την προώθηση των πωλήσεων και τη διαφήμιση, την αντιμετώπιση της απάτης και την αντιμετώπιση τεχνικών προβλημάτων του δικτύου. Στη σύγχρονη Ελεγκτική, η οποία τα τελευταία χρόνια αντιμετωπίζει μεγάλες προκλήσεις, μέθοδοι προερχόμενες από τη Μηχανική Μάθηση μπορούν να χρησιμοποιηθούν σε εργασίες, όπως η πρόβλεψη χρεοκοπίας και ο εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων. Η Εξόρυξη Κειμένου και η Ανάλυση Συναισθήματος είναι χρήσιμες τεχνικές για να γνωρίσει η επιχείρηση τις απόψεις του καταναλωτικού κοινού σχετικά με τα προϊόντα της και τις υπηρεσίες της. Με την Ανάλυση Κοινωνικών Δικτύων, μέσα από ιστοθέσεις κοινωνικής δικτύωσης και blogs, επιτυγχάνεται εντοπισμός ομάδων χρηστών, ενίσχυση του προφορικού μάρκετινγκ με επικέντρωση της διαφήμισης στους δυναμικούς κόμβους, και εύρεση ειδικών για συνεργασία ή πρόσληψη. Τέλος, η ΕΔ μπορεί να χρησιμοποιηθεί για την ανακάλυψη και μοντελοποίηση των επιχειρηματικών διαδικασιών μιας επιχείρησης.

Προαπαιτούμενη Γνώση.

Το παρόν κεφάλαιο εισάγει τον αναγνώστη σε βασικές έννοιες της Εξόρυξης Δεδομένων και παρουσιάζει τα κύρια πεδία εφαρμογής της στη σύγχρονη επιχείρηση. Υπό την έννοια αυτή, δεν υπάρχει κάποια ιδιαίτερη απαίτηση για προηγούμενες γνώσεις.

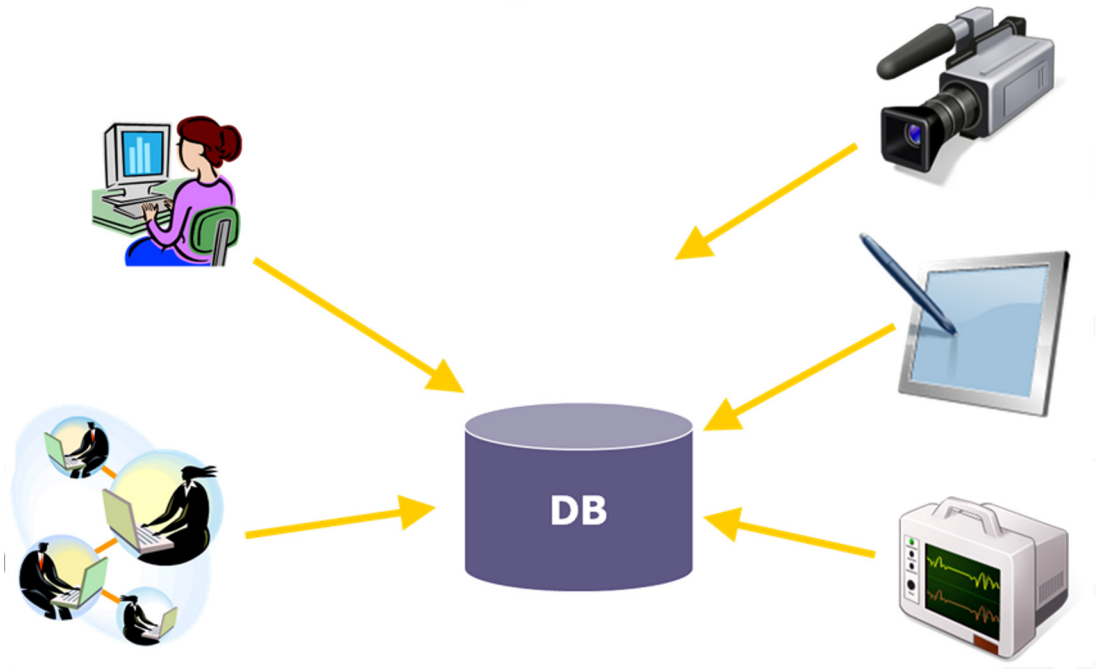
6.1 Εισαγωγή

Έχουμε κατακλυστεί από δεδομένα, όμως μας λείπει η πληροφορία. Η διαπίστωση αυτή θα μπορούσε να αποτελέσει την εναρκτήρια πρόταση ενός βιβλίου που αναφέρεται στην Εξόρυξη Δεδομένων, γιατί αποτυπώνει με ακρίβεια τη γενεσιουργό αιτία της ανάπτυξης αυτού του νέου επιστημονικού πεδίου. Αφετηριακό σημείο είναι η παραγωγή και αποθήκευση δεδομένων με καταιγιστικό ρυθμό. Η διαπίστωση αυτή δεν είναι πρόσφατη. Την αναγνωρίζει κάθε εποχή, τοποθετώντας την όμως στα δικά της μέτρα. Μόλις το 1996, οι Fayyad, Piatetsky-Shapiro and Smyth (1996) ισχυρίζονται ότι τα δεδομένα συλλέγονται και συσσωρεύονται με δραματικό ρυθμό. Για τη σύγχρονη κοινωνία μπορούμε να πούμε ότι η εφαρμογή της πληροφορικής είναι γενικευμένη. Δύσκολα θα μπορούσε να φανταστεί κανείς μια ανθρώπινη δραστηριότητα, στην οποία δεν εμπλέκεται η πληροφορική. Οι λόγοι γι' αυτό το φαινόμενο είναι πολλοί. Ο πρώτος σχετίζεται με τις **δυνατότητες και το κόστος του υλικού** (hardware) των υπολογιστών. Ένας συνηθισμένος σημερινός οικιακός υπολογιστής θα φάνταζε σαν εξοπλισμός σε σκηνικό ταινίας επιστημονικής φαντασίας πριν από ελάχιστες δεκαετίες. Αρκεί να αναλογιστεί κανείς ότι το βασικό μέσο αποθήκευσης ενός προσωπικού υπολογιστή πριν από 25 περίπου χρόνια ήταν οι δισκέτες με χωρητικότητα 720 KB, ενώ ένας σημερινός υπολογιστής αποθηκεύει δεδομένα σε σκληρό δίσκο χωρητικότητας εκατοντάδων GB. Οι σύγχρονοι υπολογιστές προσφέρουν πολύ μεγάλη υπολογιστική ισχύ και ικανότητα αποθήκευσης δεδομένων, ενώ ταυτόχρονα διατίθενται σε τιμές προσιτές στον απλό καταναλωτή. Επιπλέον όμως, η αγορά σήμερα κατακλύζεται και από ένα πλήθος συσκευών, όπως ευφυή τηλέφωνα (smartphones), tablets, ψηφιακές φωτογραφικές μηχανές κλπ. οι οποίες ενσωματώνουν τεχνολογίες υπολογιστών. Όλες αυτές οι συσκευές καθημερινά παράγουν και αποθηκεύουν δεδομένα. Παράλληλη με την ανάπτυξη του υλικού ήταν και η **ανάπτυξη του λογισμικού** (software). Ξεκινώντας από τα πρώτα απλά προγράμματα που εκτελούσαν βασικούς υπολογισμούς, έχουμε φτάσει στη σημερινή εποχή, όπου πάμπολλα λογισμικά κάθε είδους και σκοπού μπορούν να εκτελούν περίπλοκες εργασίες και χρησιμοποιούνται καθημερινά από εκατομμύρια ανθρώπους. Αποτέλεσμα των παραπάνω είναι η **διείσδυση της πληροφορικής σε κάθε δραστηριότητα της σύγχρονης κοινωνίας**. Η επιστήμη, η οικονομία, η εκπαίδευση, αλλά και η διασκέδαση και οι ανθρώπινες σχέσεις, κάνουν χρήση της σύγχρονης τεχνολογίας και έχουν επηρεαστεί από αυτήν.

Ένας άλλος καθοριστικός παράγοντας ήταν η έλευση του Διαδικτύου. Το **Διαδίκτυο** με τα ιδιαίτερα χαρακτηριστικά του, τα οποία επιτρέπουν αποτελεσματική επικοινωνία πάνω σε μη αξιόπιστες γραμμές, γνώρισε ταχύτατη διάδοση και εξαπλώθηκε σε όλη τη Γη, αποτελώντας μια καθολική πλατφόρμα διασύνδεσης υπολογιστών, διαμοιρασμού πόρων, διάχυσης πληροφορίας και επικοινωνίας. Πολύ γρήγορα κρατικοί οργανισμοί, επιχειρήσεις και άλλοι φορείς το χρησιμοποίησαν για να δικτυώσουν τα υπολογιστικά τους συστήματα και να προβάλουν πληροφορίες. Σήμερα όμως, η χρήση του Διαδικτύου δεν περιορίζεται σε αυτά. Η έλευση της **Υπολογιστικής Νέφους** (Cloud Computing), η οποία μεταθέτει την ανάγκη ύπαρξης υπολογιστικών πόρων και δεδομένων από τον τοπικό υπολογιστή στο δίκτυο, χαράσσει νέους δρόμους για την τεχνολογία της πληροφορικής. Η ρήση «The network is the computer» που αποδίδεται στον John Gage αποτυπώνει τη νέα αυτή πραγματικότητα. Μια νέα πραγματικότητα για το Διαδίκτυο αποτελεί και το λεγόμενο Web 2.0. Οι ιστότοποι κοινωνικής δικτύωσης, τα απειράριθμα blogs, τα wikis κλπ. μετατρέπουν τον απλό χρήστη του Διαδικτύου από παθητικό καταναλωτή πληροφοριών σε πρωταγωνιστικό παράγοντα παραγωγής πληροφορίας. Οι επιπτώσεις του **Web2.0** στην επικοινωνία, στις κοινωνικές σχέσεις και τη συμπεριφορά των ανθρώπων, στην πολιτική και στην οικονομία είναι τεράστιες, καθώς καθημερινά πολλά εκατομμύρια ανθρώπων καταγράφουν και διακινούν τις προσωπικές τους απόψεις, **συνεισφέροντας την προσωπική τους συμβολή στη διαμόρφωση της παγκόσμιας συνείδησης**.

Συνοψίζοντας τα παραπάνω, επισημαίνουμε ότι η εξέλιξη της τεχνολογίας της πληροφορικής στη σύγχρονη τεχνολογία χαρακτηρίζεται από:

- Τη διάθεση πολύ ισχυρού και φθηνού υλικού υπολογιστών, με ιδιαίτερη μνεία στις νέες φορητές συσκευές όπως τα tablets και τα smartphones.
- Την ανάπτυξη εξελιγμένου λογισμικού για κάθε χρήση.
- Την επικράτηση του Διαδικτύου και τις νέες του δυνατότητες, όπως η υπολογιστική νέφους και το Web2.0.
- Την καθολική διάχυση και ενσωμάτωση της νέας τεχνολογίας στη σύγχρονη κοινωνία.



Σχήμα 6.1 *Μαζική καταγραφή δεδομένων*

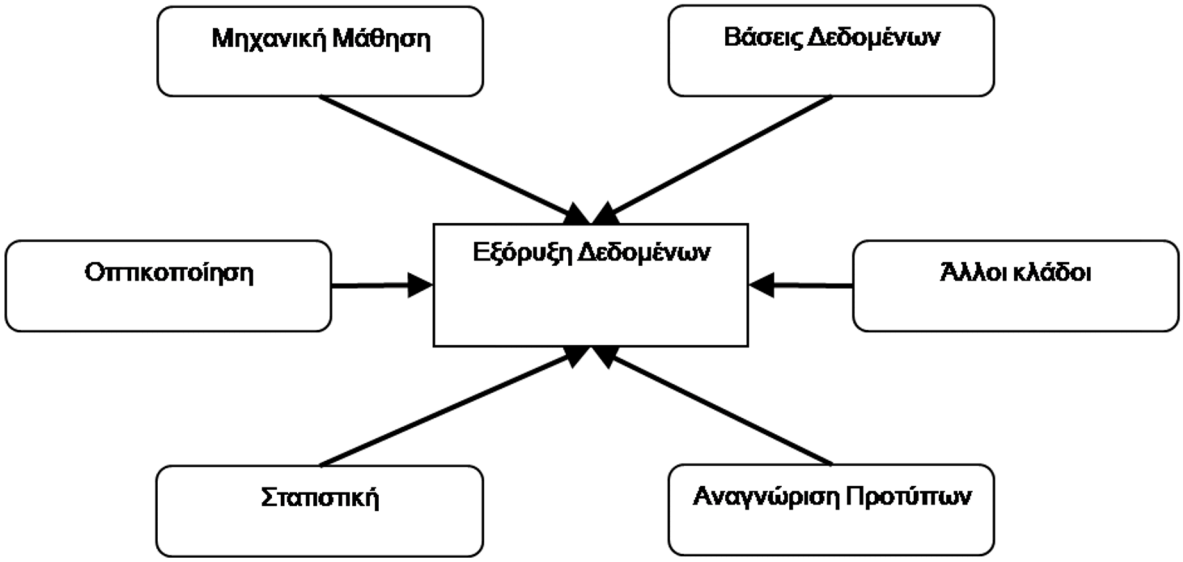
Πληροφορική σημαίνει καταρχήν δεδομένα. Η χρήση όλων αυτών των συσκευών, από εκατομμύρια ανθρώπους, για διάφορους σκοπούς και με συνεχώς επιταχυνόμενο ρυθμό, έχει σαν αποτέλεσμα την καθημερινή παραγωγή και αποθήκευση τεράστιων όγκων δεδομένων. Ο ρυθμός αύξησης της καταγραφής δεδομένων στη σημερινή εποχή είναι εκρηκτικός. Ένας μη έμπειρος αναγνώστης θα μπορούσε να θεωρήσει την προηγούμενη πρόταση υπερβολική και να την αποδώσει στη μεροληπτική άποψη ενός οπαδού της σύγχρονης τεχνολογίας. Όμως μια σύντομη περιήγηση στο διαδίκτυο δίνει εύκολα την απάντηση. Πάμπολλοι ιστότοποι επισημαίνουν το φαινόμενο της ραγδαίας αύξησης του όγκου των δεδομένων. Ενδεικτικά παραπέμπουμε σε ιστοσελίδα της IBM (Www-01.ibm.com, 2015), στην οποία αναφέρεται ότι παράγουμε 2,5 πεντάκις εκατομμύρια bytes την ημέρα και ότι το 90% των αποθηκευμένων δεδομένων έχουν παραχθεί την τελευταία διετία. Με μαθηματικούς όρους θα μπορούσαμε να πούμε ότι η συνάρτηση καταγραφής δεδομένων στον χρόνο είναι εκθετική. Επαναλαμβάνουμε λοιπόν, ότι ο ρυθμός αύξησης της καταγραφής δεδομένων είναι εκρηκτικός. Και επανερχόμενοι στην πρώτη πρόταση του κεφαλαίου την τροποποιούμε ελαφρώς. **Έχουμε κατακλυστεί από δεδομένα. Όμως που είναι η πληροφορία;.**

6.2 Εξόρυξη Δεδομένων - Ορισμός.

Η Εξόρυξη Δεδομένων αποτελεί την απάντηση στο ερώτημα – κατακλείδα του προηγούμενου υποκεφαλαίου. Ο ανθρώπινος νους έχει περιορισμένες αναλυτικές δυνατότητες. Ακόμα και χωρίς την εκρηκτική αύξηση του όγκου των δεδομένων που παρατηρήθηκε τα τελευταία χρόνια, τού είναι πολύ δύσκολο να επεξεργαστεί αποτελεσματικά τα διαθέσιμα δεδομένα. Μια τυπική εφαρμογή μηχανογραφημένης εμπορικής διαχείρισης σε μια μικρομεσαία επιχείρηση μπορεί να τηρεί στοιχεία δεκάδων χιλιάδων συναλλαγών. Η επεξεργασία των στοιχείων αυτών χωρίς εξειδικευμένα εργαλεία, αν δεν είναι αδύνατη, είναι αργή, ακριβή και εν πολλοίς υποκειμενική. Η επιστήμη της Στατιστικής προσφέρει λύσεις ανάλυσης δεδομένων, δεν λαμβάνει όμως μέριμνα για το πρόβλημα του πολύ μεγάλου όγκου τους. Επίσης η Μηχανική Μάθηση και η Αναγνώριση Προτύπων διαθέτουν τις δικές τους μεθοδολογίες, όμως και πάλι δεν αντιμετωπίζουν το πρόβλημα του όγκου των δεδομένων. Ο κλάδος των Βάσεων Δεδομένων είναι ο κατ' εξοχήν αρμόδιος για την τήρηση μεγάλου όγκου δεδομένων, όμως η σχεδιαστική φιλοσοφία του είναι προσανατολισμένη στην καταχώρηση, στη διαχείριση και στην ανάκτηση των δεδομένων, όχι όμως και στην ανάλυση τους.

Η Εξόρυξη Δεδομένων αποτελεί τέκνο της ανάγκης για επεξεργασία των αποθηκευμένων δεδομένων και εξαγωγή χρήσιμης πληροφορίας. Αντλώντας μεθοδολογίες από όλους τους επιστημονικούς κλάδους που αναφέρθηκαν παραπάνω, καθώς και από άλλους, όπως η Οπτικοποίηση, στοχεύει στην ανακάλυψη πολύτιμης γνώσης, που είναι κρυμμένη σε μεγάλους όγκους δεδομένων. Το όνομα της παραπέμπει στην εξόρυξη πολύτιμων μετάλλων, όπου οι επίδοξοι χρυσοθήρες αναζητούν ψήγματα χρυσού σε όγκους χώματος. Ως ένας

διεπιστημονικός κλάδος, η Εξόρυξη Δεδομένων προσέλκυσε επιστήμονες από διαφορετικούς χώρους. Οι διάφοροι ορισμοί που κατά καιρούς έχουν διατυπωθεί, αντανακλούν σε ορισμένο βαθμό την οπτική γωνία των συγγραφέων τους.



Σχήμα 6.2 *Η Εξόρυξη Δεδομένων ως αποτέλεσμα συμβολής άλλων κλάδων.*

Σύμφωνα με τους Witten and Frank (2000), η **Εξόρυξη Δεδομένων** (ΕΔ) (**Data Mining** (DM)) ορίζεται ως η διαδικασία ανακάλυψης προτύπων μέσα από δεδομένα, δίνοντας έτσι έμφαση στη διάσταση της Μηχανικής Μάθησης. Σύμφωνα με τους Han and Kamber (2001), η Εξόρυξη Δεδομένων συνίσταται στην ανακάλυψη ή «εξόρυξη» γνώσης από μεγάλους όγκους δεδομένων. Ο ορισμός αυτός τονίζει τη διάσταση του όγκου των δεδομένων. Άλλοι συγγραφείς, όπως οι Maimon and Rokach (2005), χρησιμοποιούν τον όρο **Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases – KDD)** για τη συνολική διαδικασία ανακάλυψης προτύπων μέσα από μεγάλα και περίπλοκα σύνολα δεδομένων. Σύμφωνα με το σκεπτικό αυτό, η ανακάλυψη γνώσης από τα δεδομένα συνίσταται σε μια διαδικασία, που ξεκινά από τα πηγαία δεδομένα και καταλήγει στην τελική διατύπωση συμπερασμάτων και στη λήψη αποφάσεων, μέσα από μια αλληλουχία διαδοχικών σταδίων. Η καθαυτό Εξόρυξη Δεδομένων αποτελεί ένα από τα στάδια αυτής της διαδικασίας και συνίσταται στον σκληρό πυρήνα της. Περιλαμβάνει την εφαρμογή αλγορίθμων και την κατασκευή μοντέλων, τα οποία στοχεύουν στην ανακάλυψη και εξαγωγή προτύπων (patterns). Η συνολική διαδικασία ανακάλυψης γνώσης καλείται να αντιμετωπίσει προβλήματα, όπως το πώς είναι αποθηκευμένα τα δεδομένα και πως επιτυγχάνεται η πρόσβαση σε αυτά, πως οι αλγόριθμοι εξαγωγής προτύπων πρέπει να κλιμακωθούν, ώστε να είναι ικανοί να χειριστούν τον όγκο των δεδομένων, πως θα γίνει η οπτικοποίηση των αποτελεσμάτων, ώστε να καταστούν κατανοητά κλπ.

Ένα ζήτημα που οφείλει να γίνει κατανοητό, είναι το τι σημαίνει «ανακάλυψη γνώσης», δηλαδή τι τελικά εξάγεται. Από τεχνικής σκοπιάς αυτό που τελικά εξάγεται είναι κανονικότητες και πρότυπα δεδομένων, που περιγράφουν ή διαφοροποιούν κατηγορίες ή περιπτώσεις και που πιθανώς να χρησιμεύσουν για τη διατύπωση προβλέψεων. Η αξιολόγηση αυτών των προτύπων και η ερμηνεία τους συνιστά τη γνώση, όπως αυτή γίνεται αντιληπτή από τον άνθρωπο.

6.3 Στάδια της Διαδικασίας Ανακάλυψης Γνώσης

Έχει επικρατήσει η πρακτική, κυρίως στη βιομηχανία λογισμικού, να θεωρούνται οι όροι «Εξόρυξη Δεδομένων» και «Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων» ως συνώνυμοι, και να χρησιμοποιούνται για να περιγράψουν τη συνολική διαδικασία ανακάλυψης γνώσης. Η πρακτική αυτή ακολουθείται χάριν ευκολίας, γιατί ο όρος «Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων» είναι δύσχρηστος. Επί της ουσίας όμως, επικρατεί η κοινή άποψη ότι η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (ΑΓΒΔ) είναι μια ευρύτερη διαδικασία και ότι η καθεαυτό Εξόρυξη Δεδομένων αποτελεί ένα στάδιο της.

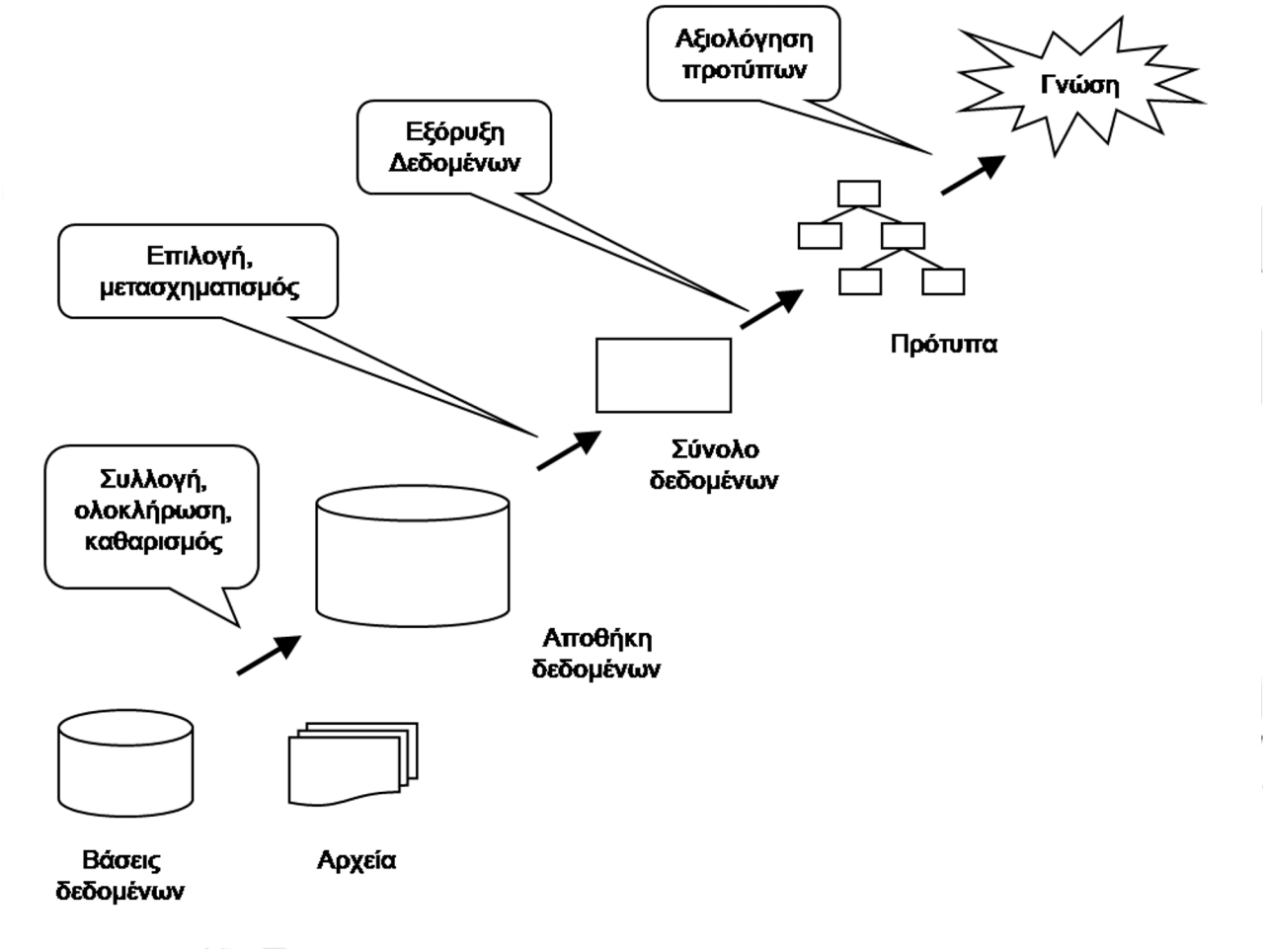
Η ΑΓΒΔ έχει ως αφετηριακό σημείο της τα πηγαία δεδομένα και ως σκοπό της την κατάλληλη επεξεργασία τους, ώστε να ανακαλυφθεί χρήσιμη γνώση. Όμως μεταξύ του αρχικού σημείου και του τελικού στόχου μεσολαβούν ενδιάμεσα στάδια, κατά τα οποία πρέπει να λάβουν χώρα συγκεκριμένες εργασίες, που επηρεάζουν σημαντικά το τελικό αποτέλεσμα. Επίσης, πρέπει να τονιστεί ότι η διαδικασία δεν είναι μονόδρομη και ότι ο αναλυτής πιθανώς θα χρειαστεί να επανέλθει σε κάποιο προηγούμενο στάδιο, να τροποποιήσει τις πρακτικές του και να επαναλάβει τα ακόλουθα στάδια. Τα στάδια της διαδικασίας ΑΓΒΔ και οι ειδικότερες εργασίες που λαμβάνουν χώρα σε κάθε ένα από αυτά είναι τα ακόλουθα:

Συλλογή, Ολοκλήρωση και Καθαρισμός των Δεδομένων. Αφετηριακό σημείο είναι τα πηγαία δεδομένα. Κατά κανόνα, τα δεδομένα αυτά είναι αποθηκευμένα σε διάφορες πηγές, όπως συστήματα παρακολούθησης συναλλαγών, ανεξάρτητες βάσεις δεδομένων, ανεξάρτητα αρχεία, εξωτερικές πηγές κλπ. Επίσης, τα δεδομένα αυτά πάσχουν από διάφορα προβλήματα όπως σφάλματα, αντιφάσεις, χαμένες τιμές κλπ. Τα αρχικά δεδομένα πρέπει να συλλεχτούν από τις διάφορες πηγές, να ομογενοποιηθούν και να καθαριστούν. Προβλήματα των πηγαίων δεδομένων, καθώς και τρόποι αντιμετώπισης ορισμένων από αυτά, έχουν παρουσιαστεί στο Κεφάλαιο 4, που αναφέρεται στις Αποθήκες Δεδομένων, και ειδικότερα στο υποκεφάλαιο που αναφέρεται στις εργασίες ETL. Πρόσθετες τεχνικές για την αντιμετώπιση προβλημάτων που επηρεάζουν τη διαδικασία ανακάλυψης προτύπων, όπως πχ το πρόβλημα των χαμένων τιμών, παρατίθενται στο Κεφάλαιο 7, το οποίο αναφέρεται στην Προεπεξεργασία των δεδομένων. Στο σημείο αυτό οφείλουμε να τονίσουμε τη σημασία της ποιότητας των δεδομένων. Προβληματικά δεδομένα, τα οποία περιέχουν εσφαλμένες, ακραίες ή χαμένες τιμές μπορεί να αποπροσανατολίσουν τους αλγόριθμους εξόρυξης και να οδηγήσουν στην εξαγωγή άκυρων και εσφαλμένων προτύπων. Ορισμένοι αλγόριθμοι εξόρυξης δεδομένων έχουν τη δυνατότητα να αντιμετωπίζουν ενδογενώς τα προβλήματα των δεδομένων, ωστόσο αυτό δεν ισχύει γενικώς και επίσης, ο τρόπος αντιμετώπισης δεν είναι πάντα ο καλύτερος. Για τον λόγο αυτό, είναι προτιμότερο ο καθαρισμός των δεδομένων να γίνει από τον αναλυτή ως ανεξάρτητη εργασία και με τρόπο ελεγχόμενο από αυτόν. Συνήθως, τα δεδομένα, αφού απαλλαγούν από τα προβλήματα τους, αποθηκεύονται σε μια Αποθήκη Δεδομένων.

Επιλογή Δεδομένων και Μετασχηματισμός τους. Για να διεξαχθεί επιτυχώς η Εξόρυξη Δεδομένων, πρέπει καταρχήν να δημιουργηθεί το κατάλληλο σύνολο δεδομένων (data set). Οι μέθοδοι της ΕΔ είναι ισχυρώς καθοδηγούμενες από τα δεδομένα (data driven). Αυτό σημαίνει ότι το όποιο αποτέλεσμα θα αντληθεί απευθείας από τα δεδομένα. Για τον λόγο αυτό, η επιλογή των κατάλληλων δεδομένων είναι κομβικής σημασίας. Επιλογή δεδομένων σημαίνει καταρχάς επιλογή των κατάλληλων γνωρισμάτων ή χαρακτηριστικών (attributes – features). Με όρους πίνακα, αυτό θα σήμαινε επιλογή των κατάλληλων στηλών. Η επιλογή των γνωρισμάτων είναι άμεσα συναρτημένη με την εργασία που εκτελεί ο αναλυτής. Ορισμένα γνωρίσματα μπορεί να είναι χρήσιμα για μια εργασία, ενώ ορισμένα άλλα για κάποια άλλη. Ο αναλυτής αρχικά επιλέγει τα γνωρίσματα τα οποία θεωρεί ότι περιέχουν ουσιαστική πληροφορία που σχετίζεται με την ανάλυση του. Αν για παράδειγμα, μελετά οικονομικά στοιχεία επιχειρήσεων και τις επιπτώσεις τους σε κάποιο φαινόμενο, όπως τη λήψη αρνητικών σχολίων από τους εξωτερικούς ελεγκτές, τότε εύκολα μπορεί να εξαιρέσει δεδομένα όπως τον τηλεφωνικό αριθμό ή τη διεύθυνση. Είναι βασικό να μην εξαιρέσει κανένα σημαντικό γνώρισμα, γιατί τότε θα αφαιρούσε χρήσιμη πληροφορία από την ανάλυση του. Σε ορισμένες περιπτώσεις πρέπει να γίνει και επιλογή παραδειγμάτων, αντικειμένων ή γραμμών του πίνακα. Η εργασία αυτή ονομάζεται **δειγματοληψία** και γίνεται όταν τα δεδομένα είναι πάρα πολλά και οι αλγόριθμοι επεξεργασίας έχουν πρόβλημα με τον χειρισμό του όγκου τους.

Η αρχική υποκειμενική επιλογή χαρακτηριστικών δεν είναι αρκετή. Στο αρχικό στάδιο, ο αναλυτής αποκλείει τα γνωρίσματα που εμφανώς δεν σχετίζονται με την ανάλυση του. Στη συνέχεια όμως, η επιλογή μπορεί να μην είναι προφανής, καθώς το ίδιο μέγεθος μπορεί να καταγράφεται με διαφορετικούς τρόπους. Η κερδοφορία πχ μιας επιχείρησης καταγράφεται μέσω απόλυτων τιμών, όπως κέρδη προ φόρων και τόκων ή καθαρά κέρδη, καθώς και μέσω αριθμοδεικτών, όπως κέρδη προς σύνολο ενεργητικού ή κέρδη προς σύνολο πωλήσεων. Είναι προς διερεύνηση ποιο από όλα αυτά τα γνωρίσματα είναι το πλέον κατάλληλο για την ανάλυση. Δεδομένου ότι αναλύονται πολλά στοιχεία, κάθε ένα από τα οποία έχει καταγραφεί με πολλούς τρόπους, καθώς και ότι μπορεί να υπάρχουν σχέσεις αλληλεξάρτησης μεταξύ των γνωρισμάτων, το πρόβλημα περιπλέκεται. Είναι χαρακτηριστικό ότι σε διάφορες μελέτες για την ανάπτυξη μοντέλων πρόβλεψης χρεοκοπίας έχουν χρησιμοποιηθεί περισσότεροι από 500 αριθμοδείκτες (Du Jardin, 2010). Τίθεται λοιπόν το ερώτημα ποιο είναι το βέλτιστο υποσύνολο γνωρισμάτων, το οποίο είναι το πλέον κατάλληλο για τη συγκεκριμένη εργασία που εκτελεί ο αναλυτής. Η εργασία επιλογής στηλών ονομάζεται επιλογή **χαρακτηριστικών** (feature selection) και για την ολοκλήρωση της εφαρμόζονται εξειδικευμένες μέθοδοι, ορισμένες από τις οποίες παρουσιάζονται στο Κεφάλαιο 7.

Στο στάδιο αυτό γίνεται και ο **μετασχηματισμός** των δεδομένων. Για παράδειγμα, μπορεί να γίνει αναγωγή αριθμητικών τιμών σε άλλες αριθμητικές τιμές ή μετατροπή αριθμητικών τιμών σε ονομαστικές τιμές. Τέτοιες εργασίες γίνονται συνήθως για να προσαρμοστούν τα δεδομένα σε απαιτήσεις των μεθόδων ανάλυσης. Για παράδειγμα, ορισμένες μέθοδοι κατηγοριοποίησης επηρεάζονται ισχυρά από πεδία με μεγάλες τιμές και ασθενώς από πεδία με μικρές τιμές. Σε τέτοιες περιπτώσεις, οι τάξεις μεγέθους των τιμών στα διάφορα πεδία πρέπει να είναι συγκρίσιμες. Το τελικό αποτέλεσμα αυτού του σταδίου είναι ένα σύνολο δεδομένων που θα χρησιμοποιηθεί για την εξαγωγή προτύπων.



Σχήμα 6.3. *Στάδια της διαδικασίας Ανακάλυψης Γνώσης*

Εξόρυξη Δεδομένων. Στο στάδιο αυτό γίνεται η καθεαυτό Εξόρυξη Δεδομένων, δηλαδή η εξαγωγή προτύπων. Αρχικά ο αναλυτής πρέπει να επιλέξει το είδος εργασίας Εξόρυξης Δεδομένων που θα εφαρμόσει. Σε γενικές γραμμές, ο χρήστης μπορεί να εκτελέσει Περιγραφική Ανάλυση (descriptive analytics) ή Προγνωστική Ανάλυση (predictive analytics). Η Περιγραφική Ανάλυση στοχεύει στην κατάδειξη ομαδοποιήσεων και ιδιοτήτων των δεδομένων, χωρίς να επιδιώκει τη διατύπωση προβλέψεων. Αντιθέτως, η προγνωστική ανάλυση στοχεύει στη διατύπωση προβλέψεων για το μέλλον, συνήθως με την οικοδόμηση κάποιου μοντέλου. Υπάρχουν διάφορα είδη εργασιών ΕΔ, όπως η κατηγοριοποίηση, η ανάλυση συστάδων και η ανάλυση κανόνων συσχέτισης. Αναλόγως με το πρόβλημα που θέλει να μελετήσει, ο αναλυτής επιλέγει το είδος εργασίας ΕΔ. Επιπλέον, για κάθε είδος εργασίας υπάρχουν πολλές μέθοδοι. Για παράδειγμα, κατηγοριοποίηση μπορεί να γίνει με χρήση Νευρωνικών Δικτύων, Δένδρων Αποφάσεων, Μπαϋεσιανών Κατηγοριοποιητών κλπ. Κάθε μια από αυτές τις μεθόδους πλεονεκτεί σε ορισμένα θέματα έναντι των άλλων και μειονεκτεί σε άλλα. Ο αναλυτής, ανάλογα με τις προτεραιότητες του, θα επιλέξει τη μέθοδο που θα χρησιμοποιήσει. Επίσης, είναι πιθανό να χρησιμοποιήσει περισσότερες από μια μεθόδους και να συγκρίνει τα αποτελέσματα. Κατά τη διάρκεια της εφαρμογής των μεθόδων, ο αναλυτής θα χρειαστεί να ρυθμίσει τις παραμέτρους τους. Για παράδειγμα, σε ένα Νευρωνικό Δίκτυο θα ρυθμίσει το πλήθος εποχών, τον ρυθμό μάθησης κλπ.

Αξιολόγηση Προτύπων και Ανακάλυψη Γνώσης. Ο αναλυτής αξιολογεί τα πρότυπα δεδομένων που προέκυψαν από το προηγούμενο στάδιο. Εάν τα αποτελέσματα δεν είναι ικανοποιητικά θα επανέλθει σε προηγού-

μενα στάδια και θα επαναλάβει τις εργασίες. Πιθανώς θα χρειαστεί να τροποποιήσει το σύνολο δεδομένων ή να χρησιμοποιήσει διαφορετικές μεθόδους ΕΔ. Για την αναπαράσταση της γνώσης μπορούν να εφαρμοστούν τεχνικές οπτικοποίησης. Με την ολοκλήρωση αυτού του σταδίου ο αναλυτής έχει εξάγει τα συμπεράσματα του και έπεται η λήψη αποφάσεων και η ανάληψη δράσης.

Στο Σχήμα 6.3 απεικονίζονται τα στάδια της διαδικασίας ανακάλυψης γνώσης.

6.4 Εργασίες Εξόρυξης Δεδομένων

Στο στάδιο της Εξόρυξης Δεδομένων ο αναλυτής εξάγει πρότυπα από το σύνολο δεδομένων που έχει δημιουργήσει. Ανάλογα με το είδος του προβλήματος και τον τύπο της ανάλυσης που θα διεξάγει, μπορεί να αναζητήσει πρότυπα διάφορων τύπων. Υπάρχουν ποικίλες εργασίες Εξόρυξης Δεδομένων. Οι εργασίες αυτές μπορούν να ταξινομηθούν με διάφορους τρόπους. Ένας διαχωρισμός των μεθόδων ΕΔ είναι σε μεθόδους **επιβλεπόμενης μάθησης** (supervised learning) και μεθόδους **μη επιβλεπόμενης μάθησης** (unsupervised learning). Η επιβλεπόμενη μάθηση έχει στόχο τη μοντελοποίηση των σχέσεων ανάμεσα σε ένα εξαρτημένο γνώρισμα – στόχο και σε άλλα ανεξάρτητα γνωρίσματα. Η ανάλυση συνίσταται στην τυποποίηση των σχέσεων ανάμεσα στην εξαρτημένη και στις ανεξάρτητες μεταβλητές, συνήθως με τη δημιουργία ενός μοντέλου, που επιτρέπει τον υπολογισμό της εξαρτημένης μεταβλητής από τις ανεξάρτητες. Το μοντέλο μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Το όνομα «επιβλεπόμενη μάθηση» σημαίνει ότι το γνώρισμα στόχος και οι τιμές του καθοδηγούν τη διαδικασία μάθησης. Στη μη επιβλεπόμενη μάθηση δεν υπάρχει κάποια στήλη στόχος και οι αλγόριθμοι προσπαθούν να ομαδοποιήσουν τα δεδομένα σε ομάδες που δεν είναι γνωστές εκ των προτέρων.

Οι βασικές εργασίες Εξόρυξης Δεδομένων είναι οι ακόλουθες:

Κατηγοριοποίηση. Η κατηγοριοποίηση είναι μια από τις συνηθέστερες εργασίες Εξόρυξης Δεδομένων. Πρόκειται για εργασία επιβλεπόμενης μάθησης. Σε προβλήματα Κατηγοριοποίησης είναι γνωστό εκ των προτέρων ότι τα δεδομένα υπάγονται σε κατηγορίες. Σε ένα από τα γνωρίσματα καταγράφεται η κατηγορία των αντικειμένων. Ένα παράδειγμα συνόλου δεδομένων, κατάλληλο για κατηγοριοποίηση, είναι τα στοιχεία αιτήσεων χορήγησης τραπεζικών δανείων. Στα γνωρίσματα καταγράφονται τα στοιχεία των πελατών, όπως ηλικία, οικονομική κατάσταση κλπ. και σε ένα γνώρισμα αναφέρεται το εάν εγκρίνεται ή απορρίπτεται το δάνειο. Η έγκριση ή η απόρριψη εξαρτάται από τα στοιχεία του κάθε πελάτη. Με την κατηγοριοποίηση δημιουργείται ένας μηχανισμός υπολογισμού της κατηγορίας του κάθε αντικειμένου από τα υπόλοιπα γνωρίσματα του. Στο παράδειγμα με τα τραπεζικά δάνεια, ένα σύνολο κανόνων, οι οποίοι ορίζουν για ποιες ηλικίες, εισοδήματα και άλλα στοιχεία εγκρίνεται το δάνειο, ενώ για ποιες όχι, είναι ένα μοντέλο κατηγοριοποίησης. Το μοντέλο δεν είναι υποχρεωτικά κανόνες, αλλά μπορεί να έχει άλλες μορφές, όπως πχ να συνίσταται σε ένα πλέγμα κόμβων και συνδέσεων ενός νευρωνικού δικτύου. Αφού δημιουργηθεί το μοντέλο, μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Για μια νέα αίτηση δανείου μπορούν να εισαχθούν τα στοιχεία του πελάτη στο μοντέλο και αυτό να προβλέψει εάν θα εγκριθεί ή θα απορριφθεί η αίτηση. Σημειωτέον ότι αυτό που θα προβλεφθεί είναι η κατηγορία του κάθε αντικειμένου, δηλαδή μια ονομαστική τιμή. Υπάρχουν πολλές δυνατότητες εφαρμογής τεχνικών κατηγοριοποίησης στον κόσμο των επιχειρήσεων. Η πρόβλεψη χρεοκοπίας επιχειρήσεων και άλλων οργανισμών είναι ίσως το πιο γνωστό παράδειγμα. Επίσης, στον χρηματοπιστωτικό τομέα, η εκτίμηση της πιστοληπτικής ικανότητας και η διαχείριση του ρίσκου είναι δύο τυπικά πεδία εφαρμογής. Σημαντικές εφαρμογές βρίσκει η κατηγοριοποίηση και στον χώρο της διαφήμισης και των πωλήσεων. Η ένταξη πελατών σε προκαθορισμένες κατηγορίες χρησιμοποιείται για την προσέλκυση πελατών με άμεσο διαπροσωπικό μάρκετινγκ, καθώς και για την κατάρτιση προγραμμάτων επιβράβευσης πελατών και εκχώρησης πελατών σε συγκεκριμένα προγράμματα. Στη διαχείριση ανθρωπίνων πόρων χρησιμοποιούνται τεχνικές κατηγοριοποίησης για την πρόβλεψη της επίδοσης των εργαζομένων.

Παλινδρόμηση. Η παλινδρόμηση είναι μια εργασία επιβλεπόμενης μάθησης που μοιάζει πολύ με την κατηγοριοποίηση. Υπάρχει πάλι ένα γνώρισμα-στόχος, οι τιμές του οποίου υπολογίζονται από τα υπόλοιπα γνωρίσματα. Οι αλγόριθμοι παλινδρόμησης εξετάζουν τις σχέσεις μεταξύ του γνωρίσματος-στόχου και των υπόλοιπων γνωρισμάτων και κατασκευάζουν έναν μηχανισμό υπολογισμού. Η διαφορά με την κατηγοριοποίηση είναι ότι στην περίπτωση της παλινδρόμησης υπολογίζονται αριθμητικές τιμές. Ένα μοντέλο που υπολογίζει το ύψος των ανθρώπων από άλλα στοιχεία τους, όπως το ύψος των γονέων τους, τις διατροφικές συνήθειες τους κλπ. είναι ένα μοντέλο παλινδρόμησης. Στις επιχειρήσεις, τεχνικές παλινδρόμησης χρησιμοποιούνται για την πρόβλεψη αριθμητικών τιμών, όπως το ύψος των πωλήσεων, το ύψος των κερδών κλπ.

Ανάλυση Συστάδων. Η Ανάλυση Συστάδων είναι και αυτή μια πολύ συνηθισμένη εργασία Εξόρυξης Δεδομένων και ανήκει στην κατηγορία εργασιών μη επιβλεπόμενης μάθησης. Στόχος της Ανάλυσης Συστάδων είναι ο επιμερισμός ενός συνόλου αντικειμένων σε ομάδες. Η διαφορά με την Κατηγοριοποίηση έγκειται στο

γεγονός ότι δεν υπάρχουν κατηγορίες, οι οποίες είναι εκ των προτέρων γνωστές, δεν υπάρχει δηλαδή ένα γνώρισμα στο οποίο καταγράφεται η κατηγορία των αντικειμένων. Οι αλγόριθμοι της Ανάλυσης Συστάδων προσπαθούν να δημιουργήσουν ομάδες αναζητώντας ομοειδή αντικείμενα. Στόχος τους είναι να μεγιστοποιήσουν την ομοιότητα εντός των ομάδων και την ανομοιότητα μεταξύ των ομάδων. Αυτό σημαίνει ότι τα αντικείμενα της μιας ομάδας πρέπει να μοιάζουν μεταξύ τους και να μην μοιάζουν με τα αντικείμενα των άλλων ομάδων. Αφού σχηματιστούν οι ομάδες, μπορούν να θεωρηθούν ως κατηγορίες και να δημιουργηθούν κανόνες που να τις περιγράφουν. Ένα πολύ συνηθισμένο παράδειγμα εφαρμογής της Ανάλυσης Συστάδων είναι ο επιμερισμός των πελατών σε ομοειδείς ομάδες. Η εφαρμογή αυτή ονομάζεται τμηματοποίηση αγοράς και είναι κεφαλαιώδους σημασίας για το μάρκετινγκ, γιατί επιτρέπει τη διεξαγωγή στοχευμένης διαφήμισης. Εάν η επιχείρηση γνωρίζει τα κοινά χαρακτηριστικά μεγάλων ομάδων πελατών, μπορεί να σχεδιάσει διαφημιστικές καμπάνιες ειδικά προσαρμοσμένες στα χαρακτηριστικά και τις απαιτήσεις αυτών των ομάδων. Ένα άλλο πολύ συνηθισμένο παράδειγμα εφαρμογής τεχνικών Ανάλυσης Συστάδων είναι η διαχείριση παραπόνων και αιτημάτων πελατών. Τα κέντρα κλήσεων των επιχειρήσεων γίνονται καθημερινά αποδέκτες χιλιάδων μηνυμάτων πελατών, οι οποίοι ζητούν οδηγίες ή τεχνική υποστήριξη, διατυπώνουν παράπονα κλπ. Η κατάλληλη ομαδοποίηση αυτών των μηνυμάτων διευκολύνει τη διαχείριση τους, επιτρέπει την ορθή δρομολόγηση τους χωρίς σφάλματα και καθυστερήσεις, και βοηθά την επιχείρηση να κατανοήσει καλύτερα τις ανάγκες των πελατών της, έτσι ώστε να προβεί στις αναγκαίες ενέργειες. Το τελικό αποτέλεσμα είναι η αύξηση της ικανοποίησης των πελατών. Ανάλυση συστάδων εφαρμόζεται και για τη μελέτη γεωγραφικής πληροφορίας, έτσι ώστε να οριστούν περιοχές και να επιτευχθεί βέλτιστη διασπορά των υποκαταστημάτων της επιχείρησης.

Ανάλυση Κανόνων Συσχέτισης. Η Ανάλυση Κανόνων Συσχέτισης θεωρείται το πιο γνήσιο τέκνο της Εξόρυξης Δεδομένων, καθώς οι άλλες μέθοδοι προέρχονται από τη Μηχανική Μάθηση, τη Στατιστική κλπ. Στόχος των Κανόνων Συσχέτισης είναι η ανακάλυψη σχέσεων μεταξύ τιμών των γνωρισμάτων, οι οποίες εμφανίζονται συχνά μαζί. Για την καλύτερη κατανόηση του αντικειμένου παραθέτουμε το παρακάτω παράδειγμα. Θεωρήστε τις πωλήσεις ενός σούπερ μάρκετ. Για κάθε συναλλαγή πώλησης (απόδειξη λιανικής) καταγράφονται τα προϊόντα που αγόρασε ο καταναλωτής. Το ερώτημα είναι εάν υπάρχουν προϊόντα τα οποία πωλούνται συχνά μαζί, εάν υπάρχουν δηλαδή ομάδες καταναλωτών που επιλέγουν να αγοράσουν κοινά προϊόντα. Οι Κανόνες Συσχέτισης ανακαλύπτουν τέτοιες σχέσεις και τις ποσοτικοποιούν, καταγράφοντας ποσοστά εμφάνισης τους. Για παράδειγμα, εξορύσσονται κανόνες που αναφέρουν ότι όταν αγοράζεται το προϊόν Α, τότε αγοράζεται ταυτόχρονα και το προϊόν Β, και παρατίθενται οι πιθανότητες εμφάνισης αυτού του γεγονότος. Οι Κανόνες Συσχέτισης μπορούν να χρησιμοποιηθούν για τη διαρρύθμιση των ραφιών ενός σούπερ μάρκετ, ώστε παροτρύνοντας τον καταναλωτή να αυξηθούν οι πωλήσεις. Γενικώς, η Ανάλυση Κανόνων Συσχέτισης εφαρμόζεται σε μεγάλο βαθμό για την επίτευξη διασταυρούμενων πωλήσεων. Ειδικά στο ηλεκτρονικό εμπόριο, όπου υπάρχει άμεση τροφοδότηση δεδομένων από τον πελάτη, καθώς και δυνατότητα άμεσης ανάλυσης αυτών των δεδομένων και αντιπαραβολής με ιστορικά στοιχεία, η προώθηση πρόσθετων προϊόντων στον πελάτη γίνεται άμεσα, την ώρα της επίσκεψης στην ιστοθέση. Ο πελάτης πραγματοποιεί αγορές και ταυτόχρονα οι αλγόριθμοι εξόρυξης δεδομένων εντοπίζουν άλλα προϊόντα, τα οποία πωλούνται συχνά μαζί με τα προϊόντα που επέλεξε ο συγκεκριμένος πελάτης. Άμεσα παρουσιάζεται στον πελάτη ένα μήνυμα της μορφής «Οι πελάτες που αγόρασαν αυτά τα προϊόντα αγόρασαν επίσης τα παρακάτω προϊόντα». Η εξόρυξη Κανόνων Συσχέτισης επιτρέπει τον εντοπισμό καταναλωτικών προτύπων και την καλύτερη κατανόηση των πραγματικών αναγκών των πελατών. Οι πληροφορίες αυτές χρησιμοποιούνται για την προσωποποιημένη απεύθυνση στον πελάτη και τη διεξαγωγή μάρκετινγκ ένα-προς-ένα.

Ανάλυση Εξαιρέσεων. Όλες οι εργασίες που αναφέρθηκαν μέχρι τώρα αφορούν τη διατύπωση «κανόνων γενικής ισχύος», την τυποποίηση δηλαδή σχέσεων που αφορούν μεγάλες ομάδες αντικειμένων. Στις εργασίες αυτές ιδιόμορφα και σπάνια γεγονότα απορρίπτονται ως μη χρήσιμη πληροφορία. Υπάρχουν όμως περιπτώσεις, που το ενδιαφέρον βρίσκεται ακριβώς στις εξαιρέσεις. Ας θεωρήσουμε την περίπτωση μιας κλεμμένης πιστωτικής κάρτας. Οι νόμιμοι κάτοχοι πιστωτικών καρτών έχουν μια «φυσιολογική συμπεριφορά» και πραγματοποιούν λογικές αγορές με κάποια συχνότητα. Μια κλεμμένη πιστωτική κάρτα έχει διαφορετική συμπεριφορά. Συνήθως ο φορέας της επιχειρεί πρώτα να πραγματοποιήσει μια συναλλαγή ευτελούς αξίας, για να ελέγξει εάν η κάρτα είναι σε ισχύ. Η συναλλαγή πρέπει να αφορά ένα πολύ μικρό ποσό, ώστε εάν υπάρχει πρόβλημα με την κάρτα να μην υπάρχουν αντιδράσεις. Στη συνέχεια, εάν η κάρτα δεν έχει απενεργοποιηθεί, ο φορέας επιχειρεί το συντομότερο δυνατό να πραγματοποιήσει σε άλλο κατάστημα αγορές υψηλής αξίας και να εξαντλήσει το πιστωτικό περιθώριο. Δεδομένου ότι οι κλεμμένες πιστωτικές κάρτες είναι ένα πολύ μικρό ποσοστό, το ενδιαφέρον σε αυτήν την περίπτωση είναι ο εντοπισμός της εξαίρεσης. Οι μέθοδοι Ανάλυσης Εξαιρέσεων επιχειρούν να εντοπίσουν τις εξαιρέσεις, χρησιμοποιώντας στατιστικές κατανομές πιθανοτήτων ή μέτρα απόστασης που βασίζονται στην ομοιότητα. Ο εντοπισμός κλεμμένων πιστωτικών καρτών

και γενικότερα η αντιμετώπιση της απάτης, είναι μια πολύ σημαντική, αλλά όχι η μοναδική εφαρμογή της Ανάλυσης Εξαιρέσεων. Η μη ομαλή συμπεριφορά διαφόρων δεικτών και ποσοτήτων μπορεί να σηματοδοτεί προβλήματα σε διάφορες λειτουργίες. Παραδείγματα περιπτώσεων μη ομαλής συμπεριφοράς δεικτών είναι η απότομη πτώση της ζήτησης, η πτώση του χρόνου ανταπόκρισης στην παροχή υπηρεσιών, η απότομη αύξηση της κυκλοφορίας σε ένα τηλεπικοινωνιακό δίκτυο ή των φορτίων σε ένα δίκτυο ηλεκτρικής ενέργειας κλπ. Ο άμεσος εντοπισμός τέτοιων ανωμαλιών επιτρέπει την έγκαιρη διάγνωση του προβλήματος και την ανάληψη κατάλληλης δράσης για τη θεραπεία του προβλήματος. Η ασφάλεια των δικτύων υπολογιστών των επιχειρήσεων και ο εντοπισμός εισβολών από μη εξουσιοδοτημένα άτομα είναι ένα ακόμα πεδίο εφαρμογής της Ανάλυσης Εξαιρέσεων.

Ανάλυση Χρονοσειρών. Υπάρχουν μεγέθη τα οποία παρουσιάζουν μια χρονική εξέλιξη. Η εξέλιξη αυτή αναπαρίσταται με τη βοήθεια χρονοσειρών, δηλαδή ακολουθιών σημείων που αποτελούν μετρήσεις του μεγέθους στη διάρκεια του χρόνου. Οι μέθοδοι ανάλυσης χρονοσειρών αναλύουν τα δεδομένα διαφορετικών χρονικών περιόδων και εξάγουν χρήσιμα συμπεράσματα για το φαινόμενο. Εάν για παράδειγμα, οι τιμές παρουσιάζουν κανονικότητες στις διακυμάνσεις τους στη διάρκεια του χρόνου, ο εντοπισμός αυτών των διακυμάνσεων μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Το συνηθέστερο παράδειγμα χρονοσειρών είναι ο δείκτης τιμών του χρηματιστήριου. Η ανάλυση χρονοσειρών έχει μελετηθεί σε μεγάλο βαθμό στα πλαίσια του κλάδου της Οικονομετρίας. Σε μια επιχείρηση, αλλά και σε μια χώρα, υπάρχει μεγάλος αριθμός ποσοτήτων που εξελίσσονται χρονικά. Η παρακολούθηση της χρονικής διακύμανσης των πωλήσεων μιας επιχείρησης είναι παράδειγμα μιας τέτοιας ποσότητας. Άλλα παραδείγματα είναι η διακύμανση του ύψους αποθεμάτων σε μια αποθήκη και η διακύμανση του ύψους των συναλλαγών σε μια τράπεζα.

6.5 Η ΕΔ στη σύγχρονη επιχείρηση

Όπως έχει αναφερθεί και προηγουμένως, η Εξόρυξη Δεδομένων είναι ένας νέος επιστημονικός κλάδος, ο οποίος στόχο έχει την ανακάλυψη γνώσης από τα δεδομένα. Για τη σύγχρονη επιχείρηση, η γνώση αποτελεί πολύτιμο κεφάλαιο γιατί δίνει τη δυνατότητα μείωσης του ρίσκου κατά τη λήψη αποφάσεων, καθώς και τη δυνατότητα καλύτερης ανταπόκρισης στις προκλήσεις των αγορών. Η ορθή χρήση της γνώσης παράγει απτή επιχειρηματική αξία. Η διαχείριση της γνώσης συνίσταται στη διαδικασία απόκτησης, δημιουργίας, σύνθεσης, διασποράς και χρήσης της πληροφορίας, της εφευρετικότητας και της εμπειρίας για την επίτευξη επιχειρηματικών στόχων. Όλες οι επιχειρήσεις, σε μικρότερο ή μεγαλύτερο βαθμό, είναι υποχρεωμένες να διαχειριστούν τη γνώση, και οι επιχειρήσεις που θα το επιτύχουν σε μεγαλύτερο βαθμό, θα αποκομίσουν τα περισσότερα οφέλη. Σε αυτά περιλαμβάνονται η αύξηση της κερδοφορίας, η μείωση του κόστους, η βελτίωση της αποτελεσματικότητας, η ταχεία ανταπόκριση και προσαρμογή και τέλος, η αύξηση της καινοτομίας. Είναι προφανές ότι σε ένα περιβάλλον όπου η γνώση έχει βαρύνουσα σημασία, η Εξόρυξη Δεδομένων αποτελεί ένα πολύτιμο επιχειρηματικό εργαλείο.

Υπάρχουν διάφοροι τρόποι να ταξινομήσει κανείς και να μελετήσει την εφαρμογή της ΕΔ στις σύγχρονες επιχειρήσεις. Ένας τρόπος είναι να τη συναρτήσει με τον τύπο της επιχείρησης (πχ εμπόριο, τράπεζες, τηλεπικοινωνίες κλπ.) και να καταγράψει τις εφαρμογές της ΕΔ για την αντιμετώπιση των ιδιαίτερων απαιτήσεων του κάθε επιχειρηματικού κλάδου. Ένας άλλος τρόπος είναι, να μελετήσει το αντικείμενο σε σχέση με τις μεθόδους ΕΔ που χρησιμοποιούνται (πχ, εξόρυξη από βάσεις δεδομένων, εξόρυξη κειμένου, ανάλυση δικτύων κλπ.). Ένα πιο γενικό μοντέλο μελετά την εφαρμογή της ΕΔ σε σχέση με τις επιχειρηματικές διαδικασίες.

Η αρχιτεκτονική μιας επιχείρησης μπορεί να μελετηθεί μέσα από τον καθορισμό των επιχειρηματικών διαδικασιών της. Σύμφωνα με τους Davenport and Short (1990), οι **επιχειρηματικές διαδικασίες** (business processes) είναι ένα σύνολο λογικά συνδεδεμένων καθηκόντων, τα οποία εκτελούνται για την επίτευξη ενός καθορισμένου επιχειρηματικού αποτελέσματος. Ο καθορισμός και η διαχείριση των επιχειρηματικών διαδικασιών έχει αποτελέσει αντικείμενο μελέτης ερευνητών και έχουν προταθεί διάφορα μοντέλα ταξινόμησης. Ένα δημοφιλές μοντέλο είναι το Process Classification Framework (PCF) της APQC (APQC, n.d.). Το PCF είναι ένα υψηλού επιπέδου και ανεξάρτητο από τον τύπο της εκάστοτε επιχείρησης μοντέλο. Προτάθηκε το 1992 και από τότε εξελίσσεται συνεχώς. Το 2008 η APQC, σε συνεργασία με την IBM, το εξειδίκευσε για συγκεκριμένους επιχειρηματικούς κλάδους. Το PCF ορίζει δώδεκα βασικές κατηγορίες επιχειρηματικών διαδικασιών. Οι κατηγορίες αυτές παρουσιάζονται στον πίνακα 6.1

Process Classification Framework	
Operating Processes	Vision & Strategy
	Products & Services
	Market & Selling
	Delivery
	Customer Service
Management & Support Process	Human Capital
	Information Technology
	Financial Resources
	Property Managemet
	Environmental Issues
	External Relationships
	Knowledge Management

Πίνακας 6.1 *Process Classification Framework της APQC*

Οι 12 βασικές κατηγορίες αναλύονται περαιτέρω σε πολλές υποκατηγορίες. Ο ενδιαφερόμενος αναγνώστης μπορεί να αναζητήσει το πλήρες μοντέλο στην ιστοθέση της APQC.

Οι επιχειρηματικές διαδικασίες μπορούν να συσχετιστούν με την ΕΔ υπό τρείς διαφορετικές έννοιες:

- Εφαρμογή τεχνικών Εξόρυξης Δεδομένων για την εξυπηρέτηση των στόχων των επιχειρηματικών διαδικασιών. Οι επιχειρηματικές διαδικασίες αντιμετωπίζονται ως πεδία εφαρμογής της ΕΔ. Από τις γενικές κατηγορίες, μπορεί να διαπιστώσει κανείς ότι η ΕΔ βρίσκει πεδία εφαρμογής στις περισσότερες από αυτές, όπως στη διαχείριση γνώσης, στη διαφήμιση και τις πωλήσεις, στα προϊόντα και τις υπηρεσίες, στην εξυπηρέτηση πελατών, στα χρηματοοικονομικά κλπ.
- Μοντέλα και τρόποι ενσωμάτωσης της Εξόρυξης Δεδομένων στις επιχειρηματικές διαδικασίες. Μελετάται ο τρόπος ολοκλήρωσης των επιχειρηματικών διαδικασιών με μεθοδολογίες Εξόρυξης Δεδομένων. Με τον τρόπο αυτό, οι επιχειρηματικές διαδικασίες εμπλουτίζονται με προηγούμενη γνώση και μετασχηματίζονται. Ερευνητές έχουν προτείνει σχετικά μοντέλα.
- Εξόρυξη των επιχειρηματικών διαδικασιών με χρήση τεχνικών Εξόρυξης Δεδομένων. Στοχεύει στην ανακάλυψη των επιχειρηματικών διαδικασιών, οι οποίες εφαρμόζονται στην επιχείρηση, με χρήση τεχνικών Εξόρυξης Δεδομένων και με ανάλυση των στοιχείων που τηρούνται στα αρχεία συμβάντων (log files) των πληροφοριακών συστημάτων.

Από τα παραπάνω γίνεται κατανοητό, ότι ο τρόπος προσέγγισης της σχέσης της Εξόρυξης Δεδομένων με τη σύγχρονη επιχείρηση είναι πολυεπίπεδος. Φυσικά, η έκταση του αντικείμενου είναι πολύ μεγάλη και δεν μπορεί να καλυφθεί στα πλαίσια ενός κεφαλαίου. Δικός μας στόχος είναι να εισάγουμε τον αναγνώστη στα βασικά πεδία και ζητήματα εφαρμογής της Εξόρυξης Δεδομένων στη σύγχρονη επιχείρηση και να τον προσανατολίσουμε σε ενδεχόμενες μελλοντικές και πιο εξειδικευμένες αναζητήσεις του.

Στην τρέχουσα κατάσταση πραγμάτων, βασικά πεδία εφαρμογής της ΕΔ είναι οι πωλήσεις και η διαφήμιση, οι τράπεζες, ο ασφαλιστικός τομέας και οι τηλεπικοινωνίες.

6.5.1 Πωλήσεις και Διαφήμιση

Οι Πωλήσεις και η Διαφήμιση είναι ίσως το πιο δημοφιλές πεδίο εφαρμογής της Εξόρυξης Δεδομένων στις επιχειρήσεις. Κάθε επιχείρηση επιδιώκει να αυξήσει τον όγκο των πωλήσεων της. Για την επίτευξη αυτού του στόχου εφαρμόζονται διάφορες τακτικές. Η πρώτη είναι η ενδυνάμωση της σχέσης του πελάτη με την επιχείρηση. Ο βαθμός προσήλωσης ενός πελάτη σε μια συγκεκριμένη εμπορική επωνυμία ονομάζεται **πίστη**. Ο πιστός πελάτης έχει λιγότερες πιθανότητες να μετακινηθεί σε μια ανταγωνίστρια εταιρεία. Βασικός παράγοντας αύξησης της πίστης του πελάτη είναι το υψηλό επίπεδο εξυπηρέτησης του και κατ' επέκταση ο βαθμός ικανοποίησης του. Η συγκράτηση της τρέχουσας πελατείας και η ελαχιστοποίηση **απώλειας πελατών** (customer churn) αποτελεί βασικό μέλημα κάθε επιχείρησης. Μια άλλη τακτική αύξησης των πωλήσεων είναι οι λεγόμενες **διασταυρούμενες πωλήσει**ς (cross selling). Διασταυρούμενες πωλήσεις είναι η πώληση επιπλέον προϊόντων ή υπηρεσιών στους ήδη υπάρχοντες πελάτες. Υπάρχουν προϊόντα που συνδυάζονται σχε-

τικά εύκολα με κάποια άλλα. Για παράδειγμα, ο ιδιοκτήτης ενός τηλεσκοπίου μπορεί να αγοράσει διαφόρων ειδών βάσεις στήριξης, προσοφθάλμιους φακούς μεγέθυνσης σε ποικίλες κλίμακες, φίλτρα προστασίας από την ακτινοβολία, ειδικές κάμερες για αστροφωτογράφηση, λογισμικό επεξεργασίας εικόνας κλπ. Σε άλλες περιπτώσεις όμως, οι συνδυασμοί δεν είναι τόσο προφανείς. Για παράδειγμα, είναι προς διερεύνηση ποιοι από τους κατόχους αποταμιευτικών λογαριασμών σε μια τράπεζα είναι διατεθειμένοι να επενδύσουν σε αμοιβαία κεφάλαια που διαχειρίζεται η τράπεζα. Οι διασταυρούμενες πωλήσεις είναι μια μέθοδος πωλήσεων αρκετά δύσκολη και περίπλοκη. Για την επιτυχία της απαιτείται ο εντοπισμός πραγματικών και σύνθετων αναγκών των πελατών, ώστε η επιχείρηση να μπορεί να προσφέρει κατάλληλα πακέτα προϊόντων ή υπηρεσιών και να ακολουθήσει μια δελεαστική τιμολογιακή πολιτική. Ένας τρίτος τρόπος αύξησης των πωλήσεων είναι η **προσέλκυση νέων πελατών**. Στην περίπτωση αυτή, καθοριστικό ρόλο παίζει η **διαφήμιση**. Επίσης, οι επιχειρήσεις προσπαθούν να εντείνουν τη σχέση και την αλληλεπίδραση τους με τους πελάτες, καθιερώνοντας **πολλαπλούς διαύλους επικοινωνίας** με αυτούς.

Για την αύξηση των πωλήσεων βασικό ρόλο παίζει η **αναγνώριση της καταναλωτικής συμπεριφοράς** των πελατών, η κατανόηση δηλαδή του τι, πότε και πως αγοράζουν οι πελάτες. Μια μέθοδος Εξόρυξης Δεδομένων που βρίσκει ευθεία εφαρμογή στην αναγνώριση της καταναλωτικής συμπεριφοράς, είναι οι Κανόνες Συσχέτισης. Οι Κανόνες Συσχέτισης εντοπίζουν συνδυασμούς προϊόντων που πωλούνται συχνά μαζί. Η εφαρμογή τους στα στοιχεία πωλήσεων σούπερ μάρκετ έχει αποδώσει εντυπωσιακά αποτελέσματα. Σχεδόν σε κάθε βιβλίο Εξόρυξης Δεδομένων αναφέρεται το παράδειγμα, όπου βρέθηκε ότι πωλούνται συχνά μαζί μπύρες και πάνες υγιεινής για βρέφη. Η εξήγηση του φαινομένου είναι αρκετά απλή. Άρρενες γονείς, όταν επισκέπτονται το σούπερ μάρκετ για τις οικογενειακές αγορές, αγοράζουν ταυτόχρονα και τις αγαπημένες τους μπύρες. Το ερώτημα είναι ποιος ειδικός πωλήσεων θα μπορούσε να φανταστεί, χωρίς τη χρήση εξελιγμένων τεχνικών, έναν τέτοιο συνδυασμό προϊόντων.

Με τη βοήθεια των Κανόνων Συσχέτισης, ο διευθυντής ενός υποκαταστήματος μπορεί να γνωρίζει συνδυασμούς προϊόντων που πωλούνται συχνά μαζί. Επιπλέον, μπορεί να γνωρίζει σε τι ποσοστό επί των συνολικών πωλήσεων εμφανίζεται ο συνδυασμός των προϊόντων, την πιθανότητα με την οποία η εμφάνιση ενός προϊόντος συνεπάγεται την εμφάνιση του άλλου προϊόντος, καθώς και το εάν η πώληση του ενός προϊόντος προκρίνει την πώληση του άλλου προϊόντος. Οι πληροφορίες αυτές είναι πολλαπλώς αξιοποιήσιμες. Καταρχήν, μπορεί να χρησιμοποιηθούν για να καθοριστεί το μίγμα των προϊόντων που διατίθενται στο υποκατάστημα. Επιπλέον, η πληροφορία αυτή αξιοποιείται για τον κατάλληλο σχεδιασμό των ραφιών και την τοποθέτηση των προϊόντων σε αυτά. Ο καταναλωτής, ο οποίος θα βρει σε γειτονικά ράφια προϊόντα που επιθυμεί, ενθαρρύνεται να προβεί σε πρόσθετες αγορές. Η επιχείρηση μπορεί να βελτιώσει περαιτέρω τις προσφορές της, προτείνοντας πακέτα προϊόντων, για τα οποία ισχύουν ειδικές εκπτώσεις. Η μέθοδος αυτή είναι ιδιαίτερα κατάλληλη για την ταχεία πώληση ευπαθών προϊόντων με κοντινή ημερομηνία λήξεως. Αν για παράδειγμα, βρεθεί ότι ένας τύπος κρασιών συνδυάζεται καλά και πωλείται συχνά με έναν συγκεκριμένο τύπο φρούτων, τότε μπορεί να γίνει μια ειδική προσφορά για τον συγκεκριμένο συνδυασμό. Ο διευθυντής μπορεί να επιλέξει να διαθέσει τα ευπαθή φρούτα σε τιμή κόστους ή και χαμηλότερα, όταν πωλούνται μαζί με το κρασί. Με τον τρόπο αυτό, θα διαθέσει πολύ γρήγορα τα φρούτα, τα οποία θα καταστρέφονταν και ταυτόχρονα, θα αποκομίσει κέρδος από την πώληση των κρασιών.

Το μοντέλο ανάλυσης που παρουσιάστηκε στην προηγούμενη παράγραφο είναι πολύ αποτελεσματικό, αλλά είναι απρόσωπο. Αναδεικνύει πρότυπα αγοράς και καταναλωτικής συμπεριφοράς και συμβάλλει στη βελτίωση της εξυπηρέτησης των πελατών, δεν αξιοποιεί όμως τα ιδιαίτερα ατομικά χαρακτηριστικά του κάθε πελάτη. Ωστόσο, η σύγχρονη επιχείρηση, χάρη στην εκτεταμένη μηχανοργάνωση, διατηρεί στοιχεία για τους πελάτες της και μπορεί να τα χρησιμοποιήσει για τη βελτίωση των διαφημιστικών πρακτικών και την προώθηση των πωλήσεων. Μια προσφιλής τακτική του σύγχρονου μάρκετινγκ είναι η λεγόμενη **τμηματοποίηση της αγοράς** (market segmentation). Η τμηματοποίηση της αγοράς συνίσταται στον επιμερισμό του καταναλωτικού κοινού σε ομάδες με ομοειδή χαρακτηριστικά. Τα μέλη μιας ομάδας παρουσιάζουν ομοιότητες σε σχέση με τα υποκειμενικά χαρακτηριστικά τους, όπως η οικονομική κατάσταση, το εργασιακό καθεστώς, η οικογενειακή κατάσταση, το μορφωτικό επίπεδο, η ηλικία, το φύλο, ο τόπος κατοικίας και βεβαίως η καταναλωτική συμπεριφορά. Γνωρίζοντας τη σύνθεση του καταναλωτικού κοινού και τα χαρακτηριστικά της κάθε ομάδας, η επιχείρηση μπορεί να υλοποιήσει μια στρατηγική σχεδιασμού και διάθεσης προϊόντων και υπηρεσιών, η οποία θα εξυπηρετεί τις ιδιαίτερες ανάγκες της κάθε ομάδας. Επιπλέον, μπορεί να οργανώσει διαφημιστικές εκστρατείες, οι οποίες θα απευθύνονται στοχευμένα σε επιλεγμένες ομάδες. Η τακτική αυτή ονομάζεται **στοχευμένη διαφήμιση** (target marketing).

Τα στοιχεία των πελατών τηρούνται στις βάσεις δεδομένων της επιχείρησης. Με την εφαρμογή μεθόδων Εξόρυξης Δεδομένων, όπως οι μέθοδοι Ανάλυσης Συστάδων, επιτυγχάνεται ο διαμοιρασμός των πελατών σε

τμήματα και ο καθορισμός των ομάδων. Επίσης, με χρήση μεθόδων κατηγοριοποίησης μπορούν να δημιουργηθούν προγνωστικά μοντέλα ικανά να εκτιμήσουν εάν ένας πελάτης θα ανταποκριθεί θετικά σε μια διαφημιστική καμπάνια. Το μοντέλο μπορεί να έχει ως έξοδο μια δυαδική μεταβλητή της μορφής Ναι/Όχι ή να υπολογίζει μια πιθανότητα θετικής ανταπόκρισης του πελάτη, πχ 80%. Ο όρος **database marketing** περιγράφει την ανάλυση βάσεων δεδομένων με στοιχεία πελατών, με στόχο την τμηματοποίηση της αγοράς και την ανάπτυξη προγνωστικών μοντέλων, τα οποία επιλέγουν υποψήφιους πελάτες με στοχευμένο τρόπο. Η σημασία της αξιοποίησης των στοιχείων των πελατών είναι τόσο μεγάλη, ώστε ορισμένοι οργανισμοί ανανεώνουν τα προφίλ των πελατών, ακόμα και σε ημερήσια βάση, με εργασίες δεσμών (batch) κατά τη διάρκεια της νύχτας.

Η στοχευμένη προσέγγιση των πελατών είναι μια πιο αποδοτική τακτική. Ας θεωρήσουμε το παράδειγμα της διαφήμισης με ταχυδρομική αποστολή διαφημιστικών φυλλαδίων ή καταλόγων. Συνήθως, τα φυλλάδια είναι σχεδιασμένα έτσι ώστε να έχουν γενική απεύθυνση και αποστέλλονται στο σύνολο των πελατών ή ίσως και στο σύνολο του πληθυσμού μιας περιοχής. Αυτή η μέθοδος είναι αρκετά ακριβή και συνήθως παρουσιάζει μικρό βαθμό ανταπόκρισης. Χρησιμοποιώντας τεχνικές Εξόρυξης Δεδομένων, εντοπίζονται υποψήφιοι πελάτες οι οποίοι έχουν μεγάλη πιθανότητα ανταπόκρισης. Το μοντέλο αναλύει στοιχεία υπαρκτών πελατών και ιστορικά στοιχεία προηγούμενων διαφημιστικών εκστρατειών και επιλέγει τους πελάτες στους οποίους θα αποσταλεί η αλληλογραφία. Με τον τρόπο αυτό, συμπιέζεται το κόστος της διαφήμισης και αυξάνεται το ποσοστό ανταπόκρισης. Επιπλέον, η γνώση των χαρακτηριστικών του target group επιτρέπει την προσαρμογή του διαφημιστικού υλικού στις ανάγκες και τις αισθητικές προτιμήσεις του, αυξάνοντας περαιτέρω τον βαθμό ανταπόκρισης. Η στοχευμένη διαφήμιση έχει και ορισμένα μειονεκτήματα. Ο πελάτης μπορεί να θεωρήσει ότι μια απεύθυνση ειδικά προσαρμοσμένη σε αυτόν, προέκυψε από επεξεργασία των στοιχείων του και αυτό να το εκλάβει ως παραβίαση της ιδιωτικής ζωής του.

Όλα τα παραδείγματα που παρουσιάστηκαν παραπάνω ανήκουν στην κατηγορία του λεγόμενου **εξερχόμενου μάρκετινγκ** (outbound marketing). Στο εξερχόμενο μάρκετινγκ, η επιχείρηση, με δική της πρωτοβουλία, διακινεί το διαφημιστικό μήνυμα, πχ με δημοσιευμένες καταχωρήσεις, τηλεοπτική διαφήμιση, αφίσες, αλληλογραφία κλπ. Μια νέα τάση στη διαφήμιση είναι το **εισερχόμενο μάρκετινγκ** (inbound marketing). Το εισερχόμενο μάρκετινγκ γίνεται με πρωτοβουλία του πελάτη, όταν αυτός προσεγγίζει την επιχείρηση, όταν πχ επισκέπτεται την ιστοθέση της ή όταν πραγματοποιεί μια τηλεφωνική κλήση για κάποιο θέμα του. Με τη χρήση μεθόδων Εξόρυξης Δεδομένων, το εισερχόμενο μάρκετινγκ μπορεί να αποδώσει εντυπωσιακά αποτελέσματα. Κατά την επαφή του πελάτη με την επιχείρηση, αναγνωρίζονται τα χαρακτηριστικά του και ο πελάτης κατηγοριοποιείται. Στη συνέχεια, υπολογίζεται η πιθανότητα να ανταποκριθεί θετικά σε ένα διαφημιστικό μήνυμα ή σε μια προσφορά και ακολούθως, του προωθείται η πλέον κατάλληλη πληροφόρηση. Η όλη διαδικασία γίνεται σε πραγματικό χρόνο κατά τη διάρκεια της επαφής του πελάτη με την επιχείρηση. Με τον τρόπο αυτό, η πρωτόβουλη επίσκεψη του πελάτη μετατρέπεται σε προσωποποιημένη διαφήμιση και επιτυγχάνεται η καλύτερη εξυπηρέτηση των αναγκών του.

Μια άλλη σύγχρονη τακτική στη διαφήμιση είναι το λεγόμενο **καθοδηγούμενο από γεγονότα μάρκετινγκ** (ΚΓΜ) (event driven marketing ή triggered marketing). Σύμφωνα με το ΚΓΜ, η αποστολή διαφημιστικού μηνύματος ενεργοποιείται όταν διαπιστωθεί μια σημαντική μεταβολή στο προφίλ του χρήστη. Η μεταβολή αυτή συνεπάγεται αλλαγή στην καταναλωτική συμπεριφορά του πελάτη. Για παράδειγμα, η μεταβολή της οικογενειακής κατάστασης, από άγαμος σε έγγαμος, έχει προφανείς επιπτώσεις στην καταναλωτική του συμπεριφορά. Μόλις διαπιστωθεί η μεταβολή των χαρακτηριστικών, ενεργοποιούνται τα μοντέλα Εξόρυξης Δεδομένων και επαναπροσδιορίζουν την κατηγορία του πελάτη, τις νέες ανάγκες του και την πιθανότητα να ανταποκριθεί θετικά σε νέα διαφημιστικά μηνύματα και προσφορές. Με τον τρόπο αυτό, επιτυγχάνεται η ταχύτατη εξυπηρέτηση των νέων αναγκών του. Το ΚΓΜ συνδυάζεται πολύ αποτελεσματικά με το εισερχόμενο μάρκετινγκ.

Περισσότερες πληροφορίες σχετικά με την εφαρμογή μεθόδων Εξόρυξης Δεδομένων στις πωλήσεις και στη διαφήμιση μπορεί να βρει ο αναγνώστης στα βιβλία των Linoff and Berry (2011) και των Venkatesan, Farris and Wilcox (2014).

6.5.2 Ηλεκτρονικό Εμπόριο

Οι επιχειρήσεις Ηλεκτρονικού Εμπορίου έχουν ιδιαίτερα χαρακτηριστικά και γι' αυτό εξετάζονται ξεχωριστά. Η ιδιομορφία έγκειται στο ότι η αγοροπωλησία διεξάγεται ηλεκτρονικά. Το γεγονός αυτό έχει σαν συνέπεια την παραγωγή και καταγραφή μεγάλων ποσοτήτων δεδομένων, όπως πχ προηγούμενες αγορές πελατών. Εκτός όμως από τέτοιες πληροφορίες, οι οποίες θα μπορούσαν να καταγραφούν και σε συστήματα συμβατικού εμπορίου, στην περίπτωση των ηλεκτρονικών καταστημάτων υπάρχει η δυνατότητα καταγραφής πληροφοριών που αφορούν τη χρήση της ιστοθέσης. Τέτοιες πληροφορίες είναι η διαδρομή περιήγησης του χρήστη,

το ρεύμα των κλικ, προηγούμενες αναζητήσεις κλπ. Η ανάλυση αυτών των στοιχείων αποφέρει πρόσθετες πληροφορίες σχετικά με το προφίλ του χρήστη, την καταναλωτική του συμπεριφορά και τις προτιμήσεις του, ακόμα και σε εξατομικευμένο επίπεδο λεπτομέρειας. Η χρήση των στοιχείων αυτών δεν διαφέρει από την αντίστοιχη στις επιχειρήσεις συμβατικού εμπορίου. Επιδίωξη είναι η εξειδίκευση, ακόμα και σε βαθμό εξατομίκευσης, της διαφήμισης και οι διασταυρούμενες πωλήσεις. Μια επιχείρηση δημιουργεί σχέσεις με τους πελάτες της όταν παρατηρεί τις ανάγκες τους και θυμάται τις προτιμήσεις τους, καθώς και την καταναλωτική συμπεριφορά τους.

Οι διασταυρούμενες πωλήσεις είναι μια πολύ διαδεδομένη τακτική στο ηλεκτρονικό εμπόριο και συνήθως γίνονται με τη μορφή συστάσεων. Οι επισκέπτες των ηλεκτρονικών καταστημάτων πολύ συχνά τροφοδοτούνται με μηνύματα του τύπου «οι πελάτες που αγόρασαν το τάδε προϊόν αγόρασαν επίσης τα παρακάτω προϊόντα». Πίσω από τα μηνύματα αυτά κρύβονται τεχνικές Εξόρυξης Δεδομένων. Ταυτόχρονα με την περιήγηση του χρήστη, διεξάγεται ανάλυση Κανόνων Συσχέτισης και εντοπίζονται προϊόντα, που πουλήθηκαν ταυτόχρονα με αυτά που έχει επιλέξει ο χρήστης. Επίσης, αλγόριθμοι k-πλησιέστερων γειτόνων εντοπίζουν άλλα προϊόντα με παρόμοια χαρακτηριστικά. Ο αναγνώστης μπορεί να βρει λεπτομέρειες σχετικά με αλγορίθμους για τη διατύπωση συστάσεων σε ιστοθέσεις ηλεκτρονικού εμπορίου στο Sarwar, Karypis, Konstan and Riedl (2000).

Η Εξόρυξη Δεδομένων βρίσκει ενδιαφέρουσες εφαρμογές και στους ηλεκτρονικούς πλειστηριασμούς. Εφαρμόζοντας ανάλυση τάσεων μπορεί να προβλεφθεί το πλήθος και η αξία των προϊόντων που θα διατεθούν στον πλειστηριασμό, καθώς και το πλήθος των συμμετεχόντων. Επίσης, στους ηλεκτρονικούς πλειστηριασμούς, εξαιτίας της ανωνυμίας και των χαλαρότερων νομικών περιορισμών, είναι αυξημένη η πιθανότητα απάτης. Με τη χρήση τεχνικών όπως τα Δένδρα Αποφάσεων, είναι δυνατόν να αντιμετωπιστεί σε ορισμένο βαθμό αυτό το φαινόμενο. Ένα επιπλέον έμμεσο όφελος, που μπορεί να προκύψει από την εφαρμογή της ΕΔ, είναι η βελτίωση του σχεδιασμού των ιστοσελίδων. Η επιχείρηση, γνωρίζοντας τα ενδιαφέροντα και τις προτιμήσεις των χρηστών, διαμορφώνει κατάλληλα τις ιστοσελίδες. Η διαμόρφωση των ιστοσελίδων είναι δυνατόν να φθάσει στον βαθμό της εξατομίκευσης και να προσαρμόζεται στις ανάγκες και επιθυμίες του εκάστοτε ξεχωριστού χρήστη.

6.5.3 Τράπεζες

Οι Τράπεζες είναι οργανισμοί κατεξοχήν κατάλληλοι για την εφαρμογή μεθόδων Εξόρυξης Δεδομένων. Ένας από τους λόγους είναι ότι διατηρούν αναλυτικά δεδομένα για όλες στις συναλλαγές των πελατών τους. Επίσης, σύμφωνα με τις νέες κανονιστικές διατάξεις, που διέπουν τη λειτουργία τους μετά την πρόσφατη οικονομική κρίση, είναι υποχρεωμένες να διατηρούν αναλυτικά στοιχεία των πελατών τους. Τα στοιχεία αυτά τηρούνται κυρίως για την αντιμετώπιση περιπτώσεων ξεπλύματος χρήματος (money laundering). Η ύπαρξη ποιοτικών δεδομένων και ο μεγάλος όγκος τους καθιστούν ιδανικές τις μεθόδους ΕΔ για τη διεξαγωγή αναλύσεων.

Η εφαρμογή της ΕΔ στα πλαίσια ενός τραπεζικού οργανισμού ενδείκνυται για την αντιμετώπιση τεσσάρων προβλημάτων:

- την προώθηση πωλήσεων και διαφήμιση,
- τη διαχείριση του ρίσκου,
- την απάτη πιστωτικών καρτών,
- το ξέπλυμα χρήματος.

Προώθηση πωλήσεων και διαφήμιση. Η προώθηση πωλήσεων και η διαφήμιση στα πλαίσια ενός τραπεζικού οργανισμού δεν διαφέρει ριζικά από τις αντίστοιχες εργασίες σε άλλες επιχειρήσεις. Τα σχετικά ζητήματα παρουσιάστηκαν αναλυτικά στο υποκεφάλαιο που αναφέρεται στις πωλήσεις και τη διαφήμιση. Στο σημείο αυτό να αναφέρουμε ότι και οι τράπεζες αναλύουν τα στοιχεία των πελατών τους, δημιουργούν τα προφίλ τους και διεξάγουν αναλύσεις τμηματοποίησης της αγοράς. Επιδίωξη τους είναι οι διασταυρούμενες πωλήσεις και η στοχευμένη διαφήμιση. Φυσικά, οι εργασίες αυτές είναι προσαρμοσμένες στο ιδιαίτερο αντικείμενο των τραπεζών. Οι διασταυρούμενες πωλήσεις σε μια τράπεζα αφορούν την ταυτόχρονη πώληση προϊόντων όπως δάνεια, πιστωτικές κάρτες, e-banking κλπ. Μια ιδιαιτερότητα των τραπεζών είναι ότι εξαιτίας της οικονομικής κρίσης, αλλά και του σκληρού ανταγωνισμού, έχουν υποχρεωθεί να ενισχύσουν την κεφαλαιακή τους βάση. Για τον λόγο αυτό, ενδιαφέρονται ιδιαίτερα να προσελκύσουν πελάτες που θα έχουν μια μακροχρόνια και σταθερή σχέση με την τράπεζα, αγοράζοντας προϊόντα όπως μακροχρόνιους προθεσμιακούς λογαριασμούς ή διάφορα επενδυτικά σχήματα. Η προσέλευση τέτοιων πελατών επιτυγχάνεται προνομιακά με τη διεξαγωγή στοχευμένου μάρκετινγκ.

Διαχείριση ρίσκου. Η διαχείριση του ρίσκου είναι η καρδιά των τραπεζικών εργασιών. Η φύση των τραπεζικών εργασιών είναι τέτοια που εμπεριέχει το στοιχείο του κινδύνου. Οι τράπεζες, χρησιμοποιώντας εξειδικευμένες μεθοδολογίες, αναγνωρίζουν, αναλύουν, μετρούν και ελέγχουν το ρίσκο που αναλαμβάνουν, έτσι ώστε το συνολικό επίπεδο του να μην θέτει σε κίνδυνο την ασφάλεια και την ομαλή λειτουργία της τράπεζας. Οι τραπεζικοί οργανισμοί αντιμετωπίζουν διαφόρων ειδών κινδύνους. Σε αυτούς περιλαμβάνονται ο πιστωτικός κίνδυνος, ο κίνδυνος ελλιπούς ρευστότητας, ο κίνδυνος από διακυμάνσεις ισοτιμιών νομισμάτων, ο κίνδυνος από μεταβολές επιτοκίων και ο κίνδυνος διαταραχής της φήμης του οργανισμού. Βαρύνουσα είναι η σημασία του πιστωτικού κινδύνου. Ο πιστωτικός κίνδυνος (credit risk) είναι η πιθανότητα να αποτύχει ο δανειολήπτης να αποπληρώσει το δάνειο σύμφωνα με τους συμφωνημένους όρους. Το πρόβλημα συνίσταται στο γεγονός ότι οι τράπεζες πρέπει να επιδιώκουν και να ενθαρρύνουν τη χορήγηση όσων περισσότερων δανείων μπορούν, ενώ ταυτόχρονα πρέπει να αποφεύγουν τη χορήγηση επισφαλών δανείων. Για τον λόγο αυτό, η σωστή εκτίμηση του πιστοληπτικού κινδύνου καθίσταται κομβικής σημασίας. Επισημαίνεται ότι η πρόσφατη οικονομική κρίση πυροδοτήθηκε από την κατάρρευση της αγοράς στεγαστικών δανείων στις ΗΠΑ και τη συνακόλουθη απαξίωση σύνθετων προϊόντων, όπως των περιβόητων CDOs (Collateralized Debt Obligations). Το αποτέλεσμα ήταν η χρεοκοπία εκατοντάδων μικρότερων τραπεζών και η επιλεκτική διάσωση μεγαλύτερων. Το παράδειγμα αυτό καταδεικνύει τις επιπτώσεις που μπορεί να έχει η ανεπαρκής διαχείριση του ρίσκου.

Η εφαρμογή μεθόδων Εξόρυξης Δεδομένων, ειδικά εκείνων που προέρχονται από τον χώρο της Μηχανικής Μάθησης και είναι ικανές να αυτοεκπαιδεύονται, βασιζόμενες σε ιστορικά στοιχεία, μπορεί να προσφέρει μέγιστες υπηρεσίες στη βελτίωση των τρεχουσών πρακτικών εκτίμησης πιστοληπτικού κινδύνου. Τα ιστορικά στοιχεία που αναλύονται αφορούν τον δανειολήπτη (ηλικία, επάγγελμα, υγεία, οικονομικές υποχρεώσεις, προηγούμενη πιστωτική συμπεριφορά και αποπληρωμή δανείων κλπ.), τις ιδιαιτερότητες του δανείου (ποιότητα υποθήκης, επιτόκια, χρόνος αποπληρωμής), καθώς και μακροοικονομικά στοιχεία. Η ερευνητική βιβλιογραφία περιλαμβάνει πάμπολλες μελέτες που ασχολούνται με την εφαρμογή τεχνικών ΕΔ για την εκτίμηση του πιστοληπτικού κινδύνου. Ενδεικτικά αναφέρουμε τις Chen, Xiang, Liu and Wang (2012), Bekhet and Eletter (2014), Zhang, Gao and Shi (2014) και Oreski and Oreski (2014).

Απάτη Πιστωτικών Καρτών. Στη σημερινή εποχή, η εκτεταμένη χρήση του ηλεκτρονικού χρήματος έχει επιφέρει την ευρεία διάδοση των πιστωτικών καρτών. Δυστυχώς, ταυτόχρονα παρουσιάστηκαν και αντίστοιχες πρακτικές εξαπάτησης. Οι πρακτικές αυτές στηρίζονται είτε στη φυσική κλοπή της κάρτας είτε στην υποκλοπή των στοιχείων της. Οι τράπεζες ενδιαφέρονται να εντοπίσουν τις συναλλαγές με κλεμμένες πιστωτικές κάρτες για να προστατέψουν τους πελάτες τους. Μέθοδοι ΕΔ, όπως η Ανάλυση Συστάδων και η Ανάλυση Εξαιρέσεων, χρησιμοποιούνται για τον εντοπισμό των παράνομων καρτών. Καταγράφονται πρότυπα συναλλαγών που πραγματοποιούν οι πελάτες και όταν διαπιστωθεί απόκλιση από τις συνήθεις πρακτικές, τότε ελέγχεται η νομιμότητα της κάρτας.

Ξέπλυμα Χρήματος. Όπως ήδη αναφέρθηκε, σύμφωνα με την πρόσφατη νομοθεσία, οι τράπεζες οφείλουν να ελέγχουν τις συναλλαγές των πελατών τους για τον εντοπισμό περιπτώσεων ξεπλύματος χρήματος, νομιμοποίησης δηλαδή χρήματος που προέρχεται από παράνομες δραστηριότητες. Το πρόβλημα είναι σημαντικό γιατί σχετίζεται με το εμπόριο ναρκωτικών και με τη χρηματοδότηση τρομοκρατικών ενεργειών. Ο εντοπισμός περιπτώσεων ξεπλύματος χρήματος μπορεί να γίνει με δύο τρόπους. Ο ένας είναι να ελεγχθούν οι προηγούμενες συναλλαγές του πελάτη για να διαπιστωθεί εάν υπάρχουν αποκλίσεις από τα συνηθισμένα πρότυπα. Ο δεύτερος είναι να ελεγχθεί το δίκτυο με το οποίο συναλλάσσεται ο ύποπτος λογαριασμός. Μπορούν να χρησιμοποιηθούν μέθοδοι ανάλυσης συστάδων, κατηγοριοποίησης, ανάλυσης εξαιρέσεων, καθώς και μέθοδοι ανάλυσης δικτύων (Gao & Ye, 2007).

6.5.4 Ασφάλειες

Ο κλάδος των ασφαλειών είναι ένας ακόμα επιχειρηματικός τομέας, στον οποίο η Εξόρυξη Δεδομένων έχει βρει πεδία εφαρμογής. Οι ασφαλιστικές εταιρείες μπορούν να μειώσουν το κόστος, να αυξήσουν τα κέρδη, να αποκτήσουν καινούργιους πελάτες, να διατηρήσουν τους υπάρχοντες πελάτες, να αναπτύξουν νέα προϊόντα και κυρίως να διαχειριστούν το ρίσκο, το οποίο αποτελεί και το κεντρικό ζήτημα του κλάδου. Μελέτες έχουν δείξει ότι οι πελάτες που έχουν συνάψει περισσότερα συμβόλαια είναι πιο πιστοί. Εφαρμόζοντας Κανόνες Συσχέτισης, οι ασφαλιστικές εταιρείες βρίσκουν προϊόντα που πωλούνται συχνά μαζί και τα προσφέρουν ως ομαδοποιημένα πακέτα (πχ ασφάλεια ζωής, σπιτιού και αυτοκινήτου) σε προνομιακές τιμές. Με τον τρόπο αυτό, αυξάνουν την αξία των προσφορών τους καθώς και την ικανοποίηση και την προσήλωση των πελατών. Με τη βοήθεια τεχνικών Ανάλυσης Συστάδων, εντοπίζουν ομάδες ανασφάλιστων υποψήφιων πελατών, εφαρμόζουν στοχευμένη διαφήμιση και προτείνουν τα κατάλληλα προϊόντα στην εκάστοτε ομάδα. Για παράδειγμα,

προτείνουν ασφάλειες ζωής σε νέους πελάτες και επενδυτικά προγράμματα σε ώριμους επιχειρηματίες. Τα στοιχεία που αναλύουν αφορούν την ηλικία, την οικονομική κατάσταση, το μορφωτικό επίπεδο κλπ.

Η διαχείριση του κινδύνου είναι η ουσία των ασφαλιστικών εργασιών. Οι ασφαλιστικές εταιρείες εκτιμούν την πιθανότητα να διεκδικήσουν οι πελάτες τους αποζημιώσεις, καθώς και το ύψος των αποζημιώσεων, σχεδιάζουν αναλογιστικά μοντέλα και καθορίζουν το ύψος των ασφαλίστρων. Τα μοντέλα στηρίζονται στα υποκειμενικά στοιχεία του πελάτη και ποσοτικοποιούν τον κίνδυνο ανάλογα με τα χαρακτηριστικά του. Είναι προς διερεύνηση ποια από τα χαρακτηριστικά είναι σημαντικά και πως το κάθε ένα από αυτά επηρεάζει τον συνολικό κίνδυνο. Για παράδειγμα, ένα μοντέλο πρέπει να υπολογίζει την πιθανότητα να εμπλακεί σε αυτοκινητιστικό δυστύχημα ένας νέος και άπειρος οδηγός, καθώς και την πιθανότητα για έναν έμπειρο αλλά ηλικιωμένο οδηγό. Προγνωστικά μοντέλα κατηγοριοποίησης μπορούν να προβλέψουν με μεγάλη επιτυχία τον βαθμό του κινδύνου και επιτρέπουν στις ασφαλιστικές εταιρείες να καθορίσουν με μεγαλύτερη ακρίβεια το ύψος των ασφαλίστρων και να επιτύχουν έτσι ανταγωνιστικό πλεονέκτημα.

Οι ασφαλιστικές εταιρείες χάνουν κάθε χρόνο εκατομμύρια εξαιτίας δόλιων και απατηλών διεκδικήσεων από πελάτες. Επιπλέον κόστος επιφέρει η απασχόληση προσωπικού επιφορτισμένου με την αντιμετώπιση της απάτης. Μέθοδοι Εξόρυξης Δεδομένων, όπως η Ανάλυση Συστάδων, η Ανάλυση Εξαιρέσεων και η Κατηγοριοποίηση, χρησιμοποιούνται για την αντιμετώπιση της απάτης στις ασφάλειες. Παραδείγματα εφαρμογής συγκεκριμένων μεθόδων Εξόρυξης Δεδομένων για ασφαλιστικά θέματα υπάρχουν στο Devale and Kulkarni (2012).

6.5.5 Χρηματιστήριο

Η πρόβλεψη της διακύμανσης των τιμών των μετοχών και των δεικτών στα χρηματιστήρια αποτελεί κάτι σαν το ιερό δισκοπότηρο για τους επενδυτές. Πολυάριθμοι ερευνητές έχουν ασχοληθεί διεξοδικά με το θέμα. Φυσικά οι δυνατότητες πρόβλεψης που παρέχουν οι μέθοδοι της Εξόρυξης Δεδομένων έχουν αξιοποιηθεί στο έπακρο. Είναι χαρακτηριστικό ότι μια αναζήτηση στη Scopus με λέξεις κλειδιά data mining και stock market, αποδίδει περισσότερες από 60 δημοσιευμένες εργασίες για κάθε χρόνο. Μέθοδοι Ανάλυσης Τάσεων, Κατηγοριοποίησης και Ανάλυσης Συστάδων, καθώς και ευφάνταστες υβριδικές τεχνικές που συνδυάζουν βασικές μεθόδους με διάφορους τρόπους, έχουν επανειλημμένα εφαρμοστεί με μεγαλύτερη η μικρότερη επιτυχία σύμφωνα με τα δημοσιευμένα αποτελέσματα. Ο ενδιαφερόμενος αναγνώστης μπορεί εύκολα να βρει μεγάλο πλήθος σχετικών εργασιών. Ενδεικτικά αναφέρουμε τις εργασίες των Barak and Modarres (2015) και των Hu, Feng, Zhang, Ngai and Liu (2015).

6.5.6 Τηλεπικοινωνίες

Οι επιχειρήσεις που δραστηριοποιούνται στον χώρο των τηλεπικοινωνιών είναι από τους κορυφαίους οργανισμούς στη χρήση τεχνολογιών πληροφορικής. Ως αποτέλεσμα αυτού του γεγονότος, καταγράφουν τεράστιους όγκους δεδομένων. Τα δεδομένα αυτά είναι μια πολύτιμη και πολλαπλώς αξιοποιήσιμη πηγή πληροφοριών και οι μέθοδοι Εξόρυξης Δεδομένων είναι το εργαλείο για την αξιοποίηση τους.

Οι εταιρείες τηλεπικοινωνιών καταγράφουν δεδομένα που εντάσσονται σε τρεις κατηγορίες:

- Στοιχεία πελατών. Σε αυτά περιλαμβάνονται ατομικά στοιχεία όπως όνομα, διεύθυνση, οικογενειακά στοιχεία, τύπος συνδρομής και στοιχεία για το ιστορικό πληρωμών.
- Στοιχεία κλήσεων. Καταγράφονται ο καλών και ο καλούμενος συνδρομητής, η ημερομηνία και η χρονική στιγμή της κλήσης, η διάρκεια της κλήσης και το κόστος της.
- Τεχνικά στοιχεία λειτουργίας των τηλεπικοινωνιακών δικτύων. Τα τηλεπικοινωνιακά δίκτυα συγκροτούνται από συστατικά μέρη υλικού και λογισμικού. Τα μέρη αυτά παράγουν αυτόματα τεχνικά δεδομένα, τα οποία αφορούν τη λειτουργία του δικτύου και την εμφάνιση βλαβών και προβλημάτων.

Οι εταιρείες τηλεπικοινωνιών χρησιμοποιούν αυτά τα δεδομένα για στις παρακάτω εργασίες:

- πωλήσεις και διαφήμιση,
- αντιμετώπιση απάτης,
- αντιμετώπιση προβλημάτων δικτύου.

Σε ότι αφορά τις πωλήσεις και τη διαφήμιση, γίνεται τμηματοποίηση της αγοράς, καθορισμός προφίλ πελάτη, στοχευμένη διαφήμιση και διασταυρούμενες πωλήσεις. Οι εργασίες αυτές προσαρμόζονται στις ιδιαίτερες

απαιτήσεις και δυνατότητες της συγκεκριμένης αγοράς. Σημειωτέον ότι ο ανταγωνισμός στον συγκεκριμένο χώρο είναι πολύ έντονος εξαιτίας της εμπορευματοποίησης του εύρους ζώνης, της εισόδου νέων ανταγωνιστών, των διαδοχικών εξαγορών και συγχωνεύσεων και της απαίτησης για ακριβές τεχνικές υποδομές. Για τον λόγο αυτό, η βελτιστοποίηση των τεχνικών μάρκετινγκ για τη συγκράτηση των τρεχόντων πελατών και την προσέλκυση νέων είναι απαραίτητη. Η ανάλυση των στοιχείων κλήσεων, η γνώση δηλαδή του ποιος καλεί ποιόν και για πόσο χρόνο, μπορεί να αναδείξει πρότυπα κλήσεων, τα οποία θα χρησιμοποιηθούν για τον σχεδιασμό κατάλληλων συνδρομητικών υπηρεσιών, που να ταιριάζουν στις ανάγκες των χρηστών, καθώς και για τη διαμόρφωση προσφορών και εκπτωτικών πακέτων. Το γεγονός ότι η επιχείρηση κατέχει και ελέγχει έναν μόνιμο δίαυλο επικοινωνίας με τον πελάτη, καθιστά εφικτή την προώθηση κάθε είδους διαφημιστικών μηνυμάτων με προνομιακούς όρους και χωρίς χρέωση.

Η απάτη είναι ένα σοβαρό και καθόλου σπάνιο πρόβλημα για τις εταιρείες τηλεπικοινωνιών. Η απάτη στις τηλεπικοινωνίες είναι κάθε δραστηριότητα με την οποία ο δράστης αποκτά πρόσβαση σε υπηρεσίες τηλεπικοινωνιών, χωρίς να έχει πρόθεση να τις πληρώσει. Σύμφωνα με τον Weiss (2005), υπάρχουν δύο ειδών απάτες, η απάτη συνδρομής και η απάτη υπέρθεσης. Απάτη συνδρομής υπάρχει όταν ο δράστης ανοίγει έναν λογαριασμό χωρίς να σκοπεύει να πληρώσει. Η απάτη αυτού του είδους δεν αφορά μόνο ιδιώτες, οι οποίοι σκοπεύουν να χρησιμοποιήσουν τις υπηρεσίες και να μην καταβάλλουν το αντίτιμο, αλλά και σε οργανωμένες απάτες που στοχεύουν να αποσπάσουν χρηματικά ποσά από την εταιρεία τηλεπικοινωνιών. Για παράδειγμα, μια μεγάλη εταιρεία τηλεπικοινωνιών συνάπτει συμβόλαιο με άλλη εταιρεία, η οποία θα παρέχει υπηρεσίες, όπως πχ τηλεφωνική διαφήμιση, υπηρεσίες ενηλίκων κλπ. (οι γνωστές γραμμές 0800…). Η δεύτερη εταιρεία πραγματοποιεί πολλές πλαστές κλήσεις υψηλού κόστους με εικονικούς πελάτες και με αυτοματοποιημένο τρόπο, πχ με χρήση υπολογιστών και εισπράττει το αντίτιμο από την πρώτη εταιρεία. Η πρώτη εταιρεία, όταν αναζητήσει τους πελάτες που πραγματοποίησαν τις κλήσεις, θα διαπιστώσει ότι είναι ανύπαρκτοι και ότι έχει πέσει θύμα απάτης. Απάτη υπέρθεσης υπάρχει όταν ο δράστης αποκτά πρόσβαση στον λογαριασμό κάποιου άλλου νόμιμου πελάτη και πραγματοποιεί κλήσεις, τις οποίες χρεώνει στον πελάτη – θύμα. Η απάτη αυτού του είδους δεν προκαλεί, τουλάχιστον άμεσα, οικονομική βλάβη στην εταιρεία τηλεπικοινωνιών, προκαλεί όμως την εντονότατη δυσαρέσκεια των πελατών θυμάτων και μπορεί να οδηγήσει σε απώλεια πελατών. Οι εταιρείες τηλεπικοινωνιών χρησιμοποιούν μεθόδους Εξόρυξης Δεδομένων για να αναλύσουν τα στοιχεία των συνδρομητών και τα στοιχεία των κλήσεων και να εντοπίσουν κλήσεις, που αποκλίνουν από τα πρότυπα των φυσιολογικών συνδιαλέξεων και που σηματοδοτούν υψηλή πιθανότητα διεξαγωγής απάτης. Ένα σύστημα εντοπισμού απάτης τηλεπικοινωνιών με χρήση έμπειρου συστήματος προτείνεται στο Hilas (2009).

Η ανάλυση των στοιχείων λειτουργίας των δικτύων είναι ένα επιπλέον πεδίο εφαρμογής μεθόδων Εξόρυξης Δεδομένων. Τα δεδομένα αυτά παράγονται με αυτόματο τρόπο από διάφορα υποσυστήματα του δικτύου και ο όγκος τους είναι τέτοιος, που καθιστά αδύνατη την επεξεργασία τους χωρίς τη χρήση εξελιγμένων τεχνικών. Μια συγκεκριμένη βλάβη στο δίκτυο μπορεί να προκαλέσει πολλά και διαφορετικά μηνύματα σφάλματος. Με την εφαρμογή τεχνικών κατηγοριοποίησης και ανάλυσης αλληλουχίας μπορεί να γίνει συσχετισμός των μηνυμάτων και εντοπισμός της βλάβης. Τα στοιχεία λειτουργίας του δικτύου μπορούν να χρησιμοποιηθούν και για την ανάλυση της ποιότητας των υπηρεσιών. Περισσότερες λεπτομέρειες σχετικά με τη χρήση μεθόδων Εξόρυξης Δεδομένων στις τηλεπικοινωνίες υπάρχουν στο Madhouri (2013) και στο Kabakchieva (2009)

6.5.7 Λογιστική - Ελεγκτική

Η σύγχρονη εποχή θέτει επιτακτικά νέα και πιο απαιτητικά καθήκοντα στους ελεγκτικούς μηχανισμούς των επιχειρήσεων. Στο παρελθόν δεν ήταν λίγα τα παραδείγματα αποτυχίας. Χαρακτηριστική είναι η περίπτωση του σκανδάλου Enron, το οποίο εκτός των άλλων προκάλεσε και την κατάρρευση της ελεγκτικής εταιρείας Arthur Andersen, ενός από τους πυλώνες του παγκόσμιου ελεγκτικού συστήματος και μέλους της λεγόμενης ομάδας των πέντε μεγάλων ελεγκτών (Big 5). Δυστυχώς οι αποτυχίες δεν περιορίστηκαν εκεί. Η πρόσφατη οικονομική κρίση πυροδοτήθηκε από την κατάρρευση αμερικανικών τραπεζικών κολοσσών, όπως η Fannie Mae και η Freddie Mac, και οι ελεγκτικοί μηχανισμοί απέτυχαν να προβλέψουν και να εμποδίσουν αυτό το φαινόμενο. Οι αρμόδιοι κυβερνητικοί και άλλοι φορείς, στην προσπάθεια τους να αντιμετωπίσουν τέτοια φαινόμενα, θεσπίζουν νέα κανονιστικά πλαίσια, τα οποία διέπουν την εταιρική διακυβέρνηση και ορίζουν ελεγκτικές διαδικασίες.

Το έργο των εξωτερικών ελεγκτών είναι ιδιαίτερα δύσκολο, καθώς καλούνται να λάβουν μη δομημένες αποφάσεις σε συνθήκες υψηλού βαθμού αβεβαιότητας. Το έργο τους καθίσταται ακόμα δυσκολότερο σε περιπτώσεις όπου τα διοικητικά στελέχη των επιχειρήσεων εμπλέκονται σε καταχρηστικές πρακτικές. Τα στελέχη αυτά διαθέτουν την κατάλληλη εμπειρία, αλλά και το κίνητρο, ώστε να παρέμβουν και να αποπροσανατο-

λίσουν την ελεγκτική διαδικασία. Οι πολλαπλές δυσκολίες, αλλά και η μεγάλη σημασία του αντικειμένου, καθιστούν τη διαρκή αναβάθμιση των ελεγκτικών πρακτικών αδήριτη ανάγκη. Σε αυτήν την προσπάθεια, η συμβολή των μεθοδολογιών της Εξόρυξης Δεδομένων μπορεί να αποδειχθεί αποφασιστικής σημασίας.

Δύο μεγάλα προβλήματα της Ελεγκτικής, στα οποία βρίσκουν εφαρμογή οι τεχνικές Εξόρυξης Δεδομένων και ειδικότερα οι τεχνικές κατηγοριοποίησης, είναι η πρόβλεψη χρεοκοπίας και ο εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων. Η **πρόβλεψη χρεοκοπίας** είναι ένα από τα σημαντικότερα προβλήματα λήψης αποφάσεων. Εκτός από την Ελεγκτική σχετίζεται και με τραπεζικά ζητήματα, όπως την εκτίμηση πιστοληπτικού κινδύνου. Οι χρεοκοπίες επιχειρήσεων επιφέρουν μεγάλες οικονομικές ζημίες σε επενδυτές και πιστωτές, ενώ σε ακραίες περιπτώσεις μπορούν να επηρεάσουν ολόκληρες κοινωνίες ή και το παγκόσμιο οικονομικό σύστημα. Εξαιτίας της σημασίας των επιπτώσεων, οι εξωτερικοί ελεγκτές είναι υποχρεωμένοι να διατυπώσουν την άποψη τους σχετικά με την ικανότητα να συνεχίσει τις δραστηριότητες της για ένα ουσιαστικό χρονικό διάστημα μετά τη δημοσίευση των οικονομικών της εκθέσεων (σχόλια τύπου going concern). Η υποχρέωση αυτή των ελεγκτών ορίζεται με σαφήνεια στα λεγόμενα Ελεγκτικά Πρότυπα (Statement of Auditing Standards - SAS) και ειδικότερα στα SAS 59, 64, 77 και 96. Η ακαδημαϊκή κοινότητα, θεωρώντας ότι η χρεοκοπία είναι ένα φαινόμενο που εξελίσσεται στη διάρκεια του χρόνου και όχι ένα στιγμιαίο συμβάν, συμβάλλει με τη διατύπωση μοντέλων ικανών να προβλέψουν έγκαιρα τις περιπτώσεις χρεοκοπίας (early warning predictors). Η σχετική έρευνα ξεκίνησε ήδη από τη δεκαετία του ’60 με τις εργασίες των Beaver και Altman. Στη σημερινή εποχή πλήθος ερευνητών έχουν δημοσιεύσει εργασίες στις οποίες χρησιμοποιούν Νευρωνικά Δίκτυα, Μηχανές Διανυσμάτων Υποστήριξης, Δένδρα Αποφάσεων και άλλους κατηγοριοποιητές για την πρόβλεψη της χρεοκοπίας και τα αποτελέσματα είναι πολύ ικανοποιητικά.

Οι πρακτικές «μαγειρέματος των βιβλίων» (book cooking practices) είναι ένα μάλλον διαδεδομένο και σε καμία περίπτωση … γεωγραφικά περιορισμένο φαινόμενο. Ο Wells (1997) εκτιμά ότι η απάτη κοστίζει στην αμερικανική οικονομία 400 δισεκατομμύρια δολάρια ετησίως, ενώ ο Koskivaara (2004) αποκαλεί το έτος 2002 «φριχτή χρονιά» ως προς την τήρηση των βιβλίων και ισχυρίζεται ότι η χειραγώγηση συνεχίζεται. Ο διαχωρισμός της ιδιοκτησίας από τη διοίκηση στις σύγχρονες μεγάλες επιχειρήσεις δημιουργεί κίνητρα στα διοικητικά στελέχη να δράσουν προς ίδιον όφελος. Η **παραποίηση των χρηματοοικονομικών καταστάσεων** είναι μια σημαντική πρακτική διοικητικής απάτης και η αντιμετώπιση της εντάσσεται στα καθήκοντα των εξωτερικών ελεγκτών. Ειδικότερα, σύμφωνα με το ελεγκτικό πρότυπο 82 (Statement of Auditing Standards 82 – SAS82) επιβάλλει στους εξωτερικούς ελεγκτές να εκτιμήσουν τον κίνδυνο απάτης κατά τη διάρκεια των ελέγχων. Η ανάλυση των χρηματοοικονομικών καταστάσεων με χρήση μεθόδων Εξόρυξης Δεδομένων έχει αποδώσει μοντέλα ικανά να εντοπίζουν τις περιπτώσεις απάτης. Ενδεικτικά αναφέρουμε τις εργασίες των Fanning and Cogger (1998) και των Kirkos, Spathis and Manolopoulos (2007).

6.5.8 Εξόρυξη Κειμένου

Όλα τα παραδείγματα και οι εφαρμογές που παρουσιάστηκαν μέχρι τώρα αφορούσαν την ανάλυση δεδομένων, τα οποία προέρχονται από σχεσιακές βάσεις και είναι αριθμητικά ή ονομαστικά. Τα δεδομένα αυτά είναι γνωστά ως δομημένα δεδομένα. Τον τελευταίο καιρό προσελκύουν το ενδιαφέρον τεχνικές που αφορούν την επεξεργασία αδόμητων δεδομένων, όπως πχ κειμένου. Ακαδημαϊκοί ερευνητές και πάροχοι λογισμικού συντονίζουν τις ενέργειες τους και προτείνουν μεθόδους ικανές να αναλύσουν κείμενα, να εξάγουν χρήσιμα συμπεράσματα και να ολοκληρώσουν τα αποτελέσματα τους με αυτά άλλων μεθόδων που επεξεργάζονται δομημένα δεδομένα. Η τάση αυτή οφείλεται στο γεγονός ότι τα αδόμητα δεδομένα είναι πολύ περισσότερα. Αναφέρεται ότι ο όγκος τους αντιστοιχεί στο 80% του συνολικού όγκου των δεδομένων. Επίσης, η σύγχρονη πληροφορική και το Web 2.0 παράγουν συνεχώς αδόμητα δεδομένα. Τα δεδομένα αυτά περιέχουν πολύτιμη πληροφορία και η ανάκτηση της μπορεί να αποδειχθεί εξαιρετικά χρήσιμη.

Οι επιχειρήσεις αξιοποιούν τις τεχνικές Εξόρυξης Κειμένου με διάφορους τρόπους. Μια πιθανή εφαρμογή είναι η δρομολόγηση κλήσεων. Οι μεγάλες επιχειρήσεις δέχονται καθημερινά χιλιάδες emails από πελάτες που διατυπώνουν παράπονα, υποβάλλουν διάφορα αιτήματα, ζητούν οδηγίες χρήσης και τεχνική υποστήριξη κλπ. Η χειροκίνητη δρομολόγηση των μηνυμάτων στον κατάλληλο αποδέκτη είναι εξαιρετικά αργή. Σε πολλές περιπτώσεις, η ανάγνωση του θέματος δεν είναι αρκετή και χρειάζεται να γίνει ανάγνωση του σώματος του μηνύματος. Με τη χρήση εργαλείων Εξόρυξης Κειμένου, γίνεται κατανοητό το περιεχόμενο και το μήνυμα δρομολογείται στον αποδέκτη με αυτόματο τρόπο.

Ένα άλλο πεδίο εφαρμογής είναι η αντιμετώπιση της απάτης. Η εγκατάσταση συστημάτων ανάλυσης κειμένων σε βασικά κανάλια γραπτής επικοινωνίας, όπως είναι το email, μέσα σε οργανισμούς, συμβάλλει στην αντιμετώπιση περιστατικών απάτης και προστατεύει τους εργαζόμενους. Επίσης, ασφαλιστικές εταιρείες μπο-

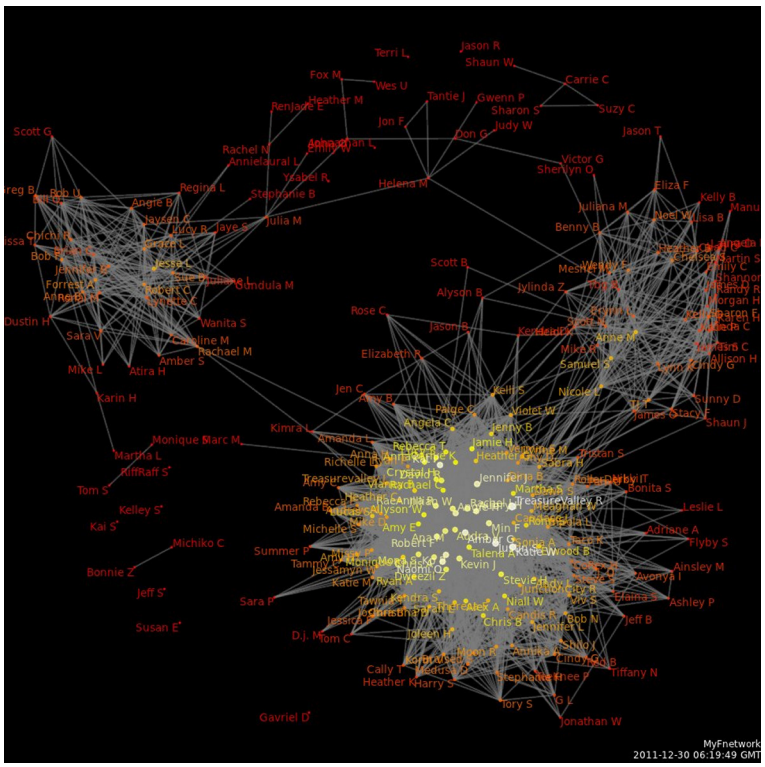
ρούν να χρησιμοποιήσουν τέτοιο λογισμικό για να αντιμετωπίσουν περιπτώσεις ψευδών αιτήσεων για αποζημιώσεις. Τα λογισμικά αυτά αναλύουν εκθέσεις εκτιμητών, ιατρικές γνωματεύσεις και άλλα ντοκουμέντα και αποφαίνονται σχετικά με την ειλικρίνεια του αιτήματος.

Μια δημοφιλής εργασία Εξόρυξης Κειμένου είναι η λεγόμενη **Ανάλυση Συναισθήματος** (Sentiment Analysis). Η ανάλυση συναισθήματος χρησιμοποιεί τεχνικές Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing), Στατιστικής και Μηχανικής Μάθησης για να αναγνωρίσει το συναισθηματικό περιεχόμενο ενός κειμένου, να αποφανθεί δηλαδή εάν ο συντάκτης είναι θετικά η αρνητικά διακείμενος σε μια άποψη, ένα πρόσωπο, μια κατάσταση κλπ. Η αποκωδικοποίηση του πραγματικού περιεχομένου ενός κειμένου δεν είναι πάντα εύκολη και η λεξικογραφική ανάλυση δεν επαρκεί. Οι άνθρωποι εκφράζονται με σύνθετο τρόπο. Ρητορικά σχήματα, ιδιωματισμοί, ειρωνικές εκφράσεις μπορούν να αποπροσανατολίσουν τις μεθόδους. Ένα μήνυμα της μορφής «Είμαι ενθουσιασμένος με το προϊόν του ανταγωνιστή σας. Διαθέτει επιδόσεις, εμφάνιση και αξιοπιστία. Συγχαρητήρια!!!» είναι πολύ εύκολο να παρερμηνευθεί.

Οι επιχειρήσεις χρησιμοποιούν την ανάλυση συναισθήματος για να αποκτήσουν πληροφορίες, όπως ποια είναι γνώμη του κοινού για την εταιρεία και τα προϊόντα της, ποια χαρακτηριστικά τους αρέσουν και ποια όχι, τι νομίζουν για την ποιότητα, την αξιοπιστία και την τιμή των προϊόντων, τι εντύπωση έκανε μια διαφημιστική εκστρατεία κλπ. Τα δεδομένα διατίθενται άφθονα στο Διαδίκτυο. Οι άνθρωποι εκφράζουν τη γνώμη τους με κριτικές προϊόντων σε ιστοθέσεις όπως το Amazon, σε tweets, σε blogs, σε ιστοτόπους κοινωνικής δικτύωσης και σε emails. Το υλικό αυτό είναι τεράστιο σε όγκο, καταγράφει απόψεις εκατομμυρίων ή και δισεκατομμυρίων ανθρώπων και ανανεώνεται καθημερινά. Η αξιοποίηση του υλικού αυτού δίνει πρωτόγνωρες δυνατότητες. Για παράδειγμα, ήταν σχετικά εύκολο για μια επιχείρηση να μάθει τις απόψεις των πελατών της, αλλά ήταν εξαιρετικά δύσκολο να μάθει τις απόψεις των μη πελατών της. Με τις τεχνικές ανάλυσης συναισθήματος και με δεδομένα του Διαδικτύου αυτό είναι τώρα εφικτό. Με τη χρήση των τεχνικών Εξόρυξης Κειμένου είναι δυνατή η **ανάλυση τάσεων,** η αναγνώριση δηλαδή της μεταβολής των απόψεων ανθρώπων με την πάροδο του χρόνου. Οι επιχειρήσεις μπορούν να γνωρίζουν τι μεταβολές επέφεραν στις απόψεις των ανθρώπων γεγονότα, όπως το λανσάρισμα στην αγορά ενός ανταγωνιστικού προϊόντος ή η διεξαγωγή μιας διαφημιστικής εκστρατείας.

6.5.9 Ανάλυση Κοινωνικών Δικτύων

Τα τελευταία χρόνια μεγάλη διάδοση γνωρίζει και η Ανάλυση Κοινωνικών Δικτύων. Η ανάλυση κοινωνικών δικτύων χρησιμοποιεί τη θεωρία δικτύων για να αναλύσει κοινωνικές σχέσεις. Χρησιμοποιείται ευρύτατα στη σύγχρονη κοινωνιολογία, αλλά σχετίζεται και με την ανθρωπολογία, την κοινωνική ψυχολογία, τη βιολογία και άλλους επιστημονικούς κλάδους. Μεγάλη ώθηση στην ανάλυση κοινωνικών δικτύων έδωσε η ραγδαία διάδοση των ιστοτόπων κοινωνικής δικτύωσης, όπως το Facebook, και των ιστοτόπων διαμοιρασμού μέσων, όπως το FlickR.



Σχήμα 6.4 *Σχέσεις φιλίας στο Facebook. (Αναπαραγωγή από Wikimedia Commons. Ιδιοκτήτης Kenneth Freeman/ Kencf0618)*

Ένα κοινωνικό δίκτυο, ένα σύνολο δηλαδή ανθρώπων οι οποίοι έχουν σχέσεις μεταξύ τους, αναπαρίσταται ως ένας γράφος που αποτελείται από κόμβους και συνδέσεις. Ένας κόμβος αντιστοιχεί σε ένα άτομο, ενώ οι συνδέσεις δηλώνουν την ύπαρξη σχέσης μεταξύ των ατόμων. Οι σχέσεις μεταξύ των ατόμων προκύπτουν από αλληλεπιδράσεις μεταξύ τους. Σε ιστοτόπους κοινωνικής δικτύωσης, σχέσεις μεταξύ χρηστών δηλώνονται από τους ίδιους τους χρήστες με τον καθορισμό «φίλων». Σχέσεις όμως προκύπτουν και όταν ένας χρήστης εντάσσεται σε μια ομάδα, κάνει like σε απόψεις ή follow σε άλλον χρήστη, μοιράζεται εικόνες ή άλλα μέσα κλπ. Άδηλες σχέσεις μπορούν να εξαχθούν και από την ομοιότητα των χρηστών. Χρήστες που χρησιμοποιούν συχνά τα ίδια tags ή έχουν άλλα κοινά στοιχεία συμπεριφοράς μπορούν να θεωρηθούν ως «όμοιοι» και να συνδεθούν με συνδέσεις.

Παλαιότερα η μελέτη κοινωνικών δικτύων περιοριζόταν σε μικρές ομάδες, γιατί ήταν αδύνατη η συλλογή περισσότερων δεδομένων. Η εκρηκτική όμως διάδοση ιστοτόπων κοινωνικής δικτύωσης, όπως το Facebook και το LinkedIn, άλλαξε ριζικά τις συνθήκες. Πλέον είναι διαθέσιμος ένας τεράστιος όγκος δεδομένων, που αφορά υπαρκτούς ανθρώπους και τις σχέσεις τους. Τα δεδομένα αυτά, με τη μορφή γράφων, είναι διαφορετικού τύπου από τα δεδομένα που τηρούνται στις παραδοσιακές σχεσιακές βάσεις δεδομένων, και η επεξεργασία τους συνιστά μια πρόκληση, καθώς μπορούν να αποτελέσουν πηγή χρήσιμης και διαφορετικής πληροφορίας. Τα δεδομένα των κοινωνικών δικτύων αξιοποιούνται από διάφορους φορείς και βεβαίως μπορούν να αξιοποιηθούν από τις επιχειρήσεις.

Μια πληροφορία που μπορεί να εξαχθεί από ένα κοινωνικό δίκτυο είναι η αξιολόγηση του κάθε κόμβου σχετικά με την αξιοπιστία του. Η αξιολόγηση μπορεί να είναι ολική και να προκύπτει από ολόκληρο το δίκτυο ή τοπική και να προκύπτει από τις αξιολογήσεις άλλων κόμβων. Η εκτίμηση της αξιοπιστίας του κάθε κόμβου έχει επιχειρηματικές εφαρμογές. Οι πάροχοι υπηρεσιών ηλεκτρονικών πλειστηριασμών υπολογίζουν και δημοσιοποιούν σκορ αξιοπιστίας για τους πωλητές. Οι αγοραστές προτιμούν τους πωλητές με καλή αξιοπιστία και είναι διατεθειμένοι να πληρώσουν ακριβότερο αντίτιμο. Η εκτίμηση της αξιοπιστίας χρησιμοποιείται και από πωλητές υπηρεσιών και προϊόντων για την αντιμετώπιση της απάτης. Επιχειρήσεις εντοπίζουν πελάτες χαμηλής αξιοπιστίας, οι οποίοι σκοπεύουν να προβούν σε αγορές χωρίς να καταβάλλουν το αντίτιμο. Τέλος, η αξιοπιστία του ατόμου μέσα σε ένα δίκτυο χρησιμοποιείται και για τον εντοπισμό «ειδικών». Επιχειρήσεις που αναζητούν στελέχη εξειδικευμένα σε ένα αντικείμενο για συνεργασία ή πρόσληψη, διευκολύνονται στην αναζήτηση τους, αξιοποιώντας τα κοινωνικά δίκτυα και την αξιοπιστία.

Η δομή του δικτύου είναι μια πολύτιμη πηγή πληροφοριών, καθώς ο αναλυτής μπορεί να εντοπίσει «κοινότητες» χρηστών. Ο καθορισμός των κοινοτήτων βασίζεται στην ύπαρξη πυκνών δεσμών μεταξύ των μελών

της κοινότητας και αραιότερων δεσμών με τα μη μέλη. Στο Σχήμα 6.4 είναι εύκολα διακριτές τρεις κοινότητες. Το πρόβλημα εύρεσης κοινοτήτων δεν είναι απλό και μοιάζει με πρόβλημα συσταδοποίησης γράφων. Σημειωτέον ότι οι κοινότητες είναι δυναμικές και μεταβάλλονται με τον χρόνο. Ο καθορισμός κοινοτήτων μέσω ενός δικτύου είναι ένας διαφορετικός τρόπος ομαδοποίησης των πελατών μιας επιχείρησης, δηλαδή τμηματοποίησης της αγοράς. Οι επιχειρήσεις χρησιμοποιούν αυτήν την πληροφορία για σχεδιασμό προϊόντων και υπηρεσιών κατάλληλων για μια ομάδα, για στοχευμένη διαφήμιση, για σχεδιασμό εκπτωτικών πακέτων κλπ. Σε συστήματα ηλεκτρονικού εμπορίου μπορεί να χρησιμοποιηθεί για τη διατύπωση συστάσεων. Σε μέλη μιας κοινότητας προτείνονται προϊόντα και υπηρεσίες που επέλεξαν άλλα μέλη της κοινότητας.

Πέρα από τη στατική δομή του δικτύου, ενδιαφέρον παρουσιάζει και η δυναμική ροή πληροφοριών μέσα σε αυτό. Οι σχέσεις μεταξύ των κόμβων αποτελούν κανάλια διάδοσης πληροφοριών και αλληλεπίδρασης. Ένα μήνυμα μεταδίδεται μέσα στο δίκτυο και επηρεάζει τα μέλη. Σε ότι αφορά τη διαφήμιση και τις πωλήσεις, οι χρήστες επηρεάζονται από τους φίλους τους και είναι αρκετά πιθανόν να αγοράσει κάποιος προϊόντα και υπηρεσίες που αγόρασαν φίλοι του. Οι φίλοι ανταλλάσσουν μεταξύ τους απόψεις και κάποιος μπορεί να συστήσει ένα προϊόν στους φίλους του. Σε περιπτώσεις ηλεκτρονικών συναλλαγών, υπάρχει η δυνατότητα να ενημερώνονται αυτόματα οι φίλοι για τις αγορές ενός ατόμου. Επίσης, στα δίκτυα τηλεφωνίας, λόγω προσφορών χαμηλότερων τιμών για κλήσεις εντός του δικτύου ενός παρόχου, είναι πολύ πιθανόν ένας καταναλωτής να επιλέξει το δίκτυο που χρησιμοποιούν άτομα με τα οποία σχετίζεται. Ορισμένοι κόμβοι, λόγω προσωπικής μόρφωσης, ενδιαφερόντων, έντονης δραστηριότητας στο δίκτυο κλπ. επηρεάζουν σε σημαντικότερο βαθμό άλλους κόμβους απ' ότι πιο αδρανείς κόμβοι. Ο εντοπισμός και η επιρροή των δυναμικών κόμβων επιτρέπει την ταχύτερη και αποτελεσματικότερη διάδοση ιδεών στο δίκτυο. Μια διαφημιστική εκστρατεία που στοχεύει στους δυναμικούς κόμβους, επιτυγχάνει την ταχεία διάδοση των μηνυμάτων της με μικρότερο κόστος.

Πέρα από τους ιστοτόπους κοινωνικής δικτύωσης, ένα πολύ ισχυρό κανάλι διάδοσης ιδεών είναι τα blogs. Άτομα που συμμετέχουν σε blogs δημιουργούν ένα δίκτυο και αλληλεπιδρούν μεταξύ τους ανταλλάσσοντας πληροφορίες και ιδέες. Δυναμικοί bloggers, οι οποίοι καταχωρούν συχνά κείμενα και συμμετέχουν σε πολλά blogs έχουν μεγαλύτερη επιρροή. Οι επιχειρήσεις, χρησιμοποιώντας πληροφορίες από τα δίκτυα, μπορούν να παρακολουθούν την πορεία της φήμης της επιχείρησης, να αντλούν πληροφορίες για παράπονα πελατών και για αρνητικά σχόλια σχετικά με τα προϊόντα τους, να εντοπίζουν δυσαρεστημένους πελάτες που είναι πιθανόν να εγκαταλείψουν την εταιρεία, αλλά και δυσαρεστημένους πελάτες ανταγωνιστών, να κατανοήσουν τους λόγους απώλειας πελατών, να αναγνωρίσουν καταναλωτικές τάσεις, που θα τους επιτρέψουν να σχεδιάσουν προϊόντα και να επενδύσουν σε υποδομές. Στο Bonchi, Castillo, Gionis and Jaimes (2011) γίνεται μια συνοπτική αλλά ουσιαστική παρουσίαση της χρήσης της ανάλυσης κοινωνικών δικτύων για επιχειρηματικούς σκοπούς.

6.5.10 Ενσωμάτωση Της Εξόρυξης Δεδομένων στις Επιχειρηματικές Διαδικασίες

Η Εξόρυξη Δεδομένων είναι ένας νέος επιστημονικός κλάδος, που έχει βρει πλήθος εφαρμογών στη σύγχρονη επιχείρηση, όπως έγινε εμφανές στα προηγούμενα υποκεφάλαια. Η ικανότητα των μεθόδων ΕΔ να μετατρέπουν τα δεδομένα σε πηγή πολύτιμων πληροφοριών και να παράγουν αξία για τον οργανισμό, τις καθιστά απαραίτητο εργαλείο για τη σύγχρονη επιχείρηση. Ένα ζήτημα που χρίζει περαιτέρω διερεύνησης είναι ο τρόπος ενσωμάτωσης των μεθόδων ΕΔ στις επιχειρηματικές διαδικασίες.

Η ΕΔ μπορεί να χρησιμοποιηθεί για αναλυτικές διοικητικές διαδικασίες λήψης αποφάσεων, αλλά και για καθημερινές εργασίες του λειτουργικού επιπέδου. Στη δεύτερη περίπτωση, οι μέθοδοι ΕΔ μετατρέπονται σε εργαλείο καθημερινής λειτουργίας και έτσι πολλαπλασιάζεται ο ρυθμός ανάκτησης πληροφορίας. Κατά κανόνα, ο εμπλουτισμός των επιχειρησιακών διαδικασιών με νέες τεχνολογίες προκαλεί τον ανασχεδιασμό τους, σε μικρότερο ή μεγαλύτερο βαθμό. Το ίδιο συμβαίνει και με την περίπτωση της ΕΔ. Η απλούστερη εκδοχή είναι να ενσωματωθεί η ΕΔ στις επιχειρησιακές διαδικασίες χωρίς ουσιαστική μεταβολή τους. Συνήθως όμως, η ανακάλυψη γνώσης θέτει ζητήματα των οποίων η επίλυση απαιτεί τον ευρύτερο ανασχεδιασμό των διαδικασιών. Για παράδειγμα, αν ανακαλυφθεί ότι ένας τρόπος πληρωμής παρουσιάζει κενά ασφαλείας και καθιστά δυνατή την απάτη, τότε αυτός ο τρόπος πληρωμής μεταβάλλεται. Σε περίπτωση ευρύτερου ανασχεδιασμού των επιχειρησιακών διαδικασιών προκύπτει μεγαλύτερο όφελος για την επιχείρηση, καθώς οι διαδικασίες της βελτιώνονται. Η πλέον ακραία εκδοχή, είναι η ανακάλυψη γνώσης να προκαλέσει τον στρατηγικό επαναπροσανατολισμό του οργανισμού, οπότε προκύπτει και το μεγαλύτερο όφελος. Η εισαγωγή της ΕΔ στις επιχειρηματικές διαδικασίες λειτουργεί και ως τρόπος διαμοιρασμού της πληροφορίας και γνώσης στο εσωτερικό του οργανισμού και συνήθως προκαλεί ανάγκες για νέους ρόλους, όπως για ειδικούς ΕΔ, για ειδικούς βάσεων δεδομένων και για ειδικούς στατιστικούς αναλυτές.

Σε ότι αφορά τον τρόπο εισαγωγής της ΕΔ στις επιχειρησιακές διαδικασίες, έχουν προταθεί σχετικά μοντέλα. Οι Rupnik and Jaklic (2009) ασχολούνται με την ενσωμάτωση της ΕΔ στις λειτουργικές διαδικασίες. Οι συγγραφείς προτείνουν ένα μεθοδολογικό πλαίσιο, στο πρώτο στάδιο του οποίου αξιολογούνται οι επιχειρηματικές διαδικασίες και οι εμπλεκόμενοι εργαζόμενοι, ως προς τον βαθμό ετοιμότητας τους να εφαρμόσουν την ΕΔ. Στο δεύτερο στάδιο γίνεται η εισαγωγή της ΕΔ στις επιχειρηματικές διαδικασίες και εκτελούνται δύο παράλληλες δραστηριότητες: α) ο σχεδιασμός της μεταβολής των διαδικασιών και β) η ανάλυση και ο σχεδιασμός της εφαρμογής. Επίσης, στο πλαίσιο του δεύτερου σταδίου γίνεται και η υλοποίηση της εφαρμογής και η ολοκλήρωση της με τις επιχειρηματικές διαδικασίες. Το τρίτο στάδιο περιλαμβάνει τη λειτουργία των ανασχεδιασμένων επιχειρηματικών διαδικασιών με τη χρήση της ΕΔ. Οι συγγραφείς επίσης παρουσιάζουν ένα πολύ αναλυτικό παράδειγμα εφαρμογής ΕΔ για direct marketing, μαζί με τη διαδικασία ενσωμάτωσης της ΕΔ.

6.5.11 Εξόρυξη Επιχειρηματικών Διαδικασιών

Μια ακόμα εφαρμογή της ΕΔ στις σύγχρονες επιχειρήσεις είναι η Εξόρυξη των Επιχειρηματικών Διαδικασιών (Business Process Mining (BPM)). Τα σύγχρονα επιχειρησιακά πληροφοριακά συστήματα, όπως τα ERP, καταγράφουν όλα τα συμβάντα και τις ενέργειες στα αρχεία συμβάντων (log files). Ειδικότερα, στα αρχεία συμβάντων καταγράφονται πληροφορίες σχετικά με μια δραστηριότητα και με μια περίπτωση. Η περίπτωση είναι ένα θέμα, το οποίο διαχειρίζεται το σύστημα, όπως πχ μια παραγγελία πελάτη, μια εντολή παραγωγής κλπ. Η δραστηριότητα είναι κάποια λειτουργία για την περίπτωση. Επίσης, για κάθε γεγονός καταγράφεται ο χρόνος που συμβαίνει. Επιπλέον, όταν παρεμβαίνουν άνθρωποι, καταγράφεται και το πρόσωπο που εκτελεί τη δραστηριότητα. Το Business Process Mining αναλύει τα Log files για να ανακαλύψει διαδικασίες, ελέγχους, δεδομένα και επιχειρησιακές και κοινωνικές δομές.

Στις σύγχρονες επιχειρήσεις, η βελτιστοποίηση των διαδικασιών αποτελεί μόνιμη επιδίωξη, ώστε να επιτευχθεί η μεγιστοποίηση της αποτελεσματικότητας και της αποδοτικότητας. Το Business Process Mining στοχεύει στην αυτόματη κατασκευή μοντέλων, τα οποία εξηγούν τη συμπεριφορά που παρατηρείται στα log files. Αναλύοντας τις δραστηριότητες των περιπτώσεων, κατασκευάζονται μοντέλα διαδικασιών. Το μοντέλο μπορεί να αναπαρασταθεί ως ένα δίκτυο Petri. Σε ότι αφορά τη μοντελοποίηση των διαδικασιών, εξετάζεται η ροή ελέγχου, δηλαδή η αλληλουχία των δραστηριοτήτων. Ο στόχος είναι να βρεθεί ένας χαρακτηρισμός για όλους τους δυνατούς δρόμους δράσης και να εκφραστεί με τη μορφή μοντέλου. Η εξόρυξη επιχειρηματικών διαδικασιών μπορεί επίσης να χρησιμοποιηθεί για την ανακάλυψη ρόλων των εργαζομένων και την τυποποίηση των σχέσεων μεταξύ τους. Ένα μοντέλο τέτοιου τύπου καταγράφει τη μεταφορά εργασίας από άτομο σε άτομο. Επίσης, μπορεί να ελεγχθεί η απόδοση των διαδικασιών, δηλαδή η χρονική εξέλιξη τους.

Βιβλιογραφία / Αναφορές

APQC. (n.d.) *Process Classification Framework*. Retrieved 24 September, 2015, from http://www.apqc.org/pcf.

Barak, S., & Modarres, M. (2015). Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems with Applications*, *42*(3), 1325-1339. doi: 10.1016/j.eswa.2014.09.026

Bekhet, H., & Eletter, S. (2014). Credit Risk Assessment Model for Jordanian Commercial Banks: Neural Scoring Approach. *Review of Development Finance*, *4*(1), 20-28. doi: 10.1016/j.rdf.2014.03.002

Bonchi, F., Castillo, C., Gionis, A., & Jaimes, A. (2011). Social Network Analysis and Mining for Business Applications. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 22-37. doi: 10.1145/1961189.1961194

Chen, W., Xiang, G., Liu, Y., & Wang, K. (2012). Credit Risk Evaluation by Hybrid Data Mining Technique, *Procedia Systems Engineering*, *3*, 194-200. doi: 10.1016/j.sepro.2011.10.029

Davenport, T. H., & Short, J. E. (1990). The new industrial engineering: information technology and business process redesign. *Sloan Management Review*, *31*(4), 11-27.

Devale, A. B., & Kulkarni, D. R. V. (2012). Application of Data Mining Techniques in Life Insurance. *International Journal of Data Mining and Knowledge Management Process*, 2(4), 31-40. doi: 10.5121/ijdkp.2012.2404

Du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. *Neurocomputing*, *73*(10-12), 2047–2060. doi: 10.1016/j.neucom.2009.11.034

Fanning, K., & Cogger, K. (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, *7*(1), 21–41. doi: 10.1002/(SICI)1099-1174(199803)7:1<21::AID-ISAF138>3.0.CO;2-K

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *Artificial Intelligence Magazine*, *17*(3), 37-54. doi: 10.1609/aimag.v17i3.1230

Gao, Z., & Ye, M. (2007). A Framework for Data Mining – Based Anti – Money Laundering Research. *Journal of Money Laundering Control*, *10*(2), 170-179. doi: 10.1108/13685200710746875

Han, J., & Kamber, M. (2001). *Data Mining Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Hilas, C. (2009). Designing and Expert System for Fraud Detection in Private Telecommunications Networks. *Expert Systems with Applications*, *36*(9), 11559-11569. doi: 10.1016/j.eswa.2009.03.031

Hu, Y., Feng, B., Zhang, X., Ngai, E., & Liu, M. (2015). Stock trading rule discovery with an evolutionary trend following model. *Expert Systems with Applications*, *42*(1), 212-222. doi: 10.1016/j.eswa.2014.07.059

Kabakchieva, D. (2009). Business Intelligence Applications and Data Mining Methods in Telecommunications: A Literature Review. *Bulgarian OpenAIRE Repository*. Retrieved 15 February, 2015, from http://hdl.handle.net/10867/44

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications*, *32*(4), 995-1003. doi: 10.1016/j.eswa.2006.02.016

Koskivaara, E. (2004). Artificial Neural Networks in Analytical Review Procedures. *Managerial Auditing Journal*, *19*(2), 191–223. doi: 10.1108/02686900410517821

Linoff, G., & Berry, M. (2011). *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management*. Indianapolis, IN: Wiley Publishing Inc.

Madhouri, V. (2013). Data Mining and Business Intelligence Applications in Telecommunication Industry. *International Journal of Engineering and Advanced Technology*, *2*(3), 525-528.

Maimon, O., & Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer Science + Business Media Inc.

Oreski, S., & Oreski, G. (2014). Genetic Algorithm-Based Heuristic for Feature Selection in Credit Risk Assessment. *Expert Systems with Applications*, *41*(4), 2052-2064. doi: 10.1016/j.eswa.2013.09.004

Rupnik, R., & Jaklic, J. (2009). The Deployment of Data Mining into Operational Business Process. In J. Ponce & A. Korahoca (Eds.), *Data Mining and Knowledge Discovery in Real Life Applications* (pp.

373-388). Vienna, Austria: I-Tech Education and Publishing. doi: 10.5772/6460

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Analysis of Recommendation Algorithms for E-Commerce. *Proceedings of the 2nd ACM Conference on Electronic Commerce*, 158-167. New York, NY: ACM. doi: 10.1145/352871.352887

Venkatesan, R., Farris, P., & Wilcox, R. T. (2014). *Cutting Edge Marketing Analytics: Real World Cases and Data Sets for Hands on Learning*. Upper Saddle River, NJ: Pearson Education Inc.

Weiss, G. (2005). *Data Mining in Telecommunications. In: O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers* (pp. 1189-1201). New York, NY: Springer Science + Business Media Inc.

Wells, J. T. (1997). *Occupational Fraud and Abuse*. Austin, TX: Obsidian Publishing.

Witten, I. H., & Frank, E. (2000). *Data Mining Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, CA: Morgan Kaufmann Publishers.

Www-01.ibm.com. (2015). IBM – *What is big data*?. Retrieved 18 February, 2015, from http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Zhang, Z., Gao, G., & Shi, Y. (2014). Credit Risk Evaluation Using Multi-Criteria Optimization Classifier with Kernel, Fuzzification and Penalty Factors. *European Journal of Operational Research*, 237(1), 335-348. doi: 10.1016/j.ejor.2014.01.044

7 Προεπεξεργασία Δεδομένων

Σύνοψη

Αντικείμενο του παρόντος Κεφαλαίου είναι η προεπεξεργασία των δεδομένων, οι εργασίες δηλαδή προετοιμασίας των δεδομένων, οι οποίες εκτελούνται πριν την καθαυτό εξόρυξη γνώσης. Η προεπεξεργασία των δεδομένων είναι απαραίτητη, καθώς τα αρχικά δεδομένα πάσχουν από διαφόρων ειδών προβλήματα. Σε αυτά συγκαταλέγονται η ύπαρξη αλληλοσυγκρουόμενων πληροφοριών, η ύπαρξη ασυνεπειών ως προς την κωδικοποίηση, την ονοματοδοσία πεδίων και τις μονάδες μέτρησης, καθώς και η ύπαρξη χαμένων τιμών και θορύβου, τυχαία δηλαδή κυμαινόμενων δεδομένων χωρίς ουσιαστικό περιεχόμενο. Τα προβληματικά αυτά δεδομένα καλούνται «ακάθαρτα» και η διαδικασία αντιμετώπισης των προβλημάτων τους καλείται «καθαρισμός δεδομένων». Η προεπεξεργασία των δεδομένων περιλαμβάνει τον καθαρισμό τους, αλλά δεν περιορίζεται σε αυτόν. Ειδικές απαιτήσεις των μεθόδων επεξεργασίας συχνά επιβάλλουν τον μετασχηματισμό των δεδομένων. Δύο συνήθεις εργασίες μετασχηματισμού είναι η διακριτοποίηση και η κανονικοποίηση. Ο όρος διακριτοποίηση αναφέρεται στον μετασχηματισμό αριθμητικών τιμών σε ονομαστικές τιμές. Η κανονικοποίηση είναι η μετατροπή αριθμητικών τιμών σε άλλες, πιο «κατάλληλες», αριθμητικές τιμές. Ένα επιπλέον θέμα που εμπίπτει στην προεπεξεργασία των δεδομένων είναι η μείωση του όγκου τους. Ειδική περίπτωση μείωσης των δεδομένων, με βαρύνουσα σημασία, είναι η επιλογή σημαντικών χαρακτηριστικών, η επιλογή δηλαδή εκείνων των μεταβλητών ή πεδίων που είναι απαραίτητες για την εξόρυξη της γνώσης.

Στο παρόν κεφάλαιο παρουσιάζονται οι βασικές τεχνικές, οι οποίες εφαρμόζονται στα πλαίσια της προεπεξεργασίας των δεδομένων. Ο αναγνώστης έχει την ευκαιρία να κατανοήσει τη θεωρία και τη λογική αυτών των τεχνικών και να γνωρίσει τις δυνατότητες τους, έτσι ώστε να μπορεί να τις εφαρμόσει στην πράξη. Αρχικά παρουσιάζεται το πρόβλημα των χαμένων τιμών και παρατίθενται τρόποι αντιμετώπισης του, όπως η αντικατάσταση με τη μέση ανά κλάση τιμή ή με την πιθανότερη τιμή, την οποία υπολογίζει ένα μοντέλο. Ακολούθως, γίνεται αναφορά στον θόρυβο των δεδομένων και προτείνονται τρόποι εξάλειψης του, όπως ο κατακερματισμός σε διαστήματα με ταυτόχρονη αντικατάσταση τιμών και ο εντοπισμός εξαιρέσεων. Τεχνικές κανονικοποίησης που καλύπτονται είναι η κανονικοποίηση ελάχιστου-μέγιστου, η κανονικοποίηση z-score και η κανονικοποίηση δεκαδικής κλιμάκωσης. Αναφορά γίνεται στο ζήτημα κατασκευής νέων πεδίων και παρουσιάζεται πραγματική περίπτωση που καταδεικνύει τη σημασία τους. Η επιλογή σημαντικών χαρακτηριστικών (feature selection) είναι από τις βασικότερες εργασίες προεπεξεργασίας των δεδομένων και σημαντικό πεδίο διεξαγωγής έρευνας. Στο παρόν κεφάλαιο γίνεται εκτεταμένη αναφορά στις μεθόδους επιλογής χαρακτηριστικών. Ειδικότερα, παρουσιάζονται το t στατιστικό τεστ και η ανάλυση διακύμανσης, η βηματική πρόσθια επιλογή και οπίσθια εξάλειψη, η επιλογή με χρήση αλγορίθμου, η μέθοδος CFS και η κατηγορία μεθόδων τύπου wrapper. Το τελευταίο υποκεφάλαιο αναφέρεται στο πρόβλημα της διακριτοποίησης των δεδομένων και παρουσιάζονται διάφορες τεχνικές, όπως τα διαστήματα ίσου πλάτους και ίσης συχνότητας, η βασισμένη στην εντροπία διακριτοποίηση, η τμηματοποίηση με φυσική κατάτμηση και η ασαφής διακριτοποίηση.

Προαπαιτούμενη γνώση

Αντικείμενο του Κεφαλαίου είναι η προεπεξεργασία των δεδομένων. Σε πολλά σημεία γίνονται αναφορές σε έννοιες και τεχνικές της Εξόρυξης Δεδομένων. Για τον λόγο αυτό, θεωρούμε απαραίτητη την προηγούμενη εμπέδωση των περιεχομένων του Κεφαλαίου 6, το οποίο εισάγει τον αναγνώστη σε αυτές τις βασικές έννοιες και τεχνικές. Οι εργασίες Εξαγωγής Μετασχηματισμού και Φόρτωσης (ΕΜΦ) (Extract Transform Load (ETL)) των δεδομένων είναι συγγενές αντικείμενο με την προεπεξεργασία. Θέματα εργασιών ΕΜΦ παρουσιάζονται σε σχετικό υποκεφάλαιο στα πλαίσια του Κεφαλαίου 4, το οποίο ασχολείται με τις Αποθήκες Δεδομένων. Ο αναγνώστης μπορεί να αναζητήσει πρόσθετες πληροφορίες στο εν λόγω υποκεφάλαιο, ώστε να αποκτήσει μια πιο ολοκληρωμένη εικόνα σχετικά με την προετοιμασία των δεδομένων. Η προηγούμενη γνώση ορισμένων ειδικών θεμάτων θα συμβάλλει στην καλύτερη κατανόηση αντικειμένων του παρόντος Κεφαλαίου. Βασικές έννοιες τεχνικών κατηγοριοποίησης, καθώς και η εντροπία των Δένδρων Αποφάσεων, που παρατίθενται στο Κεφάλαιο 9, είναι χρήσιμες για την εμπέδωση τεχνικών επιλογής χαρακτηριστικών, ζητημάτων κατασκευής νέων πεδίων καθώς και για την επιβλεπόμενη διακριτοποίηση. Για την κατανόηση της μεθόδου Ανάλυσης Κύριων Συνιστωσών προαπαιτούνται βασικές γνώσεις Γραμμικής Άλγεβρας και ειδικότερα πολλαπλασιασμού πινάκων. Τέλος, για πρόσθετο υλικό σχετικά με την προεπεξεργασία των δεδομένων, παραπέμπουμε τον αναγνώστη στο βιβλίο του Pyle (1999), του Svolba (2006), το οποίο όμως σε σημαντικό βαθμό αναφέρεται σε λογισμικά της SAS, και στο βιβλίο των Han, Kamber and Pei (2011).

7.1 Η αναγκαιότητα της προεπεξεργασίας δεδομένων

Ένας επίδοξος αναλυτής, ο οποίος αναλαμβάνει εργασίες ανάλυσης πραγματικών δεδομένων ενός οργανισμού, πολύ σύντομα θα διαπιστώσει ότι τα δεδομένα που τηρούνται στα διάφορα πληροφοριακά συστήματα πάσχουν από πολλά και διαφορετικά προβλήματα. Ένας αρχάριος αναλυτής πιθανότατα θα δυσφορούσε με όλα αυτά τα προβλήματα που καλείται να αντιμετωπίσει, πριν ακόμα αρχίσει την καθαυτό εργασία του. Ένας πιο έμπειρος αναλυτής όμως γνωρίζει ότι **η ύπαρξη προβλημάτων είναι ο κανόνας στα δεδομένα του πραγματικού κόσμου**.

Μια πρώτη αναφορά στα προβλήματα των δεδομένων έγινε στο Κεφάλαιο 4 και ειδικότερα στο υποκεφάλαιο, το οποίο αναφέρεται στις εργασίες Εξαγωγής Μετασχηματισμού και Φόρτωσης (ΕΜΦ) (Extract, Transform, Load (ETL)). Στα δεδομένα που τηρούνται στα πηγαία συστήματα, πιθανόν να γίνεται χρήση διαφορετικών ονομάτων για το ίδιο αντικείμενο ή του ίδιου ονόματος για διαφορετικά αντικείμενα, να χρησιμοποιούνται διαφορετικές μονάδες μέτρησης, να εφαρμόζεται διαφορετικός τρόπος κωδικοποίησης της ίδιας πληροφορίας, να χρησιμοποιείται διαφορετικός τύπος δεδομένων για την ίδια πληροφορία(πχ ακέραιος ή πραγματικός αριθμός), να υπάρχουν διαφορετικά επίπεδα συναθροίσεων (πχ πωλήσεις ανά ημέρα ή πωλήσεις ανά μήνα). Τα προβλήματα αυτά υπάρχουν εξαιτίας του γεγονότος ότι τα δεδομένα είναι διάσπαρτα σε διάφορα συστήματα. Εάν ο οργανισμός διαθέτει Αποθήκη Δεδομένων, τότε αυτά τα προβλήματα έχουν αντιμετωπιστεί στα πλαίσια των εργασιών ΕΜΦ. Εάν όμως δεν υπάρχει Αποθήκη Δεδομένων και ο αναλυτής πρέπει να ανατρέξει στα πηγαία συστήματα, τότε τα προβλήματα αυτά πρέπει να αντιμετωπιστούν από την αρχή. Τρόποι αντιμετώπισης ορισμένων προβλημάτων, όπως η ύπαρξη διαφορετικών κωδικών για το ίδιο αντικείμενο σε δύο ή περισσότερες πηγές, παρουσιάζονται στο Κεφάλαιο 4.

Δυστυχώς τα προβλήματα των δεδομένων δεν περιορίζονται μόνο σε αυτά που προκύπτουν από την ανάγκη συγχώνευσης διάσπαρτων δεδομένων, τα οποία είναι αποθηκευμένα σε πολλές πηγές. Αν θεωρήσει κανείς τα δεδομένα ενός μόνον συστήματος, θα διαπιστώσει ότι υπάρχουν διπλοκαταχωρημένες εγγραφές, εγγραφές με αντικρουόμενο περιεχόμενο, τιμές που παραβιάζουν λογικούς κανόνες, εσφαλμένες τιμές (πχ αρνητικές ποσότητες πωλήσεων), χρήση συνωνύμων τιμών σε κάποιο πεδίο (πχ η «Θεσσαλονίκη» μπορεί να καταχωρείται ολογράφως, ή ως «Θεσ/νικη») κλπ. Ένα πολύ συνηθισμένο πρόβλημα είναι η ύπαρξη χαμένων τιμών, η έλλειψη δηλαδή τιμών σε ορισμένα πεδία καταχωρημένων εγγραφών. Οι **χαμένες τιμές** μπορεί να οφείλονται σε πλήθος διαφορετικών λόγων. Για παράδειγμα, ορισμένες πληροφορίες σχετικά με τον πελάτη μπορεί να μην είναι διαθέσιμες τη στιγμή δημιουργίας της καρτέλας του και να καταχωρηθούν μόνο τα διαθέσιμα στοιχεία. Διαγραφή δεδομένων μπορεί να γίνει από ανθρώπινο χειριστικό λάθος ή και από αστοχία του εξοπλισμού. Δεν είναι σπάνιες οι περιπτώσεις, που ορισμένες πληροφορίες θεωρήθηκαν «μη σημαντικές » και δεν καταγράφηκαν από τους χειριστές των συστημάτων ή περιπτώσεις κακής συνεννόησης μεταξύ των χειριστών και των προϊσταμένων τους. Σε κάθε περίπτωση, είναι σχεδόν σίγουρο ότι ο αναλυτής θα συναντήσει χαμένες τιμές και πρέπει να είναι ικανός να αντιμετωπίσει το πρόβλημα.

Ένα άλλο πρόβλημα των δεδομένων είναι ο λεγόμενος «**θόρυβος**». Τα δεδομένα μπορεί να περιέχουν λανθασμένες τιμές. Λόγοι ύπαρξης λαθών είναι τα σφάλματα των χειριστών τη στιγμή της καταχώρησης των δεδομένων ή της χρήσης του λογισμικού, αλλά και τεχνικά προβλήματα, όπως η κακή λειτουργία συσκευών που καταγράφουν δεδομένα (πχ συσκευών ανάγνωσης ετικετών RFID), καθώς και πιθανά προβλήματα στη μετάδοση των δεδομένων μέσω ενός δικτύου. Πέραν της ύπαρξης σφαλμάτων, ένα άλλο συγγενές πρόβλημα είναι η ύπαρξη δεδομένων με ακραίες τιμές. Αυτά τα δεδομένα-εξαιρέσεις δεν προσφέρουν στην ανάλυση χρήσιμη πληροφορία, καθώς περιγράφουν σπάνιες και μεμονωμένες περιπτώσεις και δεν εκφράζουν κάποια κανονικότητα. Αντιθέτως, σε πολλές περιπτώσεις μπορεί να αποπροσανατολίσουν τους αλγορίθμους εξόρυξης και να τους οδηγήσουν σε εσφαλμένα ή μεροληπτικά συμπεράσματα. Βεβαίως, οφείλουμε να αναφέρουμε ότι σε ορισμένες εργασίες εξόρυξης δεδομένων, το ενδιαφέρον επικεντρώνεται αποκλειστικά στα δεδομένα που αποκλίνουν από το «κανονικό» ή το συνηθισμένο. Τέτοιες περιπτώσεις είναι τα προβλήματα εντοπισμού απάτης. Ωστόσο, κατά κανόνα τα δεδομένα με ακραίες τιμές θεωρούνται πρόβλημα και πρέπει να τύχουν ειδικού χειρισμού. Τα δεδομένα που περιέχουν σφάλματα και ακραίες τιμές, που περιέχουν δηλαδή μη χρήσιμη πληροφορία, χαρακτηρίζονται **θορυβώδη**. Η αντιμετώπιση του θορύβου είναι τμήμα των εργασιών, οι οποίες πραγματοποιούνται στα πλαίσια της προεπεξεργασίας των δεδομένων.

Ένας όρος που έχει επικρατήσει στη βιβλιογραφία της εξόρυξης δεδομένων και περιγράφει δεδομένα με χαμένες τιμές, θόρυβο και άλλα προβλήματα, είναι ο όρος «**ακάθαρτα δεδομένα**» (dirty data). Οι οργανισμοί, σε μια προσπάθεια βελτίωσης της ποιότητας των δεδομένων τους, καθιερώνουν κανόνες καταγραφής δεδομένων. Τέτοιες πρακτικές, αν και περιορίζουν σημαντικά το πρόβλημα, δεν μπορούν να το εξαλείψουν τελείως. Επίσης, ιστορικά δεδομένα, που έχουν συλλεχτεί σε προηγούμενο χρόνο, διατηρούν τα προβλήματα τους. Σύμφωνα με μελέτες, το ύψος των σφαλμάτων των δεδομένων κατά την απόκτηση τους φθάνει το

5% (Orr, 1998), ενώ το συνολικό ποσοστό των δεδομένων, που με τον ένα ή με τον άλλο τρόπο μπορούν να χαρακτηριστούν ακάθαρτα, αγγίζει το 40% (Fayyad, Piatetsky-Shapiro & Uthurusamy, 2003). Τα ακάθαρτα δεδομένα μπορούν να προκαλέσουν σύγχυση στους αλγορίθμους εξόρυξης. Για τον λόγο αυτό, απαιτείται η αντιμετώπιση των προβλημάτων σε χρόνο προηγούμενο από την καθαυτό ανάλυση. Η διαδικασία αντιμετώπισης των χαμένων τιμών, του θορύβου, των ασυνεπειών και άλλων προβλημάτων των δεδομένων ονομάζεται «**καθαρισμός δεδομένων**» (data cleansing) και αποτελεί μέρος των εργασιών της προεπεξεργασίας τους.

Ο καθαρισμός των δεδομένων δεν είναι το μοναδικό αντικείμενο της προεπεξεργασίας. Πολλές φορές είναι αναγκαία η προσαρμογή των δεδομένων στις απαιτήσεις των μεθόδων επεξεργασίας. Ορισμένες μέθοδοι δεν μπορούν να χειριστούν συνεχείς τιμές (αριθμούς), αλλά χρειάζονται ονομαστικές τιμές. Εάν κάποιος επιθυμεί να χρησιμοποιήσει τέτοιες μεθόδους και τα δεδομένα του περιλαμβάνουν αριθμητικά πεδία, τότε πρέπει να μετατρέψει τις αριθμητικές τιμές σε ονομαστικές. Η διαδικασία αυτή ονομάζεται **διακριτοποίηση** (discretization). Άλλες μέθοδοι που χειρίζονται συνεχόμενες τιμές, αντιμετωπίζουν προβλήματα εάν ορισμένες μεταβλητές περιέχουν πολύ μεγάλες τιμές, ενώ άλλες μεταβλητές περιέχουν μικρές τιμές. Για παράδειγμα, ο αριθμοδείκτης Αμοιβές Εξωτερικών Ελεγκτών προς Σύνολο Ενεργητικού περιέχει πολύ μικρούς δεκαδικούς αριθμούς, ενώ η μεταβλητή Πωλήσεις περιέχει τιμές που ανέρχονται στο ύψος εκατομμυρίων ή και δισεκατομμυρίων. Ορισμένες μέθοδοι, όπως η μέθοδος κατηγοριοποίησης k-Πλησιέστεροι Γείτονες, είναι ιδιαίτερα ευπαθείς στην ύπαρξη τέτοιων συνθηκών και οι μεταβλητές με τις μεγάλες τιμές θα καθορίσουν το αποτέλεσμα, ενώ η επιρροή των μεταβλητών με τις μικρές τιμές θα είναι ανύπαρκτη. Επίσης, τα Νευρωνικά Δίκτυα λειτουργούν καλύτερα όταν οι τιμές που επεξεργάζονται κυμαίνονται στην περιοχή [0..1]. Σε τέτοιες περιπτώσεις, πρέπει να γίνει αναγωγή των αριθμητικών τιμών σε άλλες αριθμητικές τιμές, που να κυμαίνονται εντός των ορίων της επιθυμητής περιοχής. Η διαδικασία αυτή ονομάζεται **κανονικοποίηση** (normalization). Τέλος, τα δεδομένα μπορούν να τροποποιηθούν έτσι ώστε να εκφράζουν γενικεύσεις, πχ συγκεντρωτικές πωλήσεις ανά περιοχή. Εργασίες μετατροπής των δεδομένων όπως η κανονικοποίηση και η γενίκευση ονομάζονται **μετασχηματισμός** των δεδομένων (data transformation) και αποτελούν μια ακόμα μορφή προεπεξεργασίας των δεδομένων.

Ένα άλλο σύνολο εργασιών, που εκτελούνται στα πλαίσια της προεπεξεργασίας, αφορούν τη **μείωση των δεδομένων**. Η ικανότητα χειρισμού μεγάλου όγκου δεδομένων είναι ένα από τα βασικά χαρακτηριστικά της Εξόρυξης Δεδομένων. Όμως δεδομένα μεγάλου όγκου μπορούν να προκαλέσουν προβλήματα στις μεθόδους επεξεργασίας και μεγάλες καθυστερήσεις στη διεξαγωγή των αναλύσεων. Για τον λόγο αυτό, είναι χρήσιμη η μείωση του όγκου των δεδομένων. Η μείωση του όγκου δεν είναι μια τετριμμένη εργασία, καθώς τα αποτελέσματα της ανάλυσης των μειωμένων δεδομένων πρέπει να είναι τα ίδια ή περίπου τα ίδια με τα αποτελέσματα της ανάλυσης του συνόλου των δεδομένων. Μια ειδική περίπτωση μείωσης του όγκου είναι η **επιλογή σημαντικών χαρακτηριστικών** (feature selection). Τα διαθέσιμα δεδομένα περιλαμβάνουν πολλά χαρακτηριστικά (στήλες). Ωστόσο, για μια συγκεκριμένη εργασία εξόρυξης, δεν είναι χρήσιμα όλα αυτά τα χαρακτηριστικά. Πολλά χαρακτηριστικά περιέχουν πληροφορίες που δεν σχετίζονται με το αντικείμενο της ανάλυσης. Επίσης, μπορεί να καταγράφονται διαφορετικές εκδοχές της ίδιας πληροφορίας σε διάφορα χαρακτηριστικά. Για παράδειγμα, υπάρχει πλήθος αριθμοδεικτών που εκφράζει την κερδοφορία μιας επιχείρησης (κέρδη προς σύνολο ενεργητικού, προς πωλήσεις, προς μετοχικό κεφάλαιο κλπ.). Είναι προς διερεύνηση ποια μεταβλητή είναι η πλέον κατάλληλη για την αναλυτική εργασία που διεξάγεται. Τέλος, μεταξύ των χαρακτηριστικών μπορεί να υπάρχουν τέτοιες αλληλεξαρτήσεις που να καθιστούν την ταυτόχρονη παρουσία τους περιττή ή και επιζήμια. Για τους λόγους αυτούς, απαιτείται να γίνει πολύ προσεκτικά η επιλογή εκείνου του υποσυνόλου των χαρακτηριστικών, το οποίο είναι το πλέον κατάλληλο για τη συγκεκριμένη εργασία εξόρυξης γνώσης. Η εργασία αυτή είναι γνωστή ως επιλογή χαρακτηριστικών.

Η προεπεξεργασία των δεδομένων αποτελεί ένα βασικό στάδιο της διαδικασίας ανακάλυψης γνώσης. Στα πλαίσια της εκτελούνται εργασίες καθαρισμού των δεδομένων, ολοκλήρωσης τους, μετασχηματισμού τους καθώς και μείωσης τους. Στις επόμενες σελίδες του παρόντος κεφαλαίου θα γνωρίσουμε αρκετές τεχνικές, οι οποίες εφαρμόζονται για τη διεξαγωγή αυτών των εργασιών.

7.2 Χαμένες Τιμές

Η ύπαρξη χαμένων τιμών (missing values) είναι ένα από τα συνηθέστερα προβλήματα των δεδομένων του πραγματικού κόσμου. Οι λόγοι για αυτό το φαινόμενο είναι πολλοί. Κάποιες πληροφορίες μπορεί να μην ήταν διαθέσιμες την ώρα της καταχώρησης ή μπορεί να διαγράφηκαν αργότερα από λάθος. Μια νέα στήλη μπορεί να προστέθηκε στον πίνακα, οπότε όλες οι προηγούμενες καταχωρήσεις θα έχουν κενά στα κελιά αυτής της στήλης. Η αστοχία υλικού ή λογισμικού είναι ένας άλλος λόγος ύπαρξης χαμένων τιμών. Οι χαμένες τιμές

είναι ένα σημαντικό πρόβλημα στην εξόρυξη δεδομένων γιατί μπορεί να αποπροσανατολίσουν τους αλγορίθμους. Ορισμένες μέθοδοι εξόρυξης δεδομένων, όπως πχ τα Δένδρα Αποφάσεων τύπου C4.5, αντιμετωπίζουν ενδογενώς το πρόβλημα των χαμένων τιμών. Ωστόσο, αυτό δεν ισχύει για όλες τις μεθόδους και επιπλέον κάθε τέτοια μέθοδος αντιμετωπίζει το πρόβλημα με διαφορετικό τρόπο. Εάν ο χρήστης επιθυμεί να συγκρίνει δύο μεθόδους εξόρυξης δεδομένων, θα υπάρχει διαφοροποίηση στη διαδικασία, ήδη από τον χειρισμό των χαμένων τιμών. Για τον λόγο αυτό, είναι προτιμότερο να επιλύσει ο χρήστης το πρόβλημα των χαμένων τιμών πριν από την καθαυτό διαδικασία εξόρυξης και με τρόπο κοινό και ελεγχόμενο από αυτόν.

Έχουν προταθεί διάφοροι τρόποι για την αντιμετώπιση των χαμένων τιμών. Ο Πίνακας 7.1 περιέχει σύνολο δεδομένων με στοιχεία χορήγησης δανείων και περιλαμβάνει χαμένες τιμές.

Ετήσιο Εισόδημα	Πιστοληπτική ικανότητα	Έγκριση δανείου
15000	Μέτρια	Ναι
12000	Κακή	Όχι
	Μέτρια	Όχι
50000	Καλή	Ναι
30000		Ναι
16000	Κακή	Όχι

Πίνακας 7.1 *Σύνολο δεδομένων με χαμένες τιμές*

Ορισμένοι από τους πιθανούς τρόπους αντιμετώπισης των χαμένων τιμών είναι οι ακόλουθοι:.

- **Διαγραφή ολόκληρης της γραμμής**. Ο τρόπος αυτός δεν ενδείκνυται γιατί προκαλεί απώλεια χρήσιμης πληροφορίας. Εφαρμόζεται μόνο σε περιπτώσεις κατά τις οποίες λείπει η τιμή της κλάσης (έγκριση δανείου στο παράδειγμα μας) ή σε περιπτώσεις όπου η γραμμή περιέχει πολλές χαμένες τιμές.
- **Αναζήτηση και καταχώρηση της πραγματικής τιμής**. Θεωρητικά αυτή θα ήταν η καλύτερη λύση. Ωστόσο, συνήθως είναι η λιγότερο εφικτή. Κατά κανόνα τα δεδομένα είναι πάρα πολλά και η αναζήτηση των τιμών είναι απαγορευτικά χρονοβόρα, ενώ σε πολλές περιπτώσεις είναι αδύνατη η εύρεση της πραγματικής τιμής.
- **Χρήση μιας σταθερής τιμής για όλες τις χαμένες τιμέ**ς, όπως πχ της λέξης «άγνωστη». Ο τρόπος δεν ενδείκνυται γιατί οι αλγόριθμοι που θα επεξεργαστούν τα δεδομένα μπορεί να εκλάβουν την τιμή αυτή ως έγκυρη, να τη συμπεριλάβουν στην επεξεργασία και να οδηγηθούν σε εσφαλμένα συμπεράσματα.
- **Αντικατάσταση της χαμένης τιμής με τη μέση τιμή της στήλη**ς αν το πεδίο είναι αριθμητικό **ή με τη συνηθέστερη τιμή** αν το πεδίο είναι ονομαστικό. Στον Πίνακα 7.2 η χαμένη τιμή στο Ετήσιο Εισόδημα αντικαταστάθηκε με τη μέση τιμή των εισοδημάτων.

Ετήσιο Εισόδημα	Πιστοληπτική ικανότητα	Έγκριση δανείου
15000	Μέτρια	Ναι
12000	Κακή	Όχι
24600	Μέτρια	Όχι
50000	Καλή	Ναι
30000		Ναι
16000	Κακή	Όχι
ΜΟ=24600		

Πίνακας 7.2

- Αντικατάσταση της χαμένης τιμής με τη μέση τιμή της κλάσης αν το πεδίο είναι αριθμητικό ή με τη συνηθέστερη τιμή αν το πεδίο είναι ονομαστικό. Μπορεί να εφαρμοστεί όταν τα δεδομένα περιέχουν μια στήλη που ορίζει την κατηγορία των παρατηρήσεων. Στο παράδειγμα μας, η στήλη αυτή είναι η έγκριση του δανείου. Η χαμένη τιμή του ετήσιου εισοδήματος ανήκει στις περιπτώσεις που το δάνειο δεν εγκρίθηκε. Θα υπολογιστεί τότε ο μέσος όρος εισοδήματος μόνο των περιπτώσεων που δεν εγκρίθηκε το δάνειο. Στον Πίνακα 7.3 η χαμένη τιμή στο Ετήσιο Εισόδημα αντικαταστάθηκε με τη

μέση τιμή των εισοδημάτων όσων δεν εγκρίθηκε το δάνειο.

Ετήσιο Εισόδημα	Πιστοληπτική ικανότητα	Έγκριση δανείου
15000	Μέτρια	Ναι
12000	Κακή	Όχι
14000	Μέτρια	Όχι
50000	Καλή	Ναι
30000		Ναι
16000	Κακή	Όχι
MO = 14000		

Πίνακας 7.3

- **Αντικατάσταση της χαμένης τιμής με κάθε δυνατή τιμή**. Σύμφωνα με αυτόν τον τρόπο, προστίθενται νέες γραμμές στον πίνακα. Αν υπάρχουν *N* δυνατές τιμές για τη χαμένη τιμή τότε προστίθενται *N-1* γραμμές. Στο παράδειγμα μας για τη στήλη «Πιστοληπτική ικανότητα» υπάρχουν τρεις δυνατές τιμές, δηλαδή καλή, μέτρια και κακή. Θα προστεθούν 2 γραμμές, και στις τρεις αυτές γραμμές (οι δύο καινούργιες και η παλιά) τα δεδομένα θα είναι ίδια, εκτός από το κελί της χαμένης τιμής, όπου θα υπάρχουν οι εναλλακτικές τιμές. Στον Πίνακα 7.4, στη θέση της γραμμής με τη χαμένη τιμή για την πιστοληπτική ικανότητα, βρίσκονται τρεις γραμμές με τις τρεις εναλλακτικές τιμές.

Ετήσιο Εισόδημα	Πιστοληπτική ικανότητα	Έγκριση δανείου
15000	Μέτρια	Ναι
12000	Κακή	Όχι
	Μέτρια	Όχι
50000	Καλή	Ναι
30000	Καλή	Ναι
30000	Μέτρια	Ναι
30000	Κακή	Ναι
16000	Κακή	Όχι

Πίνακας 7.4

- **Αντικατάσταση της χαμένης τιμής με κάθε δυνατή τιμή για τις παρατηρήσεις της κλάσης**. Ο τρόπος αυτός μοιάζει με τον προηγούμενο. Η διαφορά έγκειται στο γεγονός ότι επιλέγονται μόνον οι εναλλακτικές τιμές για τη συγκεκριμένη κατηγορία, στην οποία ανήκει το αντικείμενο. Στο παράδειγμα μας, η παρατήρηση με τη χαμένη τιμή στη στήλη «πιστοληπτική ικανότητα» υπάγεται στις περιπτώσεις όπου το δάνειο εγκρίνεται. Από τα δεδομένα προκύπτει ότι όταν εγκρίνεται το δάνειο η πιστοληπτική ικανότητα είναι καλή ή μέτρια. Στον Πίνακα 7.5 στη θέση της γραμμής με τη χαμένη τιμή για την πιστοληπτική ικανότητα, βρίσκονται δύο γραμμές με τις δύο εναλλακτικές τιμές όσων εγκρίθηκε το δάνειο.

Ετήσιο Εισόδημα	Πιστοληπτική ικανότητα	Έγκριση δανείου
15000	Μέτρια	Ναι
12000	Κακή	Όχι
	Μέτρια	Όχι
50000	Καλή	Ναι
30000	Μέτρια	Ναι
30000	Καλή	Ναι
16000	Κακή	Όχι

Πίνακας 7.5

- **Πρόβλεψη της χαμένης τιμής**. Σύμφωνα με τον τρόπο αυτό, το πρόβλημα της χαμένης τιμής αντιμετωπίζεται σαν πρόβλημα κατηγοριοποίησης (αν το πεδίο είναι ονομαστικό) ή παλινδρόμησης (αν το πεδίο είναι αριθμητικό). Αναπτύσσεται ένα μοντέλο ικανό να υπολογίζει τις τιμές της στήλης με τη χαμένη τιμή από τα δεδομένα των άλλων στηλών. Η πιθανότερη τιμή που θα υπολογιστεί με τη χρήση του μοντέλου αντικαθιστά τη χαμένη τιμή. Στον Πίνακα 7.6, η χαμένη τιμή στο Ετήσιο Εισόδημα αντικαταστάθηκε με την τιμή που έχει προβλεφθεί από το μοντέλο.

Ετήσιο Εισόδημα	Πιστοληπτική ικανότητα	Έγκριση δανείου
15000	Μέτρια	Ναι
12000	Κακή	Όχι
18000	Μέτρια	Όχι
50000	Καλή	Ναι
30000		Ναι
16000	Κακή	Όχι
Πρόβλεψη μοντέλου 18000		

Πίνακας 7.6

Όλοι οι τρόποι οι οποίοι αντικαθιστούν τη χαμένη τιμή με μία άλλη, όχι την πραγματική, μπορεί να προκαλέσουν απόκλιση στα δεδομένα, αφού η τιμή αντικατάστασης κατά πάσα πιθανότητα δεν είναι η σωστή. Ωστόσο, στις περισσότερες περιπτώσεις δεν υπάρχει εναλλακτική λύση. Από όλους τους τρόπους που αναφέρθηκαν παραπάνω, ο τελευταίος μπορεί να επιτύχει μια ικανοποιητική προσέγγιση, γιατί αξιοποιεί την πληροφορία των άλλων στηλών και διατηρεί τη σχέση μεταξύ των δεδομένων. Επιτυχημένα μοντέλα κατηγοριοποίησης ή πρόβλεψης προσεγγίζουν τις πραγματικές τιμές σε ποσοστό περίπου 80%. Για τον λόγο αυτό, ο τελευταίος τρόπος προτείνεται ως μια δόκιμη πρακτική συμπλήρωσης των χαμένων τιμών.

7.3 Θορυβώδη Δεδομένα

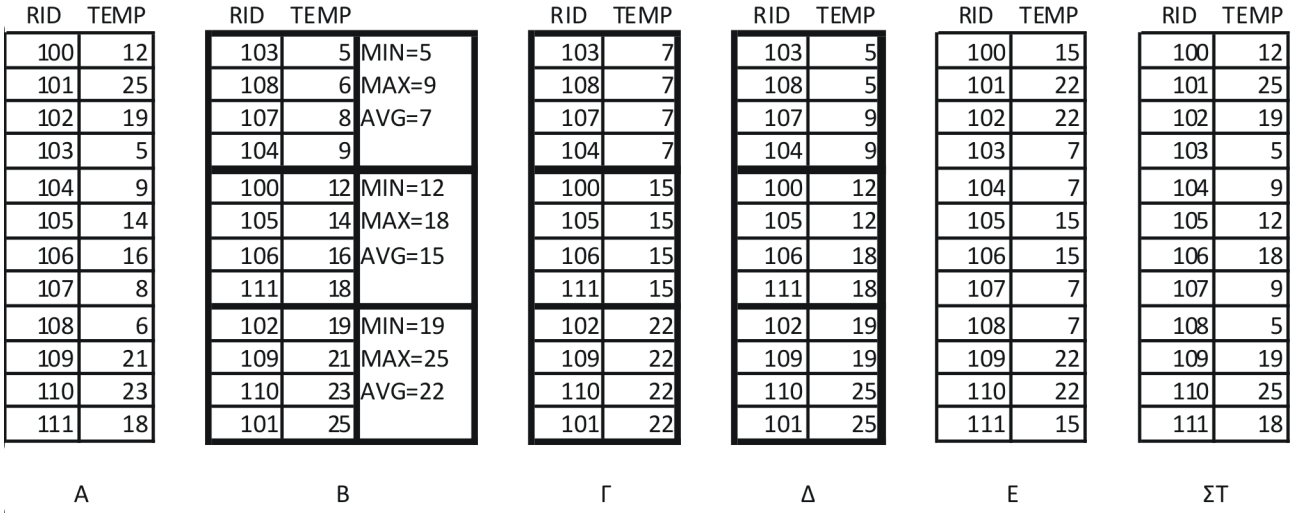
Θορυβώδη ονομάζονται τα δεδομένα τα οποία περιέχουν εσφαλμένες τιμές και τιμές-εξαιρέσεις, τιμές δηλαδή που δεν προσφέρουν χρήσιμη πληροφορία στην ανάλυση. Η ύπαρξη θορύβου προκαλεί προβλήματα στους αλγορίθμους εξόρυξης και πρέπει να αντιμετωπιστεί στα πλαίσια της προεπεξεργασίας των δεδομένων. Υπάρχουν δύο τακτικές αντιμετώπισης του θορύβου. Η πρώτη βασίζεται στην αντικατάσταση όλων των αριθμητικών τιμών με άλλες κατάλληλες τιμές. Η δεύτερη βασίζεται στον εντοπισμό των ακραίων τιμών. Αφού εντοπιστούν εγγραφές με ακραίες τιμές, υπάρχουν δύο πιθανές εκδοχές. Οι εγγραφές αυτές μπορούν να διαγραφούν από το σύνολο των δεδομένων ή μπορεί να παραμείνουν και να τροποποιηθούν οι ακραίες τιμές. Αναλυτικότερα, ορισμένες μέθοδοι αντιμετώπισης του θορύβου είναι οι ακόλουθες:

Κατακερματισμός σε διαστήματα και αντικατάσταση τιμών. Σύμφωνα με τη μέθοδο αυτή, οι τιμές μιας μεταβλητής ταξινομούνται σε αύξουσα σειρά και χωρίζονται σε διαστήματα. Τα διαστήματα μπορεί να είναι ίσου πλάτους ή ίσης συχνότητας. Τα διαστήματα ίσου πλάτους έχουν όλα το ίδιο εύρος τιμών. Τα διαστήματα ίσης συχνότητας έχουν όλα ίσο πλήθος τιμών. Αφού οριστούν τα διαστήματα, γίνεται αντικατάσταση όλων των τιμών. Υπολογίζονται νέες τιμές για κάθε διάστημα και αντικαθιστούν τις παλιές. Επειδή μια τιμή προκύπτει από τις γειτονικές της τιμές, η εξομάλυνση αυτή ονομάζεται τοπική. Υπάρχουν παραλλαγές ως προς το ποιες θα είναι οι τιμές αντικατάστασης:

- **Με την αντικατάσταση μέσων όρων**, υπολογίζεται για κάθε διάστημα ο μέσος όρος και στη συνέχεια, αντικαθιστά όλες τις τιμές του διαστήματος.
- **Με την αντικατάσταση οριακών τιμών**, κάθε τιμή αντικαθίσταται με τη μεγαλύτερη ή τη μικρότερη τιμή του διαστήματος. Αν η εκάστοτε τιμή είναι πλησιέστερα στη μικρότερη τιμή του διαστήματος, τότε αντικαθίσταται με αυτήν, διαφορετικά αντικαθίσταται με τη μεγαλύτερη τιμή του διαστήματος.

Στο Σχήμα 7.1 παρουσιάζεται παράδειγμα κατακερματισμού με διαστήματα. Στο τμήμα Α) περιλαμβάνονται δεδομένα θερμοκρασίας. Στη στήλη RID βρίσκονται οι αναγνωριστικοί αριθμοί των εγγραφών (Record ID) και στη στήλη TEMP οι τιμές θερμοκρασίας. Στο τμήμα Β) οι εγγραφές ταξινομούνται σε αύξουσα σειρά θερμοκρασίας (προσοχή στις τιμές των RID), χωρίζονται σε διαστήματα ίσης συχνότητας, με κάθε διάστημα

να περιέχει 4 τιμές και υπολογίζονται οι μεγαλύτερες, οι μικρότερες και οι μέσες τιμές κάθε διαστήματος. Στο τμήμα Γ) γίνεται αντικατάσταση μέσων όρων και στο τμήμα Δ) αντικατάσταση οριακών τιμών. Τέλος, στο τμήμα Ε) εμφανίζονται τα δεδομένα όπως θα είναι μετά την αντικατάσταση μέσων όρων και στο τμήμα ΣΤ) τα δεδομένα όπως θα είναι μετά την αντικατάσταση οριακών τιμών.



Σχήμα 7.1 *Αντικατάσταση τιμών για αντιμετώπιση θορύβου*

Στατιστικός Εντοπισμός Εξαιρέσεων. Με τη μέθοδο αυτή εντοπίζονται εγγραφές, οι οποίες σε ορισμένα πεδία τους περιέχουν ακραίες τιμές. Για κάθε πεδίο X υπολογίζεται η μέση τιμή M_X και η τυπική απόκλιση σ_X. Στη συνέχεια, εντοπίζονται οι τιμές που απέχουν από τη μέση τιμή απόσταση μεγαλύτερη από $k*\sigma_X$. Αν μια τιμή x_1 είναι μικρότερη από $M_X - k*\sigma_X$ ή μεγαλύτερη από $M_X + k*\sigma_X$ τότε θεωρείται ακραία. Ο καθορισμός του συντελεστή k γίνεται από τον χρήστη, και βασίζεται στη γνώση του σχετικά με τα δεδομένα ή το εξεταζόμενο πρόβλημα.

Χρήση Ανάλυσης Συστάδων. Οι μέθοδοι Ανάλυσης Συστάδων ομαδοποιούν αντικείμενα με βάση την ομοιότητα τους. Το αποτέλεσμα είναι ορισμένες ομάδες ομοειδών αντικειμένων, ωστόσο κάποια αντικείμενα είναι σημαντικά ανόμοια με όλα τα υπόλοιπα και δεν εντάσσονται σε καμία ομάδα. Εφαρμόζοντας τεχνικές Ανάλυσης Συστάδων σε ένα σύνολο τιμών, εντοπίζονται ορισμένες τιμές, οι οποίες δεν εντάσσονται σε καμία ομάδα και θεωρούνται εξαιρέσεις. Η Ανάλυση Συστάδων παρουσιάζεται διεξοδικά στο Κεφάλαιο 12.

Προσαρμογή των δεδομένων με χρήση μοντέλου. Με τη μέθοδο αυτή αναπτύσσεται ένα μοντέλο ικανό να προβλέπει τις τιμές του πεδίου, χρησιμοποιώντας πληροφορίες από άλλα πεδία. Ένα τέτοιο μοντέλο μπορεί να αναπτυχθεί με χρήση της Πολλαπλής Γραμμικής Παλινδρόμησης, η οποία εκφράζει ένα αριθμητικό πεδίο σαν γραμμικό συνδυασμό άλλων αριθμητικών πεδίων. Οι τιμές του πεδίου μπορούν να μεταβληθούν με βάση τις προβλέψεις του μοντέλου. Ο Teng (1999), σε μια μελέτη για την αντιμετώπιση του προβλήματος του θορύβου στα δεδομένα, χρησιμοποιεί τεχνικές κατηγοριοποίησης για να εντοπίσει θορυβώδεις τιμές στα χαρακτηριστικά που χρησιμοποιούνται για την πρόβλεψη, αλλά και στο χαρακτηριστικό της κλάσης. Οι προβλέψεις του μοντέλου χρησιμοποιούνται για τη διόρθωση των ακραίων τιμών στα δεδομένα.

7.4 Κανονικοποίηση

Η κανονικοποίηση (normalization) είναι μια διαδικασία μετασχηματισμού δεδομένων, κατά την οποία αριθμητικές τιμές αντικαθίστανται με άλλες, πιο «κατάλληλες», αριθμητικές τιμές. Η κανονικοποίηση των δεδομένων γίνεται ώστε να αντιμετωπιστούν δυσκολίες ορισμένων μεθόδων εξόρυξης. Για παράδειγμα, τα Νευρωνικά Δίκτυα λειτουργούν καλύτερα όταν οι τιμές εισόδου κυμαίνονται στην περιοχή [0,0..1,0]. Επίσης, η μέθοδος των k-Πλησιέστερων Γειτόνων, η οποία υπολογίζει αποστάσεις μεταξύ των παρατηρήσεων, αντιμετωπίζει πρόβλημα όταν ορισμένες μεταβλητές εισόδου έχουν μικρές τιμές, ενώ άλλες μεταβλητές έχουν μεγάλες τιμές. Το πρόβλημα συνίσταται στο γεγονός ότι οι μεταβλητές με τις μεγάλες τιμές καθορίζουν ουσιαστικά την απόσταση των παρατηρήσεων, ενώ οι μεταβλητές με τις μικρές τιμές επηρεάζουν την απόσταση ελάχιστα και τελικά, δεν παίζουν κανένα ρόλο στον υπολογισμό του αποτελέσματος.

Υπάρχουν διάφορες μέθοδοι κανονικοποίησης των αριθμητικών τιμών. Ορισμένες από τις πλέον χρησιμοποιούμενες είναι οι ακόλουθες:

Κανονικοποίηση ελάχιστου-μέγιστου. Με αυτήν τη μέθοδο κανονικοποίησης, οι αριθμητικές τιμές αντιστοιχίζονται με άλλες, οι οποίες κυμαίνονται εντός μιας προκαθορισμένης περιοχής τιμών. Η αντιστοίχιση γίνεται με γραμμικό μετασχηματισμό. Αν θεωρήσουμε μια μεταβλητή *A*, όπου η μεγαλύτερη τιμή της είναι η max_A και η μικρότερη τιμή της είναι η min_A, μπορούμε να αντιστοιχίσουμε όλες τις τιμές με άλλες που κυμαίνονται εντός μιας περιοχής με κατώτερο όριο την new_min_A και ανώτερο όριο την new_max_A σύμφωνα με τη Σχέση 7.1

$$x' = \frac{x - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \tag{7.1}$$

όπου *x* η εκάστοτε τιμή της μεταβλητής *A* και *x'* η νέα τιμή. Η μέθοδος αυτή έχει το πλεονέκτημα ότι ο χρήστης προκαθορίζει την περιοχή τιμών, για παράδειγμα μπορεί να μετασχηματίσει τις τιμές έτσι ώστε να κυμαίνονται στην περιοχή [0,0..1,0], ορίζοντας σαν new_min_A την τιμή 0 και σαν new_max_A την τιμή 1. Επίσης, με τη μέθοδο αυτή διατηρείται η αναλογία μεταξύ των τιμών που υπήρχε στα αρχικά δεδομένα.

Κανονικοποίηση z-score. Η μέθοδος αυτή πραγματοποιεί μετασχηματισμό των αριθμητικών τιμών, χρησιμοποιώντας τη μέση τιμή και την τυπική απόκλιση τους. Για μία μεταβλητή *A*, με μέση τιμή M_A και τυπική απόκλιση σ_A, ο μετασχηματισμός των τιμών γίνεται σύμφωνα με τη Σχέση 7.2

$$x' = \frac{x - M_A}{\sigma_A} \tag{7.2}$$

όπου *x* η εκάστοτε τιμή της μεταβλητής *A* και *x'* η νέα τιμή. Η μέθοδος αυτή είναι ιδιαίτερα κατάλληλη σε περιπτώσεις όπου τα δεδομένα περιέχουν ακραίες τιμές, γιατί η κανονικοποίηση ελάχιστου-μέγιστου θα συγκέντρωνε τη μεγάλη πλειοψηφία των τιμών σε ένα ελάχιστο τμήμα της περιοχής τιμών και θα χρησιμοποιούσε το υπόλοιπο τμήμα της περιοχής τιμών για τις εξαιρέσεις. Επίσης, η μέθοδος δίνει τιμές των οποίων η μέση τιμή ισούται με 0.

Κανονικοποίηση δεκαδικής κλιμάκωσης. Η μέθοδος αυτή πραγματοποιεί υποδεκαπλασιασμό των τιμών, διαιρώντας τες με μια δύναμη του 10. Η δύναμη του 10 υπολογίζεται με τέτοιο τρόπο ώστε η απόλυτη τιμή του νέου μέγιστου να είναι μικρότερη από 1. Ο μετασχηματισμός γίνεται σύμφωνα με τη Σχέση 7.3

$$x' = \frac{x}{10^k} \tag{7.3}$$

όπου x η εκάστοτε τιμή της μεταβλητής *A* και *x'* η νέα τιμή.

αρχικές τιμές	min-max	zscore	Δεκαδ. Κλιμ.
7188000	1,0000	2,0950	0,7188
4190200	0,5827	0,9594	0,4190
193964	0,0263	-0,5544	0,0194
44762	0,0055	-0,6109	0,0045
1150289	0,1594	-0,1921	0,1150
5000	0,0000	-0,6259	0,0005
77844	0,0101	-0,5983	0,0078
409298	0,0563	-0,4728	0,0409

M_A	1657419,6
σ_A	2639945,9
max_A	7188000
min_A	5000

Σχήμα 7.2 *Κανονικοποίηση Δεδομένων*

Στο Σχήμα 7.2 παρουσιάζεται παράδειγμα κανονικοποίησης. Στην πρώτη στήλη βρίσκονται οι αρχικές τιμές. Ακριβώς από κάτω παρουσιάζονται η μέση τιμή M_A, η τυπική απόκλιση σ_A, και η μέγιστη και ελάχιστη τιμή. Στη δεύτερη στήλη βρίσκονται οι τιμές κανονικοποιημένες με τη μέθοδο ελάχιστου-μέγιστου. Ως νέο μέγιστο ορίσαμε την τιμή 1 και ως νέο ελάχιστο την τιμή 0. Η προηγούμενη μέγιστη τιμή έχει αντικατασταθεί με την τιμή 1 και η προηγούμενη ελάχιστη τιμή με την τιμή 0. Όλες οι υπόλοιπες τιμές κυμαίνονται εντός της περιοχής τιμών [0..1]. Στην τρίτη στήλη βρίσκονται οι τιμές κανονικοποιημένες κατά z-score. Η μέση τιμή τους ισούται με 0. Στην τελευταία στήλη βρίσκονται οι τιμές κανονικοποιημένες με τη μέθοδο της δεκαδικής κλιμάκωσης.

7.5 Κατασκευή νέων πεδίων

Η ανάλυση των δεδομένων μπορεί εύκολα να οδηγήσει σε εσφαλμένα συμπεράσματα, εάν δεν τηρηθούν ορισμένες προϋποθέσεις. Μια από αυτές είναι η χρήση των «κατάλληλων» δεδομένων, δεδομένων δηλαδή που αποτυπώνουν πραγματικές καταστάσεις πραγμάτων. Για παράδειγμα, κέρδη μερικών εκατομμυρίων ευρώ θεωρούνται εξαιρετικό αποτέλεσμα για μια μικρή επιχείρηση, αλλά μπορεί να θεωρηθούν ανεπαρκές αποτέλεσμα για μια μεγάλη πολυεθνική. Το ύψος των κερδών, ως απόλυτο μέγεθος, δεν είναι επαρκές κριτήριο για την αποτίμηση της κερδοφορίας και της πορείας της επιχείρησης.

Πολλές φορές ο αναλυτής κατασκευάζει νέα πεδία για να μπορέσει να αποδώσει καλύτερα το πραγματικό περιεχόμενο των δεδομένων. Τα δεδομένα των νέων πεδίων υπολογίζονται με κατάλληλες πράξεις από τα δεδομένα άλλων πεδίων. Οι οικονομικοί αριθμοδείκτες, που υπολογίζονται μέσω πράξεων μεταξύ κάποιων οικονομικών μεγεθών, αποτελούν παράδειγμα τέτοιων πεδίων. Πολλοί αριθμοδείκτες υπολογίζονται ως πηλίκα μεταξύ μιας μεταβλητής και ενός χαρακτηριστικού οικονομικού μεγέθους. Η πράξη αυτή επιτυγχάνει την εξομάλυνση των τιμών των μεταβλητών σε σχέση με το επιλεγμένο μέγεθος. Με τον τρόπο αυτό, μπορεί να αναδειχθεί η σημαντικότητα και η κατεύθυνση επιρροής της κάθε μεταβλητής. Μια ποσότητα, που χρησιμοποιείται συχνά ως παρονομαστής για τον υπολογισμό αριθμοδεικτών, είναι το σύνολο ενεργητικού της επιχείρησης. Το σύνολο ενεργητικού μπορεί να θεωρηθεί μέτρο του μεγέθους της επιχείρησης. Στο παράδειγμα που αναφέρθηκε προηγουμένως, η μελέτη του αριθμοδείκτη «κέρδη προς σύνολο ενεργητικού», αντί του απόλυτου ποσού των κερδών, αποδίδει πολύ καλύτερα την ικανότητα της επιχείρησης να παράξει κέρδη.

Η μη χρήση των κατάλληλων δεδομένων μπορεί να οδηγήσει σε τελείως εσφαλμένα συμπεράσματα. Αναφέρουμε ενδεικτικά την περίπτωση όπου σε μελέτη για την ανάπτυξη ενός μοντέλου ικανού να προβλέψει το αποτέλεσμα του εξωτερικού ελέγχου σε επιχειρήσεις, διαπιστώθηκε ότι η μη έκδοση αρνητικών σχολίων είναι θετικά συσχετισμένη με το ύψος των αμοιβών των εξωτερικών ελεγκτών. Επιχειρήσεις που κατέβαλαν μεγαλύτερες αμοιβές στους εξωτερικούς ελεγκτές είχαν λιγότερες πιθανότητες να λάβουν δυσμενή σχόλια. Το αποτέλεσμα αυτό προέκυψε από την ανάλυση του πεδίου Αμοιβές Εξωτερικών Ελεγκτών (Audit Fees). Φυσικά, ένα τέτοιο συμπέρασμα βάλλει ευθέως στην καρδιά του ελεγκτικού συστήματος, καθώς μπορεί να θεωρηθεί ότι οι εξωτερικοί ελεγκτές χρηματίστηκαν. Ωστόσο, οι εξωτερικοί ελεγκτές, όταν καθορίζουν το ύψος

της αμοιβής τους, συνυπολογίζουν και το μέγεθος της επιχείρησης. Για τον έλεγχο μιας μεγάλης επιχείρησης απαιτούνται σαφώς περισσότεροι πόροι, και ως εκ τούτου, η αμοιβή είναι δικαιολογημένα υψηλότερη. Τίθεται πλέον το ερώτημα εάν οι επιχειρήσεις που έχουν μικρότερη πιθανότητα να λάβουν δυσμενή σχόλια είναι μεγαλύτερες (οπότε και πληρώνουν περισσότερα για τον έλεγχο) ή εάν οι επιχειρήσεις που δεν παίρνουν αρνητικά σχόλια κατέβαλλαν δυσανάλογα μεγάλες αμοιβές στους ελεγκτές, πραγματοποιώντας έτσι δωροδοκία. Σημειωτέον ότι οι μεγαλύτερες επιχειρήσεις μπορούν να αντιμετωπίσουν ευκολότερα αρκετά προβλήματα τους και συνήθως διαθέτουν πιο αποτελεσματικούς μηχανισμούς εσωτερικού ελέγχου. Το πείραμα επαναλήφθηκε αφού κανονικοποιήθηκαν οι αμοιβές των εξωτερικών ελεγκτών ως προς το μέγεθος της επιχείρησης, με τη χρήση του αριθμοδείκτη Αμοιβές Εξωτερικών Ελεγκτών / Σύνολο Ενεργητικού (Audit Fees / Total Assets). Το αποτέλεσμα που προέκυψε είναι ακριβώς το αντίθετο. Διαπιστώθηκε ότι οι εταιρείες οι οποίες έλαβαν αρνητικά σχόλια έτειναν να διαθέτουν μεγαλύτερες αμοιβές συγκριτικά με το μέγεθος τους.

7.6 Μείωση Διαστάσεων και Επιλογή Χαρακτηριστικών

Μια από τις βασικότερες ιδιότητες των δεδομένων είναι το πλήθος των στηλών. Οι στήλες αναφέρονται και ως διαστάσεις (dimensions), γνωρίσματα (attributes) και χαρακτηριστικά (features). Η ύπαρξη πολλών διαστάσεων αποτελεί πρόβλημα για τη διαδικασία εξόρυξης προτύπων. Ορισμένες στήλες μπορεί να περιέχουν άσχετη πληροφορία. Άλλες στήλες μπορεί να σχετίζονται μεταξύ τους, έτσι ώστε η ταυτόχρονη παρουσία τους να είναι περιττή. Γιατί αποτελεί πρόβλημα η ύπαρξη πολλών και περιττών στηλών, θα μπορούσε να αναρωτηθεί κανείς. Ορισμένες μέθοδοι ανάλυσης, κυρίως όσες προέρχονται από τη Στατιστική, υποθέτουν ότι στα μοντέλα περιλαμβάνονται μόνο σημαντικές στήλες και ότι οι στήλες αυτές είναι μεταξύ τους ασυσχέτιστες. Άλλες μέθοδοι (όπως οι πλησιέστεροι γείτονες) αντιμετωπίζουν πρόβλημα όταν πρέπει να χειριστούν δεδομένα με πολλά γνωρίσματα. Η ύπαρξη πολλών γνωρισμάτων αυξάνει την πολυπλοκότητα του προβλήματος και προκαλεί καθυστερήσεις στην εκπαίδευση των μοντέλων. Η αύξηση της πολυπλοκότητας και του συνακόλουθου υπολογιστικού κόστους δεν επιφέρει πάντοτε βελτίωση των αποτελεσμάτων και μείωση του λάθους. Από ένα σημείο και μετά η αύξηση του κόστους έχει μηδενική επίπτωση στη μείωση του σφάλματος (Chizi & Maimon, 2005). Το πρόβλημα των διαστάσεων είναι τόσο σημαντικό, ώστε στη βιβλιογραφία αναφέρεται ο όρος «κατάρα των διαστάσεων» (curse of dimensionality).

Μια βάση δεδομένων τυπικά περιέχει εκατοντάδες γνωρίσματα. Αν η ύπαρξη πολλών γνωρισμάτων αποτελεί πρόβλημα, τίθεται ζήτημα μιας μεθόδου για την επιλογή ορισμένων από αυτά. Σε ορισμένες περιπτώσεις η απαλοιφή κάποιων γνωρισμάτων είναι εύκολη και προφανής. Για παράδειγμα, σε ένα πρόβλημα πρόβλεψης χρεοκοπίας, ο αριθμός τηλεφώνου των επιχειρήσεων εμφανώς είναι άσχετος και μπορεί να απομακρυνθεί. Για πολλά όμως γνωρίσματα η επιλογή δεν είναι καθόλου προφανής. Ένας τρόπος αντιμετώπισης του προβλήματος είναι να προσληφθεί κάποιος ειδικός, άριστος γνώστης του ζητήματος που εξετάζεται, και να επιλέξει αυτός τις κατάλληλες μεταβλητές. Ωστόσο, η εύρεση ειδικού δεν είναι πάντα εφικτή και επιπλέον υπάρχουν προβλήματα, στα οποία η συμπεριφορά των δεδομένων δεν είναι εκ των προτέρων γνωστή και δεν υπάρχει προηγούμενη επαρκής γνώση για το ζήτημα. Επίσης, η χρήση ειδικού εισάγει ένα στοιχείο υποκειμενικότητας της κρίσης. Δεν είναι σπάνιες οι περιπτώσεις όπου οι ειδικοί έχουν διαφωνήσει μεταξύ τους. Τίθεται λοιπόν θέμα χρήσης τυπικών μεθόδων για τον περιορισμό των διαστάσεων και την επιλογή χαρακτηριστικών.

Η **μείωση των διαστάσεων** (dimensionality reduction) και η **επιλογή σημαντικών χαρακτηριστικών** (feature selection) δεν είναι ταυτόσημες έννοιες. Για την ακρίβεια, η μείωση των διαστάσεων είναι ευρύτερη έννοια και περιλαμβάνει την επιλογή χαρακτηριστικών. Η επιλογή χαρακτηριστικών συνίσταται στην επιλογή ενός υποσυνόλου Μ χαρακτηριστικών από ένα αρχικό σύνολο N χαρακτηριστικών, όπου $M<N$. Το επιλεγμένο υποσύνολο πρέπει να είναι το πλέον κατάλληλο για την εξόρυξη των προτύπων και να διατηρεί την ουσιαστική πληροφορία σχετικά με τη διασπορά και τη συμπεριφορά των δεδομένων. Μείωση των διαστάσεων μπορεί να επέλθει με την επιλογή χαρακτηριστικών, αλλά και με την προβολή των δεδομένων σε ένα διαφορετικό χώρο λιγότερων διαστάσεων. Ο χώρος αυτός έχει άλλες, διαφορετικές διαστάσεις από τον αρχικό, οι νέες διαστάσεις όμως έχουν καθοριστεί με τέτοιο τρόπο ώστε να διατηρείται ουσιαστική πληροφορία για τη συμπεριφορά των δεδομένων.

Η επιλογή σημαντικών χαρακτηριστικών δεν είναι ένα εύκολο πρόβλημα. Σε ένα σύνολο δεδομένων με N χαρακτηριστικά υπάρχουν 2^N δυνατά υποσύνολα. Αυτό σημαίνει ότι σε περίπτωση που τα δεδομένα έχουν μόλις 10 χαρακτηριστικά, υπάρχουν 1024 δυνατά υποσύνολα, ενώ αν τα χαρακτηριστικά αυξηθούν σε 20 το πλήθος των δυνατών υποσυνόλων αυξάνεται σε 1.048.576. Η άριστη μέθοδος επιλογής χαρακτηριστικών πρέπει να επιλέξει εκείνο το υποσύνολο, το οποίο είναι το πλέον κατάλληλο για το συγκεκριμένο πρόβλημα που εξετάζεται, αλλά και για τη συγκεκριμένη μέθοδο ανάλυσης που θα εφαρμοστεί. Έχουν προταθεί μέθοδοι

που βελτιστοποιούν ταυτόχρονα την επιλογή χαρακτηριστικών και τη ρύθμιση των παραμέτρων της μεθόδου ανάλυσης.

Υπάρχουν πολλές μέθοδοι επιλογής χαρακτηριστικών και διαφέρουν μεταξύ τους κατά ποικίλους τρόπους. Ένας δυνατός διαχωρισμός τους είναι σε μεθόδους τύπου filter και μεθόδους τύπου wrapper. Οι μέθοδοι τύπου **filter** βασίζονται σε χαρακτηριστικά των δεδομένων και χρησιμοποιούν μεθόδους διαφορετικές από τους αλγορίθμους που θα εφαρμοστούν για την τελική εξόρυξη των προτύπων. Χάρη στο γεγονός ότι είναι ανεξάρτητες από τον αλγόριθμο εξόρυξης, οι μέθοδοι αυτές είναι γρήγορες και μπορούν να συνδυαστούν με πολλούς αλγορίθμους. Οι μέθοδοι τύπου **wrapper** χρησιμοποιούν τον ίδιο τον αλγόριθμο εξόρυξης για να αξιολογήσουν τα υποψήφια υποσύνολα χαρακτηριστικών. Με τη βοήθεια μεθόδων τύπου wrapper μπορούν να επιτευχθούν καλύτερα αποτελέσματα, γιατί τα υποσύνολα χαρακτηριστικών είναι προσαρμοσμένα στις μεθόδους που θα χρησιμοποιηθούν για την τελική ανάλυση. Ωστόσο, οι μέθοδοι αυτές είναι σημαντικά βραδύτερες από τις μεθόδους τύπου filter.

Οι μέθοδοι επιλογής σημαντικών χαρακτηριστικών που κατά καιρούς έχουν προταθεί και εφαρμοστεί είναι πολλές. Ορισμένες από τις πλέον γνωστές είναι οι ακόλουθες:

7.6.1 Filters

7.6.1.1 t Στατιστικό Τεστ (t-test) και Ανάλυση Διακύμανσης (Analysis of Variance – ANOVA)

Πρόκειται για «κλασικές» μεθόδους που προέρχονται από τον χώρο της Στατιστικής. Εκτελούν μονομεταβλητή ανάλυση, αποτιμούν δηλαδή τη σημαντικότητα της κάθε μεταβλητής ξεχωριστά, χωρίς να ελέγχουν την πιθανή αλληλεξάρτηση μεταξύ μεταβλητών. Η αναλυτική παρουσίαση των μεθόδων αυτών είναι έξω από τα όρια του παρόντος συγγράμματος. Αναφέρουμε μόνο ότι λαμβάνουν υπόψη τους μέσους όρους και υπολογίζουν σκορ σημαντικότητας των μεταβλητών. Στην περίπτωση της ANOVA υπολογίζεται ο στατιστικός δείκτης ελέγχου F. Όσο σημαντικότερη είναι η μεταβλητή, δηλαδή όσο μεγαλύτερη διαφοροποίηση παρουσιάζουν οι τιμές, ανάλογα με την κατηγορία στην οποία ανήκουν οι παρατηρήσεις, τόσο μεγαλύτερος είναι ο δείκτης F. Αυτές οι μέθοδοι έχουν εφαρμοστεί σε πολλές μελέτες. Σε αρκετές περιπτώσεις έχουν χρησιμοποιηθεί σαν ένα πρώτο στάδιο επιλογής χαρακτηριστικών, το οποίο ακολουθήθηκε από ένα επόμενο στάδιο, όπου εφαρμόστηκαν άλλες μέθοδοι.

Στο Σχήμα 7.3 παρουσιάζονται τα αποτελέσματα Ανάλυσης Διακύμανσης για ένα σύνολο δεδομένων που αφορά τα αποτελέσματα εξωτερικών ελέγχων σε επιχειρήσεις. Από τα 16 αρχικά χαρακτηριστικά επιλέχθηκαν τα 8. Τα χαρακτηριστικά ταξινομούνται κατά φθίνουσα σειρά τιμής F (και αύξουσα σειρά τιμής p). Διαπιστώνουμε ότι η μεταβλητή Quiscore (δείκτης πιστοληπτικής ικανότητας) είναι η πιο σημαντική, δηλαδή στη μεταβλητή αυτή παρουσιάζεται η μεγαλύτερη διαφοροποίηση τιμών ανάλογα με τη λήψη ή μη αρνητικών σχολίων από τους εξωτερικούς ελεγκτές. Ακολουθούν οι άλλες μεταβλητές σε σειρά σημαντικότητας. Η ανάλυση πραγματοποιήθηκε με το ελεύθερο λογισμικό εξόρυξης δεδομένων Tanagra (Rakotomalala, 2005).

	Attributes
1	QUISCORE
2	ROSF
3	ROTA
4	SOLVENCYR
5	ZSCORE
6	IFBIG
7	GEARING
8	AUDITFEE

Calculations details

N°	Attribute	F	F (max normalized)	p-value (1,448)
1	QUISCORE	123,01		0,000000
2	ROSF	71,86		0,000000
3	ROTA	70,37		0,000000
4	SOLVENCYR	23,16		0,000002
5	ZSCORE	18,73		0,000019
6	IFBIG	17,01		0,000044

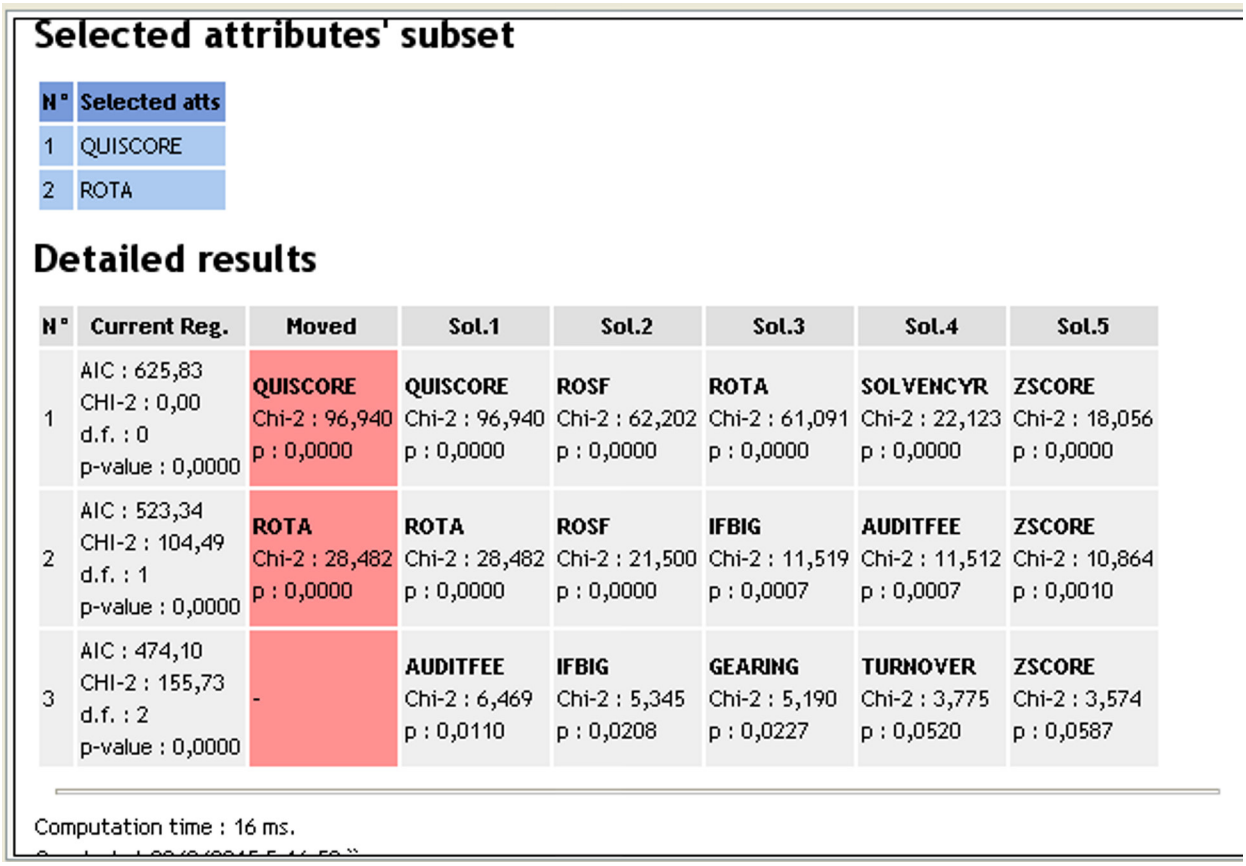
Σχήμα 7.3 *Επιλογή χαρακτηριστικών με ANOVA και δείκτη F*

7.6.1.2 Πρόσθια Επιλογή και Οπίσθια Εξάλειψη (Forward Selection and Backward Elimination)

Έχουν προταθεί διάφορα στατιστικά μέτρα εκτίμησης της σημαντικότητας του εκάστοτε γνωρίσματος. Σε ότι αφορά τον καθορισμό ενός υποσυνόλου γνωρισμάτων, υπάρχουν επίσης πολλές προσεγγίσεις. Η εξαντλητική δοκιμή όλων των δυνατών συνδυασμών δεν είναι εφικτή, αφού το πλήθος τους είναι μεγάλο. Για τον λόγο αυτό, έχουν προταθεί διάφορες ευρετικές μέθοδοι (heuristics), οι οποίες επιδιώκουν ένα «αρκετά καλό» αποτέλεσμα. Δύο τέτοιες μέθοδοι είναι η βηματική πρόσθια επιλογή (stepwise forward selection)και η βηματική οπίσθια εξάλειψη (stepwise backward elimination).

- Η **βηματική πρόσθια επιλογή** ξεκινά με ένα κενό σύνολο επιλεγμένων γνωρισμάτων. Επιλέγει από τα υπόλοιπα γνωρίσματα το πιο σημαντικό, το αφαιρεί από το αρχικό σύνολο και το προσθέτει στο σύνολο των επιλεγμένων γνωρισμάτων. Στη συνέχεια, από τα εναπομείναντα γνωρίσματα, επιλέγει το πιο σημαντικό και το προσθέτει στο σύνολο των επιλεγμένων γνωρισμάτων. Η διαδικασία επαναλαμβάνεται μέχρι να ικανοποιηθεί μια συνθήκη εξόδου.
- Η **βηματική οπίσθια εξάλειψη** αρχικά τοποθετεί όλα τα γνωρίσματα στο σύνολο των επιλεγμένων γνωρισμάτων. Στη συνέχεια, επιλέγει το λιγότερο σημαντικό γνώρισμα και το απομακρύνει από το σύνολο των επιλεγμένων γνωρισμάτων. Η διαδικασία επαναλαμβάνεται μέχρι να ικανοποιηθεί μια συνθήκη εξόδου.

Στο Σχήμα 7.4 παρουσιάζει τα αποτελέσματα βηματικής πρόσθιας επιλογής η οποία βασίζεται στη Λογιστική Παλινδρόμηση. Χρησιμοποιήθηκαν τα ίδια δεδομένα με το προηγούμενο παράδειγμα. Διαπιστώνουμε ότι επιλέχθηκαν δύο μεταβλητές, η Quiscore και η ROTA



Selected attributes' subset

N°	Selected atts
1	QUISCORE
2	ROTA

Detailed results

N°	Current Reg.	Moved	Sol.1	Sol.2	Sol.3	Sol.4	Sol.5
1	AIC : 625,83 CHI-2 : 0,00 d.f. : 0 p-value : 0,0000	**QUISCORE** Chi-2 : 96,940 p : 0,0000	**QUISCORE** Chi-2 : 96,940 p : 0,0000	**ROSF** Chi-2 : 62,202 p : 0,0000	**ROTA** Chi-2 : 61,091 p : 0,0000	**SOLVENCYR** Chi-2 : 22,123 p : 0,0000	**ZSCORE** Chi-2 : 18,056 p : 0,0000
2	AIC : 523,34 CHI-2 : 104,49 d.f. : 1 p-value : 0,0000	**ROTA** Chi-2 : 28,482 p : 0,0000	**ROTA** Chi-2 : 28,482 p : 0,0000	**ROSF** Chi-2 : 21,500 p : 0,0000	**IFBIG** Chi-2 : 11,519 p : 0,0007	**AUDITFEE** Chi-2 : 11,512 p : 0,0007	**ZSCORE** Chi-2 : 10,864 p : 0,0010
3	AIC : 474,10 CHI-2 : 155,73 d.f. : 2 p-value : 0,0000	-	**AUDITFEE** Chi-2 : 6,469 p : 0,0110	**IFBIG** Chi-2 : 5,345 p : 0,0208	**GEARING** Chi-2 : 5,190 p : 0,0227	**TURNOVER** Chi-2 : 3,775 p : 0,0520	**ZSCORE** Chi-2 : 3,574 p : 0,0587

Computation time : 16 ms.

Σχήμα 7.4 *Βηματική πρόσθια επιλογή*

7.6.1.3 Χρήση ενός αλγορίθμου ως φίλτρου για άλλον αλγόριθμο

Ένας άλλος τρόπος επιλογής χαρακτηριστικών συνίσταται στη χρήση ενός αλγορίθμου μηχανικής μάθησης. Ο αλγόριθμος αυτός αξιολογεί τα χαρακτηριστικά και επιλέγει ορισμένα από αυτά. Στη συνέχεια, διαμορφώνεται ένα σύνολο δεδομένων με τα επιλεγμένα χαρακτηριστικά, το οποίο θα χρησιμοποιηθεί από ένα άλλο αλγόριθμο μάθησης. Τα Δένδρα Αποφάσεων (παρουσιάζονται αναλυτικά στο Κεφάλαιο 8) μπορούν να χρησιμοποιηθούν για την επιλογή χαρακτηριστικών. Τα Δένδρα Αποφάσεων τύπου C4.5 αποτελούνται από κόμβους, όπου κάθε κόμβος είναι ένας έλεγχος σε ένα γνώρισμα και κάθε κλαδί είναι ένα αποτέλεσμα του ελέγχου, ενώ τα φύλα είναι αποφάσεις κατηγοριοποίησης. Για την επιλογή χαρακτηριστικών αναπτύσσεται ένα μοντέλο Δένδρου Αποφάσεων. Στο δένδρο συμμετέχουν ορισμένα μόνο γνωρίσματα. Τα γνωρίσματα αυτά θεωρούνται σημαντικά και διατηρούνται, ενώ τα υπόλοιπα εξαλείφονται. Δένδρα Αποφάσεων για την επιλογή σημαντικών χαρακτηριστικών έχουν χρησιμοποιήσει οι Cardie (1993), Min and Jeong (2009) και οι Cho, Hong and Ha (2010). Ένα γνωστό πρόβλημα με τα Δένδρα Αποφάσεων είναι ότι μικρές μεταβολές στο δείγμα μπορούν να οδηγήσουν σε ένα σημαντικά διαφορετικό δένδρο. Και άλλοι αλγόριθμοι, όπως η μέθοδος των Τραχέων Συνόλων - Rough Sets, έχουν χρησιμοποιηθεί για την επιλογή σημαντικών χαρακτηριστικών (Yeh, Chi & Hsu, 2010).

7.6.1.4 Correlation-Based Feature Selection (CFS)

Η μέθοδος CFS (Hall, 1999) ανήκει στην κατηγορία των φίλτρων και εφαρμόζεται για την επιλογή ενός υποσυνόλου χαρακτηριστικών σε προβλήματα κατηγοριοποίησης. Στόχος είναι η εύρεση χαρακτηριστικών τα οποία είναι ισχυρώς συσχετισμένα (correlated) με τη μεταβλητή της κλάσης, αλλά ασθενώς συσχετισμένα μεταξύ τους. Η μέθοδος δηλαδή όχι μόνο επιλέγει σημαντικά χαρακτηριστικά, αλλά ελέγχει και τις μεταξύ τους εξαρτήσεις, ώστε να αποδώσει ένα σύνολο ανεξάρτητων μεταβλητών. Η αποδοχή ενός γνωρίσματος εξαρτάται από την ικανότητα του να προβλέπει την κλάση παρατηρήσεων, που δεν έχουν ήδη κατηγοριοποιηθεί με χρήση των άλλων γνωρισμάτων. Παρέχοντας μια φόρμουλα, η οποία συμπεριλαμβάνει την ποσοτικοποίηση της συσχέτισης μεταξύ γνωρισμάτων και κλάσης, αλλά και την ποσοτικοποίηση της συσχέτισης μεταξύ των

γνωρισμάτων, υπολογίζει την αξία των υποψήφιων υποσυνόλων χαρακτηριστικών και ταξινομεί τα υποσύνολα με βάση την αξία τους.

Η μέθοδος μπορεί να συνδυαστεί με διαφορετικές τεχνικές μέτρησης της συσχέτισης όπως την Relief, την Symmetrical Uncertainty και την Minimum Description Length. Ως προς τη μέθοδο αναζήτησης λύσεων, συνδυάζεται με την πρόσθια επιλογή, την οπίσθια εξάλειψη και με μια ενδιάμεση τεχνική που ονομάζεται best-first. Μια αδυναμία της CFS είναι ότι εφαρμόζεται επί διακριτών δεδομένων. Εάν τα δεδομένα είναι αριθμητικά πρέπει να γίνει διακριτοποίηση. Ένα σημαντικό πλεονέκτημα της μεθόδου είναι το χαμηλό υπολογιστικό κόστος, δηλαδή η μεγάλη ταχύτητα εκτέλεσης. Υλοποιήσεις της CFS υπάρχουν στο ελεύθερο λογισμικό εξόρυξης δεδομένων WEKA, αλλά και σε άλλα λογισμικά.

7.6.2 Wrappers

Οι μέθοδοι τύπου wrapper χρησιμοποιούν για την επιλογή σημαντικών χαρακτηριστικών τον ίδιο τον αλγόριθμο που θα εφαρμοστεί για την τελική εξόρυξη προτύπων. Δεν πρόκειται δηλαδή για ανεξάρτητες μεθόδους που μπορούν να καταγραφούν. Οι μέθοδοι αυτές διαφοροποιούνται ως προς τον αλγόριθμο εξόρυξης και ως προς την τεχνική αναζήτησης λύσεων. Wrappers έχουν χρησιμοποιηθεί για Δένδρα Αποφάσεων, πλησιέστερους γείτονες, αφελείς Μπαϋεσιανούς κατηγοριοποιητές, νευρωνικά δίκτυα, μηχανές διανυσμάτων υποστήριξης κλπ. Μέθοδοι αναζήτησης λύσεων που έχουν χρησιμοποιηθεί είναι η πρόσθια επιλογή, η οπίσθια εξάλειψη, καθώς και ενδιάμεσες τεχνικές. Οι παραπάνω κατηγοριοποιητές και οι μέθοδοι αναζήτησης λύσεων έχουν χρησιμοποιηθεί σε διάφορους συνδυασμούς.

Μια ισχυρή τάση η οποία εμφανίστηκε πρόσφατα σε προβλήματα κατηγοριοποίησης, είναι η δημιουργία υβριδικών κατηγοριοποιητών, κατηγοριοποιητών δηλαδή που συνδυάζουν δύο διαφορετικές μεθόδους. Σε πολλές περιπτώσεις, μια κλασσική μέθοδος κατηγοριοποίησης, όπως πχ ένα νευρωνικό δίκτυο, συνδυάζεται με εξελικτικό αλγόριθμο, πχ γενετικό αλγόριθμο. Στόχος είναι, μέσω του εξελικτικού αλγορίθμου, να βελτιστοποιηθεί η ρύθμιση των παραμέτρων του κατηγοριοποιητή. Σε ορισμένες περιπτώσεις, ο εξελικτικός αλγόριθμος χρησιμοποιείται, εκτός από τη ρύθμιση των παραμέτρων, και για την ταυτόχρονη επιλογή σημαντικών χαρακτηριστικών. Οι Chen et al. (2011), σε μια μελέτη πρόβλεψης χρεοκοπίας, εφαρμόζουν μια μέθοδο k-πλησιέστερων γειτόνων με ασαφή λογική (fuzzy k-nearest neighbor). Για τον βέλτιστο υπολογισμό του πλήθους των πλησιέστερων γειτόνων, αλλά και για τον καθορισμό της παραμέτρου ασάφειας χρησιμοποιούν τον εξελικτικό αλγόριθμο Particle Swarm Optimization (PSO). Ο PSO χρησιμοποιείται ταυτόχρονα για την εύρεση ενός καλού υποσυνόλου χαρακτηριστικών, που θα χρησιμοποιηθούν για την κατηγοριοποίηση. Οι Chen, Ribeiro, Vieira, Duarte and Neves (2011), σε μια άλλη μελέτη πρόβλεψης χρεοκοπίας, συνδυάζουν τη μέθοδο κατηγοριοποίησης Learning Vector Quantization με γενετικούς αλγορίθμους. Η μέθοδος των γενετικών αλγορίθμων χρησιμοποιείται για να οριστούν το πλήθος των πρωτοτύπων και τα βάρη του μοντέλου, αλλά και για να επιλεγούν ταυτόχρονα τα σημαντικά χαρακτηριστικά. Το χρωμόσωμα που χρησιμοποιήθηκε περιλαμβάνει d δυαδικά γονίδια που υποδηλώνουν την ύπαρξη ή απουσία καθενός από τα d γνωρίσματα, ένα γονίδιο για το πλήθος των προτύπων και έναν αριθμό γονιδίων για τα βάρη των προτύπων.

7.6.3 Ανάλυση Κυρίων Συνιστωσών (Principal Components Analysis)

Η Ανάλυση Κυρίων Συνιστωσών (ΑΚΣ) (Principal Components Analysis (PCA)) είναι μια μέθοδος συμπίεσης των δεδομένων, η οποία μας επιτρέπει να μειώσουμε το πλήθος των διαστάσεων ενός συνόλου δεδομένων. Θεωρούμε ότι έχουμε ένα σύνολο δεδομένων με K γραμμές και N διαστάσεις (στήλες). Με την ΑΚΣ βρίσκουμε ένα σύστημα M κάθετων διανυσμάτων, όπου $M < N$, και προβάλλουμε τα δεδομένα στον νέο χώρο M διαστάσεων. Με τον τρόπο αυτό δημιουργούμε γραμμικούς συνδυασμούς των αρχικών μεταβλητών οι οποίοι:

- είναι ασυσχέτιστοι μεταξύ τους,
- περιέχουν το μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών.

Για την καλύτερη κατανόηση της μεθόδου παραθέτουμε σύντομα μερικές βασικές έννοιες.

Ας θεωρήσουμε ένα σύνολο δεδομένων 2 διαστάσεων, που απεικονίζονται σε έναν επίπεδο χώρο με άξονες X και Y. Το σύνολο περιλαμβάνει K παρατηρήσεις. Η **συνδιασπορά** (covariance) ορίζεται από την Εξίσωση 7.4

$$cov(X,Y) = \frac{\sum_{i=1}^{K}(X_i - \bar{X})\,(Y_i - \bar{Y})}{K-1}$$

(7.4)

Η συνδιασπορά υπολογίζεται πάντα ανάμεσα σε δύο διαστάσεις και αποτελεί μέτρο του πως μεταβάλλεται η μια μεταβλητή σε σχέση με την άλλη. Εάν η συνδιασπορά είναι θετική, τότε οι δύο μεταβλητές αυξάνονται ταυτόχρονα, δηλαδή όταν αυξάνονται οι τιμές της μιας μεταβλητής αυξάνονται και της άλλης. Αντιθέτως, όταν η συνδιασπορά είναι αρνητική, τότε καθώς αυξάνονται οι τιμές της μιας μεταβλητής μειώνονται οι τιμές της άλλης.

Αν έχουμε περισσότερες από δύο μεταβλητές μπορούμε να υπολογίσουμε πολλές τιμές συνδιασποράς, μια για κάθε ζευγάρι μεταβλητών. Όλες οι δυνατές τιμές συνδιασποράς καταγράφονται στον **πίνακα συνδιασποράς** (covariance matrix**).** Για τρεις μεταβλητές *X,Y,Z* ο πίνακας συνδιασποράς είναι ο ακόλουθος

$$C = \begin{pmatrix} cov(X,X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & cov(Y,Y) & cov(Y,Z) \\ cov(Z,X) & cov(Z,Y) & cov(Z,Z) \end{pmatrix}$$

(7.5)

Δεδομένου ότι *cov(X,Y) = cov (Y,X)*, ο πίνακας είναι συμμετρικός ως προς τη διαγώνιο του. Επίσης, η συνδιασπορά μιας μεταβλητής με τον εαυτό της είναι η διασπορά της. Με βάση τα παραπάνω, ουσιαστική πληροφορία περί συνδιασποράς υπάρχει μόνο στις σκιασμένες τιμές.

Ας θεωρήσουμε τώρα τον πολλαπλασιασμό πινάκων και ειδικότερα τον πολλαπλασιασμό ενός δισδιάστατου πίνακα με ένα διάνυσμα σύμφωνα με το ακόλουθο παράδειγμα:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 13 \\ 7 \end{pmatrix}$$

(7.6)

Ο δισδιάστατος πίνακας μπορεί να θεωρηθεί ως ένας πίνακας μετασχηματισμού του αρχικού διανύσματος. Το αποτέλεσμα του πολλαπλασιασμού είναι ένα διάνυσμα, το οποίο δεν είναι ίσο ή ακέραιο πολλαπλάσιο του αρχικού διανύσματος. Ας θεωρήσουμε τώρα τον πολλαπλασιασμό του ίδιου πίνακα με ένα άλλο διάνυσμα.

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

(7.7)

Στην περίπτωση αυτή, το αποτέλεσμα του μετασχηματισμού είναι ένα διάνυσμα, το οποίο είναι ακέραιο πολλαπλάσιο του αρχικού, δηλαδή ένα διάνυσμα ίδιας διεύθυνσης και διαφορετικού μήκους. Το διάνυσμα αυτό είναι ένα **ιδιοδιάνυσμα** (eigenvector) του δισδιάστατου πίνακα μετασχηματισμού και η τιμή πολλαπλασιασμού του, στο παράδειγμα μας η τιμή 4, είναι η **ιδιοτιμή (**eigenvalue). Με μαθηματικές σχέσεις αυτό καταγράφεται ως εξής:

$$A \times v = \lambda \times v$$

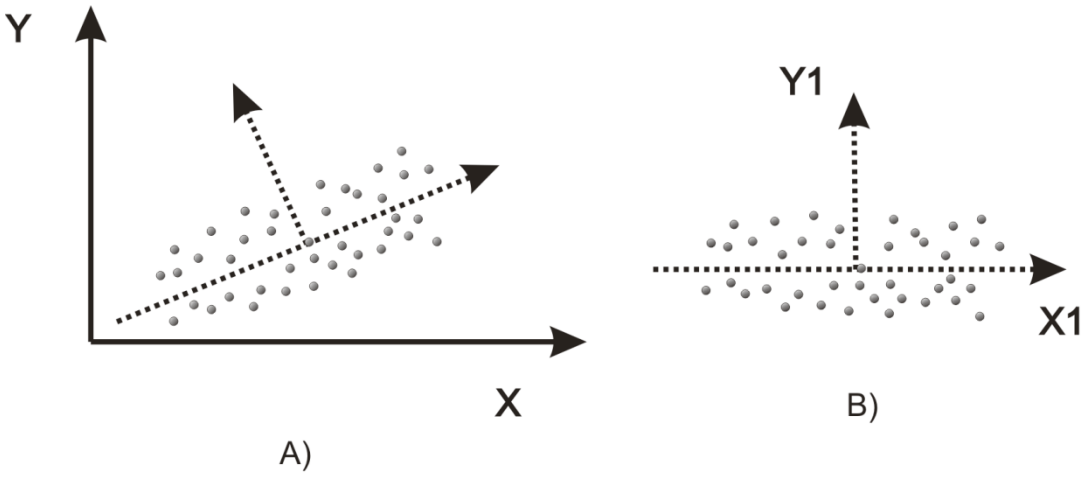
(7.8)

όπου *A* ο πίνακας μετασχηματισμού, *v* το ιδιοδιάνυσμα *και λ* η ιδιοτιμή. Μόνον τετράγωνοι πίνακες έχουν ιδιοδιανύσματα, αλλά δεν έχουν ιδιοδιανύσματα όλοι οι τετράγωνοι πίνακες. Ένας πίνακας *N X N* με ιδιοδιανύσματα, θα έχει *N* ιδιοδιανύσματα. Βασική ιδιότητα των ιδιοδιανυσμάτων ενός πίνακα είναι ότι μεταξύ τους είναι ορθογώνια.

Η **Ανάλυση Κυρίων Συνιστωσών** κάνει χρήση του πίνακα συνδιασποράς και των ιδιοδιανυσμάτων. Η διαδικασία της ΑΚΣ ακολουθεί τα παρακάτω βήματα:

- Για κάθε μεταβλητή (στήλη) υπολογίζεται η μέση τιμή και αφαιρείται από όλες τις τιμές της. Με τον τρόπο αυτό η μέση τιμή κάθε μεταβλητής γίνεται μηδέν.
- Για τον πίνακα με τις νέες τιμές υπολογίζεται ο πίνακας συνδιασποράς.
- Υπολογίζονται τα ιδιοδιανύσματα του πίνακα συνδιασποράς. Τα ιδιοδιανύσματα αυτά έχουν μήκος ίσο με ένα. Τα ιδιοδιανύσματα ονομάζονται **κύριες συνιστώσες** (principal components) και τα αρχικά δεδομένα μπορούν να εκφραστούν ως γραμμικοί συνδυασμοί των κυρίων συνιστωσών. Δηλαδή, οι βασικές συνιστώσες μπορούν να αποτελέσουν ένα νέο σύστημα αξόνων για τα δεδομένα.
- Τα ιδιοδιανύσματα ταξινομούνται με βάση τις ιδιοτιμές τους. Οι ιδιοτιμές αποτελούν μέτρο σημαντικότητας των ιδιοδιανυσμάτων. Όσο μεγαλύτερη είναι η ιδιοτιμή, τόσο σημαντικότερο είναι το ιδιοδιάνυσμα ως συνιστώσα αναπαράστασης των δεδομένων, δηλαδή η συνιστώσα αυτή περιλαμβάνει περισσότερη πληροφορία σχετικά με τη διασπορά των δεδομένων.
- Αφού ταξινομηθούν οι βασικές συνιστώσες σε σειρά σημαντικότητας, μπορούν να αφαιρεθούν οι λιγότερο σημαντικές. Αν οι ιδιοτιμές τους είναι μικρές δεν έχουμε σημαντική απώλεια πληροφορίας.
- Τα δεδομένα μπορούν να επανακαθοριστούν με βάση το νέο σύστημα αξόνων. Αν έχουμε διατηρήσει όλες τις βασικές συνιστώσες, τότε μπορούμε να ξανακατασκευάσουμε με ακρίβεια τα αρχικά δεδομένα, προβάλλοντας τα στο αρχικό σύστημα αξόνων. Αν έχουμε επιλέξει μόνον ορισμένες, τις σημαντικότερες, βασικές συνιστώσες μπορούμε να κατασκευάσουμε μια ικανοποιητική προσέγγιση των αρχικών δεδομένων.

Στο Σχήμα 7.5 βλέπουμε ένα παράδειγμα υπολογισμού των κυρίων συνιστωσών και έκφρασης των δεδομένων στο νέο σύστημα αξόνων. Στο τμήμα Α) του σχήματος παρουσιάζονται τα δεδομένα στο αρχικό σύστημα αξόνων *X* και *Y*. Υπολογίζονται οι κύριες συνιστώσες και απεικονίζονται σαν βέλη με διακεκομμένες γραμμές. Στο τμήμα Β) τα δεδομένα απεικονίζονται στο σύστημα αξόνων των κυρίων συνιστωσών. Παρατηρούμε ότι οι προβολές των σημείων στον άξονα *X1* περιέχουν πολύ πληροφορία σχετικά με τη διασπορά τους ενώ οι προβολές τους στον άξονα *Y1* περιέχουν σημαντικά λιγότερη πληροφορία. Αν εξαλείψουμε τον άξονα *Y1* και διατηρήσουμε μόνο τις τιμές των σημείων σε σχέση με τον άξονα *X1*, θα έχουμε διατηρήσει σημαντικό μέρος της πληροφορίας σχετικά με τη διασπορά των δεδομένων.



Σχήμα 7.5 *Κύριες Συνιστώσες*

Η κάθε κύρια συνιστώσα ορίζεται ως ένας γραμμικός συνδυασμός των αρχικών μεταβλητών. Αν για παράδειγμα, τα δεδομένα μας έχουν *N* μεταβλητές $(x_1,\ldots,x_n)$, τότε η κύρια συνιστώσα *i* έχει τη μορφή

$$PC_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \quad (7.9)$$

Οι συντελεστές a_{ik} καθορίζουν τον βαθμό στον οποίο κάθε μεταβλητή επηρεάζει την εκάστοτε βασική

συνιστώσα (loadings). Μια ενδιαφέρουσα οπτική αναπαράσταση του τρόπου με τον οποίο υπολογίζονται οι βασικές συνιστώσες μπορεί να βρει ο αναγνώστης στην ιστοσελίδα Principal Components Analysis explained visually (Explained Visually, 2015).

Η ανάλυση κυρίων συνιστωσών αποτελεί μέθοδο μείωσης των διαστάσεων των δεδομένων, με προβολή τους σε έναν χώρο λιγότερων, αλλά διαφορετικών, διαστάσεων και όχι μέθοδο επιλογής σημαντικών χαρακτηριστικών, τουλάχιστον με την έννοια της επιλογής ορισμένων από τις αρχικές διαστάσεις και της απόρριψης των υπόλοιπων. Ωστόσο, έχουν προταθεί τεχνικές που αξιοποιούν τους παράγοντες βαρύτητας των αρχικών μεταβλητών στις κύριες συνιστώσες (loadings), καθώς και το ποσοστό διακύμανσης που αντιπροσωπεύουν (communality), για τη χρήση της ΑΚΣ ως μεθόδου επιλογής χαρακτηριστικών. Παραδείγματα εφαρμογής της ΑΚΣ για την επιλογή σημαντικών χαρακτηριστικών υπάρχουν στο Shie, Chen and Liu (2012) και στο Chen (2011).

7.6.4 Επιλογή Χαρακτηριστικών – Συμπεράσματα

Όπως εύκολα μπορεί να διαπιστώσει ο αναγνώστης από τα περιεχόμενα αυτού του υποκεφαλαίου, υπάρχουν πολλές μέθοδοι επιλογής σημαντικών χαρακτηριστικών. Αναπόφευκτα τίθεται το ερώτημα ποια από όλες αυτές τις μεθόδους είναι η καλύτερη. Δυστυχώς στο ερώτημα αυτό δεν υπάρχει μια οριστική απάντηση. Το ζήτημα της επιλογής σημαντικών χαρακτηριστικών είναι ένα ανοικτό πεδίο έρευνας και διαρκώς προτείνονται νέες τεχνικές. Έχουν προταθεί πολλές μέθοδοι τύπου filter, ενώ η δημιουργία μεθόδων τύπου wrapper είναι ανοικτή σε κάθε είδους παραλλαγές. Σε εργασίες εξόρυξης δεδομένων και μηχανικής μάθησης δεν διαφαίνεται κάποια επικρατούσα τάση για την επιλογή μιας συγκεκριμένης μεθόδου. Χαρακτηριστικά αναφέρουμε ότι ο Kirkos (2015), σε μια μελέτη αποτίμησης σαράντα δύο ερευνητικών εργασιών με θέμα την ευφυή πρόβλεψη χρεοκοπίας, διαπίστωσε ότι σχεδόν σε κάθε εργασία εφαρμόστηκε διαφορετική μέθοδος επιλογής χαρακτηριστικών.

Για τον χρήστη, ο οποίος ενδιαφέρεται να πραγματοποιήσει οικονομικές αναλύσεις με τη χρήση μεθόδων εξόρυξης δεδομένων, αναφέρουμε ότι η εφαρμογή περίτεχνων μεθόδων τύπου wrapper απαιτεί ικανότητες προγραμματισμού και γνώση κάποιας σχετικής γλώσσας προγραμματισμού, όπως η R ή το Matlab. Λογισμικά εξόρυξης δεδομένων προσφέρουν έτοιμες σχετικά απλές μεθόδους τύπου wrapper και αρκετές μεθόδους τύπου filter. Κατά κανόνα, η εφαρμογή μιας απλής wrapper ή μιας εξελιγμένης filter, όπως η CFS, είναι αρκετή για την ανάπτυξη μοντέλων με πολύ ικανοποιητικές επιδόσεις.

7.7 Διακριτοποίηση

Η διακριτοποίηση (discretization) είναι μια διαδικασία μετασχηματισμού των δεδομένων. Για την ακρίβεια, είναι η διαδικασία μετατροπής αριθμητικών δεδομένων σε ονομαστικά δεδομένα, δεδομένα δηλαδή που οι τιμές τους αποτελούνται από ονομαστικές τιμές – λέξεις. Εναλλακτικά μπορούμε να πούμε ότι η διακριτοποίηση είναι η μετατροπή ποσοτικών σε ποιοτικά δεδομένα. Κατά κανόνα, τα αριθμητικά δεδομένα χωρίζονται σε περιοχές τιμών και δημιουργούνται νέες στήλες, όπου στη θέση της αριθμητικής τιμής εισάγεται το όνομα της περιοχής της τιμής.

Υπάρχουν πολλοί λόγοι για να διακριτοποιήσει κανείς τα δεδομένα του. Καταρχάς, ορισμένες μέθοδοι εξόρυξης δέχονται σαν είσοδο μόνο διακριτά δεδομένα. Σε περίπτωση που ο χρήστης θέλει να εφαρμόσει αυτές τις μεθόδους, είναι υποχρεωμένος να κάνει διακριτοποίηση. Επιπλέον, η διακριτοποίηση των δεδομένων μπορεί να επιταχύνει τη διαδικασία εκπαίδευσης των μοντέλων και να βελτιώσει τις επιδόσεις τους, αυξάνοντας έτσι την αποτελεσματικότητα και την αποδοτικότητα (Frank & Witten, 1999). Τέλος, η διακριτοποίηση μπορεί να οδηγήσει σε αποτελέσματα που είναι πιο κατανοητά. Για όλους αυτούς τους λόγους, η διακριτοποίηση έχει αποτελέσει αντικείμενο έρευνας, η οποία απέδωσε διάφορες τεχνικές.

Οι μέθοδοι διακριτοποίησης μπορούν να κατηγοριοποιηθούν με ποικίλους τρόπους:

- **Επιβλεπόμενες και μη επιβλεπόμενες μέθοδοι** (supervised and unsupervised methods). Οι επιβλεπόμενες μέθοδοι εφαρμόζονται σε δεδομένα, όπου μια στήλη ορίζει την κλάση (κατηγορία) των παρατηρήσεων. Οι μέθοδοι αυτές κάνουν χρήση της κλάσης των παρατηρήσεων για να ορίσουν τα διαστήματα τιμών. Για τον λόγο αυτό, οι επιβλεπόμενες μέθοδοι ενδείκνυται να χρησιμοποιούνται σε προβλήματα κατηγοριοποίησης. Οι μη επιβλεπόμενες μέθοδοι δεν χρησιμοποιούν την κλάση των παρατηρήσεων για τον καθορισμό των διαστημάτων.
- **Μονομεταβλητές και Πολυμεταβλητές μέθοδοι** (univariate and multivariate methods). Στις μο-

νομεταβλητές μεθόδους κάθε στήλη διακριτοποιείται ξεχωριστά, χωρίς να λαμβάνονται υπόψη οι τιμές των άλλων στηλών. Στις πολυμεταβλητές μεθόδους εξετάζονται σχέσεις μεταξύ περισσότερων στηλών.

- **Παραμετρικές και μη παραμετρικές μέθοδοι**. Οι παραμετρικές μέθοδοι απαιτούν τον καθορισμό κάποιας παραμέτρου από τον χρήστη. Συνήθως η παράμετρος αυτή είναι το πλήθος των διαστημάτων. Οι μη παραμετρικές μέθοδοι δεν απαιτούν καθορισμό τιμών παραμέτρων και η πληροφορία που απαιτείται για τη διακριτοποίηση αντλείται από τα δεδομένα.
- **Ιεραρχικές και μη Ιεραρχικές**. Οι ιεραρχικές μέθοδοι υλοποιούν μια σταδιακή διαδικασία διαδοχικών διαιρέσεων ή συγχωνεύσεων που αντιστοιχεί σε μια ιεράρχηση. Στις διαιρετικές μεθόδους, αρχικά, όλο το φάσμα των τιμών θεωρείται ένα διάστημα και στη συνέχεια, πραγματοποιούνται διαδοχικοί διαχωρισμοί. Στις μεθόδους συγχώνευσης αρχικά κάθε τιμή θεωρείται ένα ξεχωριστό διάστημα και ακολουθεί μια διαδικασία διαδοχικών συγχωνεύσεων. Στις μη ιεραρχικές μεθόδους δεν δημιουργείται μια ιεραρχία διαστημάτων.

Ορισμένες από τις τεχνικές διακριτοποίησης που προτάθηκαν και χρησιμοποιούνται είναι οι ακόλουθες:

Διαστήματα ίσου πλάτους (equal width discretization). Στη διακριτοποίηση με διαστήματα ίσου πλάτους ο χρήστης προκαθορίζει το πλήθος των διαστημάτων *k*. Στη συνέχεια εντοπίζονται η μικρότερη και η μεγαλύτερη τιμή της μεταβλητής που θα διακριτοποιηθεί *x* (*xmin* και *xmax* αντίστοιχα) και καθορίζονται περιοχές τιμών πλάτους *w=xmax-xmin/k*. Με τον τρόπο αυτό ορίζονται *k* διαδοχικά διαστήματα. Στα δεδομένα δημιουργείται μια καινούργια στήλη x_discr και για κάθε αριθμητική τιμή της *x* τοποθετείται στην x_discr η περιγραφή της περιοχής τιμών στην οποία ανήκει η αριθμητική τιμή.

Διαστήματα ίσης συχνότητας (equal frequency discretization). Στη διακριτοποίηση με διαστήματα ίσης συχνότητας ο χρήστης προκαθορίζει το πλήθος των διαστημάτων *k*. Στη συνέχεια, οι τιμές της μεταβλητής x η οποία θα διακριτοποιηθεί ταξινομούνται σε αύξουσα σειρά και χωρίζονται σε *k* περιοχές, έτσι ώστε όλες οι περιοχές να έχουν ίσο πλήθος τιμών. Στα δεδομένα δημιουργείται μια καινούργια στήλη x_discr και για κάθε αριθμητική τιμή της x τοποθετείται στην x_discr η περιγραφή της περιοχής τιμών στην οποία ανήκει η αριθμητική τιμή. Η διακριτοποίηση με διαστήματα ίσης συχνότητας, καθώς και η διακριτοποίηση με διαστήματα ίσου πλάτους είναι μη επιβλεπόμενες και μονομεταβλητές μέθοδοι.

Διακριτοποίηση βασισμένη στην Εντροπία (entropy based discretization). Η μέθοδος αυτή είναι ριζικά διαφορετική από τις δύο προηγούμενες. Ο χρήστης δεν προκαθορίζει το πλήθος των διαστημάτων, αλλά αυτό υπολογίζεται από τον αλγόριθμο. Η μέθοδος είναι κατάλληλη για δεδομένα, όπου μια στήλη ορίζει την κατηγορία (κλάση), στην οποία ανήκουν οι παρατηρήσεις, δηλαδή για δεδομένα κατάλληλα για κατηγοριοποίηση. Για τον καθορισμό των διαστημάτων γίνεται χρήση της στατιστικής Εντροπίας. Η Εντροπία είναι ένα μέτρο του βαθμού αταξίας ενός συστήματος. Ο καθορισμός των διαστημάτων γίνεται έτσι ώστε τα υποσύνολα των παρατηρήσεων που θα προκύψουν να έχουν αθροιστικά μικρότερη εντροπία από το σύνολο των αρχικών δεδομένων. Αυτό σημαίνει ότι με τη διακριτοποίηση μεταβαίνουμε σε υποσύνολα υψηλότερης τάξης. Ο βαθμός τάξης σχετίζεται με την κλάση των παρατηρήσεων, δηλαδή οι παρατηρήσεις των υποσυνόλων έχουν μεγαλύτερη ομοιότητα ως προς την τιμή της κλάσης τους. Η στατιστική Εντροπία και το Κέρδος Πληροφορίας παρουσιάζονται αναλυτικότερα στο Κεφάλαιο 9, το οποίο αναφέρεται στα Δένδρα Αποφάσεων. Σύμφωνα με τον αλγόριθμο που προτάθηκε από τους Fayyad and Irani (1993), για κάθε ζευγάρι διαδοχικών τιμών υπολογίζεται η μέση τιμή *m* και ελέγχονται τα δύο υποσύνολα που προκύπτουν (παρατηρήσεις με τιμή στο συγκεκριμένο πεδίο < *m* και παρατηρήσεις με τιμή > *m*) ως προς τη συνολική εντροπία τους. Η ενδιάμεση τιμή *m* που θα αποφέρει υποσύνολα ελάχιστης συνολικής εντροπίας, θα επιλεγεί ως τιμή διαχωρισμού. Η διαδικασία αυτή είναι επαναλαμβανόμενη. Το γεγονός ότι η μέθοδος χρησιμοποιεί την τιμή της κλάσης για να ορίσει τα διαστήματα, καθιστά περισσότερο πιθανό τον καθορισμό διαστημάτων που θα διευκολύνουν τους αλγόριθμους κατηγοριοποίησης. Στο Σχήμα 7.6 παρουσιάζονται δεδομένα με διακριτοποίηση ίσου πλάτους (d_eqW_ROTA_1), διακριτοποίηση ίσης συχνότητας (d_eqF_ROTA_1) και διακριτοποίηση βασισμένη στην εντροπία (d_mdlpc_ROTA_1). Διακριτοποιείται ο αριθμοδείκτης ROTA (Return On Total Assets). Τα δεδομένα ανήκουν σε πραγματικές εταιρίες και για την επιβλεπόμενη διακριτοποίηση χρησιμοποιήθηκε ως τιμή κλάσης η λήψη ή μη λήψη αρνητικών σχολίων από τους εξωτερικούς ελεγκτές.

ROTA	d_eqW_ROTA_1	d_eqF_ROTA_1	d_mdlpc_ROTA_1
-6,29	-121,56399536_=<_m_<_78,26800537	-18,88999939_=<_m_<_-0,82999998	-33,13999939_=<_m_<_1,30999994
-59,36	-121,56399536_=<_m_<_78,26800537	-59,43000031_=<_m_<_-18,88999939	m_<_-33,13999939
-0,27	-121,56399536_=<_m_<_78,26800537	-0,82999998_=<_m_<_6,96999979	-33,13999939_=<_m_<_1,30999994
3,29	-121,56399536_=<_m_<_78,26800537	-0,82999998_=<_m_<_6,96999979	m_>=_1,30999994
-198,15	-321,39599609_=<_m_<_-121,56399536	m_<_-59,43000031	m_<_-33,13999939
278,1	m_>=_78,26800537	m_>=_6,96999979	m_>=_1,30999994
-49,05	-121,56399536_=<_m_<_78,26800537	-59,43000031_=<_m_<_-18,88999939	m_<_-33,13999939
-132,17	-321,39599609_=<_m_<_-121,56399536	m_<_-59,43000031	m_<_-33,13999939
-97,41	-121,56399536_=<_m_<_78,26800537	m_<_-59,43000031	m_<_-33,13999939

Σχήμα 7.6 *Διακριτοποίηση Δεδομένων.*

Διακριτοποίηση βασισμένη στην Ανάλυση Συστάδων (cluster based discretization). Η μέθοδος αυτή ανήκει στην κατηγορία των πολυμεταβλητών μεθόδων και χρησιμοποιεί για τη διακριτοποίηση τεχνικές από την Ανάλυση Συστάδων. Λεπτομέρειες για την Ανάλυση Συστάδων μπορεί να αναζητήσει ο αναγνώστης στο Κεφάλαιο 11. Οι Chmielewski and Grzymala–Busse (1996) πρότειναν έναν αλγόριθμο δύο σταδίων. Στο πρώτο στάδιο, τα δεδομένα τα οποία διαθέτουν n γνωρίσματα θεωρούνται ως σημεία σε έναν χώρο n διαστάσεων. Χρησιμοποιώντας τη μέθοδο της διαμέσου (median method) και την Ευκλείδεια απόσταση ως μέτρο ομοιότητας, δημιουργούνται συστάδες. Αρχικά, κάθε παρατήρηση θεωρείται ότι αποτελεί μια συστάδα και στη συνέχεια, ακολουθούν διαδοχικές συγχωνεύσεις συστάδων. Οι συγχωνεύσεις συστάδων συνεχίζονται μέχρι το επίπεδο συνοχής των συστάδων να γίνει ίσο η μεγαλύτερο από το επίπεδο συνοχής των αρχικών δεδομένων. Αφού ολοκληρωθεί η δημιουργία των συστάδων, για κάθε διάσταση ελέγχονται τα σημεία των συστάδων, ώστε να βρεθεί η μικρότερη και η μεγαλύτερη τιμή. Οι δύο αυτές τιμές ορίζουν το διάστημα μέσα στο οποίο βρίσκονται όλα τα μέλη της συστάδας. Αν το διάστημα μιας συστάδας εμπεριέχεται στο διάστημα άλλης συστάδας, τότε το εσωτερικό διάστημα εξαλείφεται. Οι συστάδες αναλύονται σε σχέση με όλες τις διαστάσεις, ώστε να βρεθούν σημεία τομής για κάθε διάσταση ταυτόχρονα. Στο σημείο αυτό ολοκληρώνεται το πρώτο στάδιο. Κατά τη διάρκεια του δεύτερου σταδίου, γειτονικά διαστήματα μιας διάστασης ελέγχονται ως προς τη δυνατότητα συγχώνευσης. Κριτήριο για τη συγχώνευση διαστημάτων είναι η στατιστική εντροπία. Η διαδικασία συγχωνεύσεων επαναλαμβάνεται, έως ότου κάθε δυνατό ζεύγος γειτονικών διαστημάτων θεωρηθεί μη συγχωνεύσιμο.

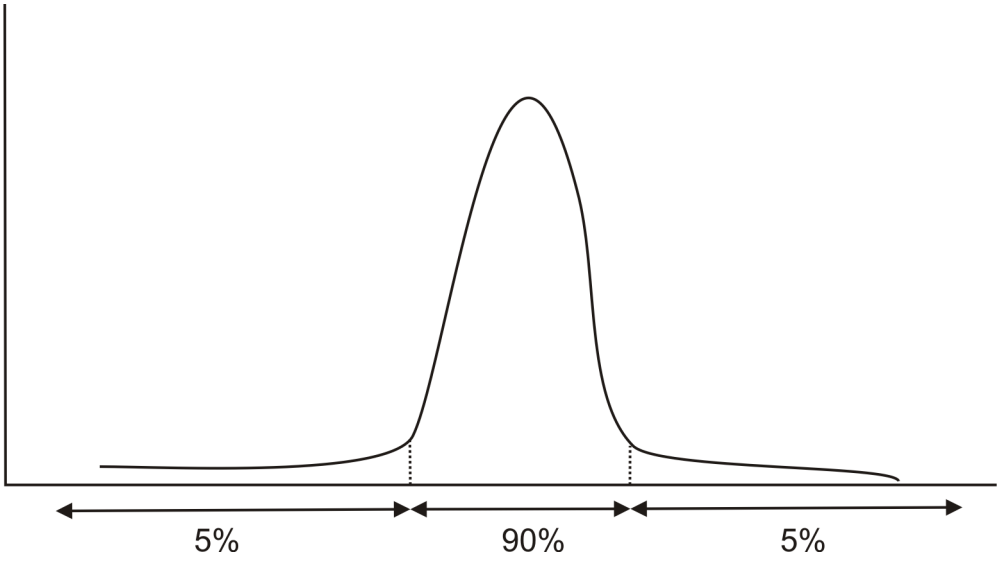
Τμηματοποίηση με φυσική κατάτμηση (segmentation by natural partitioning). Με τη φυσική κατάτμηση επιτυγχάνουμε τη δημιουργία τμημάτων, που είναι πιο κοντά στον τρόπο με τον οποίο σκέφτεται ο άνθρωπος και έχουν πιο «φυσικές» τιμές διαχωρισμού. Για παράδειγμα, είναι πιο «φυσικό» να ορίσουμε μια περιοχή τιμών 2000..3000 παρά μια περιοχή τιμών 2011..2926. Σημειωτέον ότι περίτεχνοι αλγόριθμοι διακριτοποίησης ορίζουν τις περιοχές τιμών ακόμα και με ακρίβεια δεκαδικού ψηφίου.

Η μέθοδος 3-4-5 είναι ένας εύκολος και πρακτικός τρόπος να ορίζουμε «φυσικές» περιοχές τιμών για αριθμητικά δεδομένα. Ο τρόπος καθορισμού των διαστημάτων είναι ο ακόλουθος:

- Εάν μια περιοχή τιμών καλύπτει 3,6,7 ή 9 διαφορετικές τιμές του πιο σημαντικού ψηφίου, τότε δημιουργούνται 3 διαστήματα. Αν καλύπτει 3,6 ή 9 διαφορετικές τιμές, τότε δημιουργούνται 3 διαστήματα ίσου πλάτους. Αν καλύπτει 7 διαφορετικές τιμές, τότε δημιουργούνται τρία διαστήματα όπου το πρώτο καλύπτει τις δύο μικρότερες τιμές, το επόμενο καλύπτει τις τρεις επόμενες τιμές και το τελευταίο διάστημα καλύπτει τις 2 τελευταίες τιμές. Για παράδειγμα, η περιοχή 1024..7512 καλύπτει 7 διαφορετικές τιμές για το πιο σημαντικό ψηφίο (οι αριθμοί από 1 έως 7) και θα χωριστεί στα διαστήματα 1001..3000, 3001..6000 και 6001..8000.
- Εάν μια περιοχή τιμών καλύπτει 2,4 ή 8 διαφορετικές τιμές του πιο σημαντικού ψηφίου, τότε δημιουργούνται 4 διαστήματα ίσου πλάτους. Για παράδειγμα, μια περιοχή τιμών 1024..4512 καλύπτει τέσσερεις διαφορετικές τιμές για το πιο σημαντικό ψηφίο και θα χωριστεί στα διαστήματα 1001..2000, 2001..3000, 3001..4000 και 4001..5000.
- Εάν μια περιοχή τιμών καλύπτει 1,5 ή 10 διαφορετικές τιμές του πιο σημαντικού ψηφίου, τότε δημιουργούνται 5 διαστήματα ίσου πλάτους. Για παράδειγμα, μια περιοχή τιμών 1024..5512 καλύπτει πέντε διαφορετικές τιμές για το πιο σημαντικό ψηφίο και θα χωριστεί στα διαστήματα 1001..2000, 2001..3000, 3001..4000, 4001..5000 και 5001..6000.

Σε περίπτωση που στα δεδομένα παρουσιάζονται λίγες περιπτώσεις με πολύ μεγάλες ή/και με πολύ μικρές τιμές, η μέθοδος αυτή είναι αναποτελεσματική γιατί συγκεντρώνει ελάχιστες τιμές σε ορισμένα διαστήματα

και υπερβολικά πολλές τιμές σε ορισμένα άλλα. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι να επιλέξουμε ένα μεγάλο «κεντρικό» ποσοστό των τιμών, πχ 90% και να απομονώσουμε το 5% των μεγαλύτερων τιμών, καθώς και το 5% των μικρότερων τιμών (Σχήμα 7.7). Στη συνέχεια γίνεται διακριτοποίηση της κεντρικής περιοχής και ξεχωριστή διακριτοποίηση των ακραίων περιοχών.



Σχήμα 7.7 *Κατανομή τιμών με το 90% να βρίσκονται στην κεντρική περιοχή*

Ασαφής διακριτοποίηση. Όλες οι προηγούμενες μέθοδοι αντιστοιχούσαν με απόλυτο τρόπο μια τιμή σε ένα διάστημα. Στην ασαφή διακριτοποίηση μια τιμή μπορεί να ανήκει σε δύο γειτονικές περιοχές, ανάλογα με ένα σκορ που ορίζεται από τη συνάρτηση συμμετοχής. Το σκορ συμμετοχής κυμαίνεται μεταξύ των τιμών 0 και 1, όπου τιμή 0 συνεπάγεται μη συμμετοχή μιας τιμής σε ένα διάστημα, τιμή 1 συνεπάγεται πλήρη συμμετοχή, ενώ ενδιάμεση τιμή συνεπάγεται μερική συμμετοχή. Για παράδειγμα, η ταχύτητα ενός οχήματος μπορεί να χαρακτηριστεί υψηλή σύμφωνα με την παρακάτω συνάρτηση συμμετοχής

$$\Sigma υν\Sigma υμ(x) = \begin{Bmatrix} 0 & if\ x\ < 80 \\ (x-80)/40 & if\ 80\ < x\ < 120 \\ 1 & if\ x\ > 120 \end{Bmatrix} \tag{7.10}$$

Αν το όχημα κινείται με ταχύτητα 60 km/h τότε η συνάρτηση συμμετοχής παίρνει την τιμή 0 και η ταχύτητα δεν χαρακτηρίζεται υψηλή. Αν το όχημα κινείται με 130 km/h τότε τα σκορ συμμετοχής είναι 1 και η ταχύτητα χαρακτηρίζεται υψηλή. Αν το όχημα κινείται με 100 km/h τότε η συνάρτηση συμμετοχής παίρνει την τιμή 0,5 και η ταχύτητα χαρακτηρίζεται υψηλή με βαθμό 0,5.

Για τον καθορισμό της συνάρτησης συμμετοχής μπορεί να χρησιμοποιηθεί γνώση σχετικά με το συγκεκριμένο πρόβλημα ή να επιλεγεί κάποια από τις γνωστές συναρτήσεις όπως η τραπεζοειδής συνάρτηση. Ερευνητικές εργασίες παρουσιάζουν αποδεικτικά στοιχεία ότι με χρήση ασαφούς διακριτοποίησης μπορούν να επιτευχθούν καλύτερα αποτελέσματα. Οι Roy and Pal (2003) πρότειναν μια μέθοδο ασαφούς διακριτοποίησης και πραγματοποίησαν πειράματα κατηγοριοποίησης με διάφορα σύνολα δεδομένων. Οι ερευνητές συνέκριναν τα αποτελέσματα των κατηγοριοποιητών, όταν αυτοί χρησιμοποιούσαν μη διακριτοποιημένα δεδομένα, σαφή διακριτοποιημένα δεδομένα και δεδομένα με ασαφή διακριτοποίηση. Για την κατηγοριοποίηση χρησιμοποιήθηκαν ένα νευρωνικό δίκτυο τύπου Multilayer Perceptron και η μέθοδος των Τραχέων Συνόλων (Rough Sets). Και οι δύο κατηγοριοποιητές επέτυχαν καλύτερα αποτελέσματα με δεδομένα ασαφούς διακριτοποίησης.

Για περισσότερες πληροφορίες σχετικά με τις μεθόδους διακριτοποίησης, παραπέμπουμε τον αναγνώστη στο Liu, Hussain, Tan and Dash (2002) και στο Yang, Webb and Wu (2005).

Βιβλιογραφία / Αναφορές

Cardie, C. (1993). Using Decision Trees to Improve Case-Based Reasoning. *Proceedings of the 10th International Conference on Machine Learning*, 25-32. Amherst, MA: Morgan Kaufman.

Chen, M. Y. (2011). Bankruptcy Prediction in Firms with Statistical and Intelligent Techniques and a Comparison of Evolutionary Computation Approaches. *Computers and Mathematics with Applications*, 62(12), 4514–4524. doi: 10.1016/j.camwa.2011.10.030

Chen, H. L., Yang, B., Wang, G., Liu, J., Xu, X., Wang. S. J., & Liu, D. Y. (2011). A Novel Bankruptcy Prediction Model Based on an Adaptive Fuzzy k-Nearest Neighbor Method. *Knowledge Based Systems*, 24(8), 1348–1359. doi: 10.1016/j.knosys.2011.06.008

Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J., & Neves, C. J. (2011). A Genetic Algorithm-Based Approach to Cost-Sensitive Bankruptcy Prediction. *Expert Systems with Applications*, *38*(10), 12939–12945. doi: 10.1016/j.eswa.2011.04.090

Chizi, B., & Maimon, O. (2005). Dimensions Reduction and Feature Selection. In: O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers* (pp. 1189-1201). New York, NY: Springer Science + Business Media Inc.

Chmielewski, M. R., & Grzymala–Busse, J. W. (1996). Global Discretization of Continuous Attributes as Preprocessing for Machine Learning. *Journal of Approximate Reasoning*, 15(4), 319-331. doi: 10.1016/s0888-613x(96)00074-6

Cho, S., Hong, H., & Ha, B. C. (2010). A Hybrid Approach Based on the Combination of Variable Selection Using Decision Trees and Case-Base Reasoning Using the Mahalanobis Distance: For Bankruptcy Prediction. *Expert Systems with Applications*, 37(4), 3482–3488. doi: 10.1016/j.eswa.2009.10.040

Explained Visually. (2015). *Principal Components Analysis explained visually*. Retrieved 28 February, 2015, from http://setosa.io/ev/principal-component-analysis/.

Fayyad, U. M., & Irani, K. B. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1022-1027. Chambery, FR: Morgan Kaufmanm.

Fayyad, U. M., Piatetsky-Shapiro, G., & Uthurusamy, R. (2003). Summary from the KDD-03 Panel – Data Mining: The Next 10 Years. *ACM SIGKDD Exploration Newsletter*, 5(2), 191-196. doi: 10.1145/980972.981004

Frank, E., & Witten, I. H. (1999). Making Better Use of global Discretization. *Proceedings of the 16th International Conference on Machine Learning*, 115-123. San Francisco, CA: Morgan Kaufmann.

Hall, M. A. (1999). *Correlation-Based Feature Selection for Machine Learning* (Ph.D.). University of Waikato.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques*. Waltham, MA: Morgan Kaufmann Publishers.

Kirkos, E. (2015). Assessing Methodologies for Intelligent Bankruptcy Prediction. *Artificial Intelligence Review*, 43(1), 83-123. doi: 10.1007/s10462-012-9367-6

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery,* 6(4), 393-423. doi: 10.1023/A:1016304305535

Min, J. H., & Jeong, C. (2009). A binary classification method for bankruptcy prediction. *Expert Systems with Applications*, 36(3), 5256–5263. doi: 10.1016/j.eswa.2008.06.073

Orr, K. (1998). Data Quality and Systems Theory. *Communications of the ACM*, 41(2), 66-71. doi: 10.1145/269012.269023

Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Rakotomalala, R. (2005). TANAGRA: un logiciel gratuit pour l'enseignement et la recherché. *Actes de EGC'2005, RNTI-E-3*(2), 697–702.

Roy, A., & Pal, S. (2003). Fuzzy Discretization of Feature Space for a Rough Set Classifier. *Pattern Recognition Letters, 24*(6), 895-902. doi: 10.1016/s0167-8655(02)00201-5

Shie, F. S., Chen, M. Y., & Liu, Y. S. (2012). Prediction of corporate financial distress: an application of the America banking industry. *Neural Computing and Applications*, 21(7), 1687-1696. doi: 10.1007/s00521-011-0765-5

Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York, NY: Springer-Verlag.

Svolba, G. (2006). *Data Preparation for Analytics Using SAS*. Cary, NC: SAS Institute Inc.

Teng, C. M. (1999). Correcting Noisy Data. *Proceedings of the 16th International Conference on Machine Learning*, 239-248. San Francisco, CA: Morgan Kaufmann.

Yang, Y., Webb, G. I., & Wu, X. (2005). Discretization Methods. In: O. Maimon & L. Rokach, (Eds.), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers* (pp. 113-130). New York, NY: Springer Science + Business Media Inc.

Yeh, C. C., Chi, D. J., & Hsu, M. F. (2010). A hybrid approach of DEA, rough sets and support vector machines for business failure prediction. *Expert Systems with Applications, 37*(2), 1535-1541. doi: 10.1016/j.eswa.2009.06.088

Κριτήρια Αξιολόγησης

Άσκηση Υπολογισμών 7.1

Χρησιμοποιήστε τα δεδομένα του πίνακα 7.7.

- **Συμπληρώστε τις χαμένες τιμές στη στήλη «Τέκνα», χρησιμοποιώντας τη μέση ανά κλάση τιμή.**
- **Κανονικοποιήστε τις τιμές της στήλης «Ηλικία», εφαρμόζοντας κανονικοποίηση ελάχιστου - μέγιστου με οριακές τιμές το 0 και το 1.**
- **Διακριτοποιήστε τις τιμές της στήλης «Εισόδημα» χρησιμοποιώντας διαστήματα ίσου πλάτους, με πλάτος 10000 και τιμή εκκίνησης το 0.**

RID	Ονοματεπώνυμο	Τέκνα	Ηλικία	Εισόδημα	Έγκριση δανείου
1	Νικολάου Νικόλαος	1	35	25000	ΝΑΙ
2	Γεωργίου Γεώργιος	2	30	22000	ΝΑΙ
3	Δημητρίου Δημήτριος		33	8000	ΟΧΙ
4	Ιωάννου Ιωάννης	5	45	12000	ΟΧΙ
5	Χρήστου Χρήστος	2	42	31000	ΝΑΙ
6	Θεοδώρου Θεόδωρος	0	22	12000	ΝΑΙ
7	Αντωνίου Αντώνιος	3	59	31000	ΟΧΙ
8	Αναστασίου Αναστάσιος	1	27	5000	ΟΧΙ
9	Στεργίου Στέργιος	1	57	52000	ΝΑΙ
10	Αθανασίου Αθανάσιος		44	28000	ΝΑΙ
11	Αργυρίου Αργύριος	4	41	11000	ΟΧΙ
12	Αποστόλου Απόστολος	0	21	6000	ΟΧΙ

Πίνακας 7.7 *Δεδομένα Άσκησης 1*

Λύση

Βήμα 1. Για τη συμπλήρωση των χαμένων τιμών της στήλης «Τέκνα» υπολογίζεται η μέση τιμή για τις δύο τιμές της κλάσης. Για όσους εγκρίθηκε το δάνειο, το πλήθος των τέκνων είναι (1+2+2+0+1)=6 και η μέση τιμή είναι 6/5, η οποία στρογγυλοποιείται στην τιμή 1. Η τιμή αυτή καταχωρείται στη στήλη «Τέκνα», στην εγγραφή Νο 10 με ονοματεπώνυμο «Αθανασίου Αθανάσιος». Για όσους δεν εγκρίθηκε το δάνειο, το πλήθος των τέκνων είναι (5+3+1+4+0)=13 και η μέση τιμή είναι 13/5, η οποία στρογγυλοποιείται στην τιμή 3. Η τιμή αυτή καταχωρείται στη στήλη «Τέκνα», στην εγγραφή Νο 3 με ονοματεπώνυμο «Δημητρίου Δημήτριος».

Βήμα 2. Για την κανονικοποίηση των τιμών της στήλης «Ηλικία», αρχικά εντοπίζεται η μέγιστη τιμή, η οποία είναι η «59» και η ελάχιστη τιμή, η οποία είναι η «21». Κάθε τιμή της στήλης αντικαθίσταται με μια νέα η οποία υπολογίζεται σύμφωνα με την Εξίσωση 7.1.

Τα δεδομένα, μετά τις μετατροπές των τιμών παρουσιάζονται στον πίνακα 7.8,

RID	Ονοματεπώνυμο	Τέκνα	Ηλικία	Εισόδημα	Έγκριση δανείου
1	Νικολάου Νικόλαος	1	0,37	20000-30000	ΝΑΙ
2	Γεωργίου Γεώργιος	2	0,24	20000-30000	ΝΑΙ
3	Δημητρίου Δημήτριος	3	0,32	0-10000	ΟΧΙ
4	Ιωάννου Ιωάννης	5	0,63	10000-20000	ΟΧΙ
5	Χρήστου Χρήστος	2	0,55	30000-40000	ΝΑΙ
6	Θεοδώρου Θεόδωρος	0	0,03	10000-20000	ΝΑΙ
7	Αντωνίου Αντώνιος	3	1,00	30000-40000	ΟΧΙ
8	Αναστασίου Αναστάσιος	1	0,16	0-10000	ΟΧΙ

9	Στεργίου Στέργιος	1	0,95	50000-60000	ΝΑΙ
10	Αθανασίου Αθανάσιος	1	0,61	20000-30000	ΝΑΙ
11	Αργυρίου Αργύριος	4	0,53	10000-20000	ΟΧΙ
12	Αποστόλου Απόστολος	0	0,00	0-10000	ΟΧΙ

Πίνακας 7.8 *Αποτέλεσμα Άσκησης 1*

Άσκηση Υπολογισμών 7.2

Ένα αριθμητικό πεδίο ενός συνόλου δεδομένων περιέχει τις παρακάτω τιμές «24,25,26,26,29,29,29,31,33 ,36,37,38,41,44,44,48,49,50,57,63,72» Για να εξαλείψετε τον θόρυβο, εφαρμόστε κατακερματισμό σε διαστήματα και αντικατάσταση τιμών. Συγκεκριμένα, εφαρμόστε διαστήματα ίσης συχνότητας, με πλήθος τιμών σε κάθε διάστημα ίσο με 3 και αντικατάσταση μέσων όρων.

Λύση

Το σύνολο των αριθμητικών τιμών κατακερματίζεται σε επτά διαστήματα ως εξής : [24,25,26], [26,29,29], [29,31,33], [36,37,38], [41,44,44], [48,49,50], [57,63,72].

Για κάθε ένα από αυτά τα διαστήματα υπολογίζεται ο μέσος όρος. Οι μέσοι όροι για τα επτά διαστήματα είναι 25,28,31,37,43,49,64

Οι τιμές κάθε διαστήματος αντικαθίστανται με τον μέσο όρο του διαστήματος. Μετά τις αντικαταστάσεις, τα δεδομένα θα έχουν τις ακόλουθες τιμές «25,25,25,28,28,28,31,31,31,37,37,37,43,43,43,49,49,49, 64,64,64»

Άσκηση Εφαρμογής 7.3

Χρησιμοποιήστε το αρχείο «analcatdata_bankruptcy.arff» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή StatLib). Το σύνολο δεδομένων προέρχεται από το βιβλίο του Simonoff (2003) και σχετίζεται με τη χρεοκοπία επιχειρήσεων. Υπάρχουν 50 γραμμές, κάθε μια από τις οποίες αναφέρεται σε μια επιχείρηση. Οι μισές επιχειρήσεις έχουν χρεοκοπήσει. Στο σύνολο δεδομένων υπάρχουν 7 πεδία (στήλες). Το πρώτο πεδίο περιέχει τα ονόματα των επιχειρήσεων, και ακολουθούν 5 πεδία με αριθμοδείκτες. Το τελευταίο πεδίο είναι το πεδίο της κλάσης και περιέχει μια ένδειξη («1» ή «0») για το εάν η επιχείρηση χρεοκόπησε ή εξακολουθεί τη λειτουργία της αντίστοιχα.

- **Διακριτοποιήστε το πεδίο WC/TA, εφαρμόζοντας τη μέθοδο των διαστημάτων ίσου πλάτους με πλήθος διαστημάτων ίσο με 5.**
- **Κανονικοποιήστε τα δεδομένα των αριθμητικών στηλών, αντιστοιχίζοντας τα στην περιοχή [0..1].**
- **Κανονικοποιήστε τα δεδομένα των αριθμητικών στηλών, έτσι ώστε να έχουν μέση τιμή ίση με 0 και τυπική απόκλιση ίση με 1.**
- **Τροποποιήστε τα αριθμητικά πεδία σύμφωνα με μια δική σας συνάρτηση. Μπορείτε πχ να υποδιπλασιάσετε τις τιμές.**

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «analcatdata_bankruptcy.arff» πιέζοντας το κουμπί «Open file».

Κάντε κλικ στο κουμπί «Choose» του πεδίου «Filter» και επιλέξτε weka\filters\unsupervised\arrtibute\ Discretize.

Κάντε κλικ στα περιεχόμενα του πεδίου «Filter» και στο όνομα «Discretize». Ανοίγει το παράθυρο ρύθμισης των παραμέτρων. Στο πεδίο «attributeIndices» εισάγετε την τιμή 2. Με τον τρόπο αυτό ορίζετε ότι θα διακριτοποιήσετε τη δεύτερη στήλη. Στο πεδίο «bins» εισάγετε την τιμή 5.

Κάντε κλικ στο κουμπί «Apply» για να εκτελέσετε τον αλγόριθμο.

Κάντε κλικ στην εγγραφή «WC/TA» στο πεδίο «Attributes». Στο πεδίο «selected attribute» εμφανίζονται οι περιοχές τιμών. Αν κάνετε κλικ στο κουμπί «Edit», μπορείτε να δείτε τα τροποποιημένα δεδομένα.

Βήμα 2. Για να κανονικοποιήσετε τα δεδομένα των αριθμητικών στηλών αντιστοιχίζοντας τα στην περιοχή [0..1], ανοίξτε ξανά το αρχείο «analcatdata_bankruptcy.arff».

Κάντε κλικ στο κουμπί «Choose» του πεδίου «Filter» και επιλέξτε weka\filters\unsupervised\arrtibute\ Normalize. Δεν χρειάζεται να τροποποιήσετε τις παραμέτρους.

Κάντε κλικ στο κουμπί «Apply» για να εκτελέσετε τον αλγόριθμο.

Κάντε κλικ στα αριθμητικά γνωρίσματα στο πεδίο «Attributes». Στο πεδίο «selected attribute» διαπιστώστε ότι οι μέγιστη και ελάχιστη τιμή για κάθε πεδίο είναι 1 και 0 αντίστοιχα. Αν κάνετε κλικ στο κουμπί «Edit», μπορείτε να δείτε τα τροποποιημένα δεδομένα.

Βήμα 3. Για να κανονικοποιήσετε τα δεδομένα των αριθμητικών στηλών, έτσι ώστε να έχουν μέση τιμή ίση με 0 και τυπική απόκλιση ίση με 1, ανοίξτε ξανά το αρχείο «analcatdata_bankruptcy.arff».

Κάντε κλικ στο κουμπί «Choose» του πεδίου «Filter» και επιλέξτε weka\filters\unsupervised\arrtibute\ Standarize. Δεν χρειάζεται να τροποποιήσετε τις παραμέτρους.

Κάντε κλικ στο κουμπί «Apply» για να εκτελέσετε τον αλγόριθμο.

Κάντε κλικ στα αριθμητικά γνωρίσματα στο πεδίο «Attributes». Στο πεδίο «selected attribute» διαπιστώστε ότι οι μέσες τιμές είναι ίσες με 0 και οι τυπικές αποκλίσεις είναι ίσες με 1. Αν κάνετε κλικ στο κουμπί «Edit», μπορείτε να δείτε τα τροποποιημένα δεδομένα.

Βήμα 4. Για να υποδιπλασιάσετε τις τιμές των αριθμητικών πεδίων, ανοίξτε ξανά το αρχείο «analcatdata_ bankruptcy.arff».

Κάντε κλικ στο κουμπί «Choose» του πεδίου «Filter» και επιλέξτε weka\filters\unsupervised\arrtibute\ MathExpression.

Κάντε κλικ στα περιεχόμενα του πεδίου «Filter» και στο όνομα «MathExpression». Ανοίγει το παράθυρο ρύθμισης των παραμέτρων. Στο πεδίο «Expression» εισάγετε τη συνάρτηση μετασχηματισμού των τιμών. Για να υποδιπλασιάσετε τις τιμές πληκτρολογήστε «A/2». Αφού ορίσετε τη συνάρτηση, κάντε κλικ στο κουμπί «OK».

Κάντε κλικ στο κουμπί «Apply» για να εκτελέσετε τον αλγόριθμο.

Κάντε στο κουμπί «Edit» και θα διαπιστώσετε ότι οι αριθμητικές τιμές έχουν υποδιπλασιαστεί.

Άσκηση Εφαρμογής 7.4

Χρησιμοποιήστε το αρχείο «analcatdata_japansolvent.arff» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή StatLib). Το σύνολο δεδομένων προέρχεται από το βιβλίο του Simonoff (2003) και σχετίζεται με την κατηγοριοποίηση ιαπωνικών επιχειρήσεων σε φερέγγυες (solvent) και αφερέγγυες (insolvent). Υπάρχουν 52 γραμμές, κάθε μια από τις οποίες αναφέρεται σε μια επιχείρηση. Οι 25 επιχειρήσεις χαρακτηρίζονται αφερέγγυες και οι 27 φερέγγυες. Στο σύνολο δεδομένων υπάρχουν 10 πεδία (στήλες). Το πρώτο πεδίο περιέχει τα ονόματα των επιχειρήσεων, το δεύτερο είναι το πεδίο κλάσης και ακολουθούν οκτώ αριθμοδείκτες. Οι κλάσεις κωδικοποιούνται με τις τιμές «0» για τις αφερέγγυες επιχειρήσεις και «1» για τις φερέγγυες.

- **Δημιουργήστε μια νέα στήλη, η οποία θα προκύπτει με άθροιση των αριθμοδεικτών Sales/TA και Equity/TA.**
- **Μειώστε το πλήθος των διαστάσεων και αντιστοιχείστε τις σε νέες διαστάσεις, εφαρμόζοντας τη μέθοδο Ανάλυσης Κυρίων Συνιστωσών.**
- **Εκτελέστε Επιλογή Χαρακτηριστικών (Feature Selection) με τη μέθοδο CGS Subset Evaluator στις αρχικές στήλες.**
- **Εκτελέστε Επιλογή Χαρακτηριστικών (Feature Selection) με τη μέθοδο wrapper και κατηγοριοποιητή Δένδρο Αποφάσεων τύπου C4.5, στις αρχικές στήλες.**

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «analcatdata_japansolvent.arff» πιέζοντας το κουμπί «Open file».

Κάντε κλικ στο κουμπί «Choose» του πεδίου «Filter» και επιλέξτε weka\filters\unsupervised\attribute\ AddExpression.

Κάντε κλικ στα περιεχόμενα του πεδίου «Filter» και το όνομα «AddExpression». Ανοίγει το παράθυρο ρύθμισης παραμέτρων. Στο πεδίο «expression» εισάγετε τη συνάρτηση υπολογισμού τιμών του νέου πεδίου. Για την πρόσθεση των δύο αριθμοδεικτών της εκφώνησης, πληκτρολογήστε «a5+a10». Με τον τρόπο αυτό

ορίζετε ότι θα προστεθούν οι τιμές της πέμπτης και της δέκατης στήλης. Κάντε κλικ στο κουμπί «OK».

Κάντε κλικ στο κουμπί «Apply» για να εφαρμόσετε τον αλγόριθμο.

Στο πεδίο «Attributes» διαπιστώστε ότι προστέθηκε μια νέα στήλη με όνομα «a5+a10». Κάνοντας κλικ στο κουμπί «Edit», προβάλετε τα δεδομένα και διαπιστώσετε ότι οι τιμές της νέας στήλης είναι το άθροισμα των αριθμοδεικτών Sales/TA και Equity/TA.

Βήμα 2. Για να μειώσετε το πλήθος των διαστάσεων και να τις αντιστοιχίστε σε νέες διαστάσεις εφαρμόζοντας την Ανάλυση Κυρίων Συνιστωσών, ανοίξτε ξανά το αρχείο «analcatdata_japansolvent.arff».

Στο πεδίο "Filter" επιλέξτε weka\filters\unsupervised\attibute\PrincipalComponents.

Διαγράψτε τη στήλη με το όνομα των εταιρειών. Η στήλη αυτή δεν προσφέρει κάτι χρήσιμο στην ανάλυση και επιπλέον, θα αποπροσανατολίσει τον αλγόριθμο των Κυρίων Συνιστωσών. Για να τη διαγράψετε, κάντε κλικ στο τετράγωνο μπροστά από τη στήλη «Firm» και στη συνέχεια κάντε κλικ στο κουμπί «Remove».

Ορίστε ως πεδίο κλάσης τη στήλη «Solvent».

Κάντε κλικ στο κουμπί «Apply» για να εκτελεστεί ο αλγόριθμος.

Στο πεδίο «Attributes» θα εμφανιστούν οι κύριες συνιστώσες (στη θέση των παλιών στηλών). Για κάθε κύρια συνιστώσα δίνεται ο τύπος υπολογισμού της. Κάνοντας κλικ στη κάθε συνιστώσα παρατηρήστε γραφικά την κατανομή των τιμών και των κλάσεων.

Βήμα 3. Για να εκτελέσετε Επιλογή Χαρακτηριστικών με τη μέθοδο CGS Subset Evaluator στις αρχικές στήλες, ανοίξτε ξανά το αρχείο «analcatdata_japansolvent.arff».

Ορίστε ως πεδίο κλάσης τη στήλη «Solvent».

Στο πεδίο "Filter" επιλέξτε weka\filters\supervised\attribute\AttributeSelection.

Κάντε κλικ στα περιεχόμενα του πεδίου «Filter» και στο όνομα «AttributeSelection». Ανοίγει το παράθυρο ρύθμισης των παραμέτρων. Βεβαιωθείτε ότι στο πεδίο «evaluator» είναι επιλεγμένη η μέθοδος «CFSSubsetEval», διαφορετικά επιλέξτε την κάνοντας κλικ στο κουμπί «Choose». Κάντε κλικ στο κουμπί «OK».

Κάντε κλικ στο κουμπί «Apply» για να εφαρμόσετε τη μέθοδο.

Διαπιστώστε ότι στο πεδίο «Attributes» υπάρχουν μόνο οκτώ στήλες (από τις δέκα που υπήρχαν αρχικά). Η μια στήλη είναι η κλάση (solvent).Οι υπόλοιπες επτά στήλες θεωρήθηκαν σημαντικές και μπορούν να χρησιμοποιηθούν για περαιτέρω ανάλυση.

Βήμα 4. Για να εκτελέσετε Επιλογή Χαρακτηριστικών με wrapper και Δένδρο Αποφάσεων τύπου C4.5 στις αρχικές στήλες, ανοίξτε ξανά το αρχείο «analcatdata_japansolvent.arff».

Ορίστε ως πεδίο κλάσης τη στήλη «Solvent».

Στο πεδίο "Filter" επιλέξτε weka\filters\supervised\attribute\AttributeSelection.

Κάντε κλικ στα περιεχόμενα του πεδίου «Filter» και στο όνομα «AttributeSelection». Ανοίγει το παράθυρο ρύθμισης των παραμέτρων. Στο πεδίο «evaluator» επιλέξτε τη μέθοδο «WrapperSubsetEval».

Κάντε κλικ στα περιεχόμενα του πεδίου «evaluator» και στο όνομα «WrapperSubsetEval». Ανοίγει το παράθυρο ρύθμισης των παραμέτρων. Στο πεδίο «classifier» επιλέξτε weka\classifiers\trees\J48. Κλείστε τα παράθυρα ρύθμισης παραμέτρων, κάνοντας κλικ στα κουμπιά «OK».

Κάντε κλικ στο κουμπί «Apply» για να εφαρμόσετε τη μέθοδο.

Διαπιστώσετε ότι στο πεδίο «Attributes» υπάρχουν μόνο τρεις στήλες (από τις δέκα που υπήρχαν αρχικά). Η μια στήλη είναι η κλάση (solvent).Οι υπόλοιπες δύο στήλες θεωρήθηκαν σημαντικές και μπορούν να χρησιμοποιηθούν για περαιτέρω ανάλυση.

8 Κανόνες Συσχέτισης

Σύνοψη

Η ανακάλυψη Κανόνων Συσχέτισης είναι μία από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Αντικείμενο της είναι η ανακάλυψη και διατύπωση σχέσεων, οι οποίες υπάρχουν στα δεδομένα. Οι σχέσεις αυτές προκύπτουν από τη συχνή ταυτόχρονη εμφάνιση τιμών δεδομένων. Το βασικό πεδίο εφαρμογής είναι η ανάλυση του καλαθιού αγορών, η οποία μελετά τις καταναλωτικές συνήθειες των πελατών μέσα από την ταυτόχρονη πώληση προϊόντων. Το παρόν Κεφάλαιο καλύπτει θέματα των Κανόνων Συσχέτισης. Σκοπός του συγγραφέα είναι να εισάγει τον αναγνώστη στις βασικές έννοιες και τεχνικές, οι οποίες είναι απαραίτητες για τη διεξαγωγή ανάλυσης Κανόνων Συσχέτισης. Η εξαντλητική κάλυψη των πολλών εξειδικευμένων αλγορίθμων που κατά καιρούς έχουν προταθεί, βρίσκεται έξω από τα όρια και τις επιδιώξεις του παρόντος συγγράμματος. Αναλυτική παρουσίαση γίνεται μόνο για τον βασικό αλγόριθμο Apriori. Λοιποί εξειδικευμένοι αλγόριθμοι, οι οποίοι επεκτείνουν τον Apriori, παρουσιάζονται ακροθιγώς. Για τον αναγνώστη που ενδιαφέρεται να αναζητήσει πρόσθετες λεπτομέρειες σχετικά με τους αλγόριθμους, προσφέρονται αναφορές και παραπομπές στις σχετικές πηγές.

Στο παρόν Κεφάλαιο, αρχικά γίνεται παρουσίαση των βασικών εννοιών που σχετίζονται με τους Κανόνες Συσχέτισης. Ορίζονται και εξηγούνται οι έννοιες του στοιχειοσυνόλου, του k-στοιχειοσυνόλου, της συχνότητας εμφάνισης, της υποστήριξης και της εμπιστοσύνης. Ακολούθως, προσδιορίζεται η εργασία εξόρυξης Κανόνων Συσχέτισης ως μια διαδικασία δύο σταδίων. Κατά το πρώτο στάδιο γίνεται εντοπισμός των συχνών στοιχειοσυνόλων, ενώ κατά το δεύτερο στάδιο δημιουργούνται οι κανόνες από τα συχνά στοιχειοσύνολα. Ο βασικός αλγόριθμος εντοπισμού συχνών στοιχειοσυνόλων, δηλαδή ο Apriori, παρουσιάζεται αναλυτικά και παρατίθενται σχετικά παραδείγματα και σχήματα. Στη συνέχεια, εξηγείται η διαδικασία δημιουργίας κανόνων από τα συχνά στοιχειοσύνολα. Οι κανόνες που παράγονται ικανοποιούν τα κριτήρια της υποστήριξης και της εμπιστοσύνης, δεν είναι σίγουρο όμως ότι είναι και ενδιαφέροντες. Ένα πρόσθετο στατιστικό μέτρο αξιολόγησης των κανόνων είναι το Lift, το οποίο αποδίδει τον βαθμό και το είδος συσχέτισης μεταξύ των γεγονότων του κανόνα. Εξόρυξη κανόνων μπορεί να γίνει όχι μόνο από βάσεις δεδομένων συναλλαγών, αλλά και από σχεσιακές βάσεις δεδομένων. Οι κανόνες αυτοί μπορεί να περιέχουν πολλά κατηγορήματα. Στην περίπτωση αυτή καλούνται πολυδιάστατοι. Εάν έχουν καθοριστεί ιεραρχίες εννοιών, μπορούν να δημιουργηθούν κανόνες που να αφορούν διαφορετικά επίπεδα γενίκευσης. Η χρήση διαφορετικών τιμών ελάχιστης υποστήριξης για τα διαφορετικά επίπεδα γενίκευσης διευκολύνει τη δημιουργία περιορισμένου αριθμού κανόνων για τα ανώτερα επίπεδα, καθώς και τον εντοπισμό ισχυρών κανόνων στα κατώτερα επίπεδα. Ένα σημαντικό πρόβλημα είναι η εξόρυξη κανόνων από αριθμητικά δεδομένα. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί διάφορες μέθοδοι, οι περισσότερες από τις οποίες εφαρμόζουν τη διακριτοποίηση, σε συνδυασμό με πρόσθετες τεχνικές. Με τη χρήση περιορισμών κατά την εξόρυξη των κανόνων επιτυγχάνεται επικέντρωση της ανάλυσης σε εντοπισμένα ζητήματα, αλλά και ταυτόχρονη μείωση του χώρου αναζήτησης λύσεων. Στο τελευταίο υποκεφάλαιο παρουσιάζεται μελέτη περίπτωσης, όπου οι Κανόνες Συσχέτισης εφαρμόζονται για τη μοντελοποίηση των αποφάσεων των εξωτερικών ελεγκτών.

Προηγούμενη γνώση

Οι Κανόνες Συσχέτισης κάνουν εκτεταμένη χρήση εννοιών, οι οποίες προέρχονται από τη Θεωρία Συνόλων. Τέτοιες έννοιες είναι το υποσύνολο, το γνήσιο υποσύνολο, η τομή, η ένωση κλπ. Για τον αναγνώστη που δεν είναι εξοικειωμένος με τις έννοιες αυτές, θα ήταν χρήσιμο να ανατρέξει σε κάποιο από τα πολλά βιβλία μαθηματικών, τα οποία αναφέρονται στο θέμα. Για τη δημιουργία κανόνων από πεδία αριθμητικών τιμών, γίνεται χρήση της διακριτοποίησης. Τεχνικές διακριτοποίησης παρουσιάζονται αναλυτικά στο Κεφάλαιο 7. Επίσης, η διακριτοποίηση των αριθμητικών δεδομένων συνδυάζεται με μεθόδους Ανάλυσης Συστάδων, όπως η μέθοδος k-Means ή η μέθοδος των Αυτοοργανούμενων Χαρτών. Παρουσίαση των μεθόδων αυτών γίνεται στο Κεφάλαιο 11. Η εξόρυξη Κανόνων Συσχέτισης είναι ένα ευρύ αντικείμενο και δεν μπορεί να καλυφθεί πλήρως στα πλαίσια ενός κεφαλαίου. Ο αναγνώστης, ο οποίος ενδιαφέρεται να αναζητήσει περισσότερες πληροφορίες για το θέμα αυτό, μπορεί να ανατρέξει σε κάποιο βιβλίο Εξόρυξης Δεδομένων, όπως το Han, Kamber and Pei (2011). Στους κανόνες Συσχέτισης αναφέρεται επίσης ο Hoeppner (2005), ενώ αλγόριθμοι ανακάλυψης συχνών στοιχειοσυνόλων παρουσιάζονται στο Goethals (2005). Τέλος, βιβλία εξειδικευμένα στους Κανόνες Συσχέτισης είναι το Zhang and Zhang (2002), το Gkoulalas-Divanis and Verykios (2010) και το Koh and Rountree (2009).

8.1 Εισαγωγή

Η ανακάλυψη Κανόνων Συσχέτισης είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Από πολλούς μάλιστα θεωρείται ως το πιο γνήσιο τέκνο της Εξόρυξης Δεδομένων, καθώς άλλες εργασίες εξόρυξης, μεθοδολογίες και τεχνικές προέρχονται κυρίως από τη Μηχανική Μάθηση, τη Στατιστική, τις Βάσεις Δεδομένων κλπ. Οι Κανόνες Συσχέτισης αφορούν την ανακάλυψη και διατύπωση σχέσεων που υπάρχουν στα δεδομένα. Οι σχέσεις αυτές προκύπτουν από τη συχνή ταυτόχρονη εμφάνιση τιμών δεδομένων.

Το συνηθέστερο πεδίο εφαρμογής αλλά και το συνηθέστερο παράδειγμα Κανόνων Συσχέτισης είναι η ανάλυση του καλαθιού αγορών (market basket analysis). Η ανάλυση του καλαθιού αγορών αναφέρεται στη μελέτη των αγορών που πραγματοποιούν οι πελάτες ενός καταστήματος. Κάθε αγορά περιλαμβάνει ένα σύνολο προϊόντων. Αναλύοντας τα σύνολα προϊόντων των πωλήσεων, μπορούν να βρεθούν ομάδες προϊόντων οι οποίες πωλούνται συχνά μαζί. Από αυτές τις ομάδες υπολογίζονται κανόνες της μορφής «εάν ένας πελάτης αγοράσει το προϊόν Α, έχει 60% πιθανότητες να αγοράσει ταυτόχρονα και το προϊόν Β, ενώ η ταυτόχρονη πώληση των προϊόντων Α και Β παρουσιάζεται στο 5% του συνόλου των πωλήσεων». Η πληροφορία αυτή είναι πολλαπλώς αξιοποιήσιμη από τον υπεύθυνο πωλήσεων. Μπορεί να χρησιμοποιηθεί για τη διαρρύθμιση και την τοποθέτηση των προϊόντων μέσα στο κατάστημα. Τοποθετώντας δύο προϊόντα, τα οποία πωλούνται συχνά μαζί, σε γειτονικές θέσεις, επιτυγχάνεται αύξηση των πωλήσεων. Ο πελάτης, ο οποίος περιπλανιέται στο κατάστημα, όταν αγοράσει το προϊόν Α, έχει περισσότερες πιθανότητες και προτρέπεται να αγοράσει το συγγενές προϊόν Β, εάν το βρει σε μια γειτονική θέση. Σε αντίθετη περίπτωση είναι πιθανόν να ξεχάσει το προϊόν Β και να το αγοράσει κάποια άλλη χρονική στιγμή από άλλο κατάστημα. Άλλος τρόπος αξιοποίησης αυτής της πληροφορίας είναι ο σχεδιασμός προσφορών. Ο υπεύθυνος πωλήσεων, προσφέροντας μια δελεαστική τιμή για το προϊόν Α, προσελκύει πελάτες και αυξάνει τις πωλήσεις του προϊόντος Α. Αν τοποθετήσει σε γειτονικά ράφια το συγγενές προϊόν Β, θα επιτύχει αύξηση πωλήσεων του Β, καθώς οι πελάτες, γνωρίζοντας ότι εξοικονόμησαν χρήματα από την αγορά του Α, πιθανόν να προβούν σε πρόσθετες αγορές. Τέτοιες τεχνικές διασταυρούμενων πωλήσεων εφαρμόζονται και για την ταχεία πώληση ευπαθών προϊόντων με κοντινή ημερομηνία λήξης.

Η ένταση του ανταγωνισμού μεταξύ των επιχειρήσεων έχει καταστήσει την προσέλκυση νέων πελατών αρκετά δύσκολη. Για τον λόγο αυτό, οι επιχειρήσεις προσανατολίζονται όλο και περισσότερο στην καλύτερη εξυπηρέτηση των ήδη υπαρκτών πελατών και στην αύξηση των πωλήσεων μέσα από την πώληση περισσότερων προϊόντων στους ίδιους πελάτες. Για την επίτευξη αυτού του στόχου, είναι αναγκαία η βαθιά κατανόηση της καταναλωτικής τους συμπεριφοράς. Η ανάλυση κανόνων συσχέτισης μπορεί να αποδώσει εντυπωσιακά αποτελέσματα και πρωτόγνωρα συμπεράσματα. Το παράδειγμα με τις πάνες και τις μπύρες έχει καταστεί θρύλος. Ανάλυση κανόνων συσχέτισης σε σούπερ μάρκετ κατέδειξε ότι οι πάνες υγιεινής για βρέφη και οι μπύρες πωλούνται συχνά μαζί. Ο εντοπισμός τέτοιων, παράδοξων με μια πρώτη ματιά, ωστόσο έγκυρων συσχετίσεων συμβάλλει στην καλύτερη κατανόηση της καταναλωτικής συμπεριφοράς. Η γνώση των πραγματικών αναγκών του πελάτη αποτελεί ακρογωνιαίο λίθο του μάρκετινγκ, καθώς επιτρέπει τον σχεδιασμό κατάλληλων λύσεων και την ανάληψη της απαιτούμενης δράσης. Αν για παράδειγμα, η ανάλυση εντοπίσει την αυξημένη και ταυτόχρονη πώληση προϊόντων υγιεινής διατροφής, αυτό σημαίνει την ύπαρξη μιας μερίδας καταναλωτικού κοινού με αντίστοιχες προτιμήσεις. Ανταποκρινόμενος σε αυτήν την τάση, ο υπεύθυνος πωλήσεων μπορεί να εγκαινιάσει στο κατάστημα του μια νέα πτέρυγα βιολογικών προϊόντων και να εισάγει νέες κατηγορίες προϊόντων υγιεινής διατροφής. Με τον τρόπο αυτό, θα εξυπηρετήσει καλύτερα τους πελάτες του, θα αυξήσει την κατανάλωση της πελατειακής του βάσης και πιθανόν να προσελκύσει και νέους πελάτες. Η μεγάλη χρησιμότητα της ανάλυσης Κανόνων Συσχέτισης την έχει καταστήσει μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων.

8.2 Ορισμοί

Για την ανάλυση των πωλήσεων ενός καταστήματος, που διαθέτει k εμπορεύματα $E_1,\ldots,E_k$, μπορούμε να θεωρήσουμε ένα σχήμα βάσης δεδομένων με k πεδία (στήλες), όπου κάθε στήλη αντιστοιχεί σε ένα εμπόρευμα. Οι τιμές των πεδίων είναι δυαδικές, δηλαδή *0* και *1*. Κάθε πώληση καταγράφεται ως μια γραμμή. Μια πώληση, η οποία περιλαμβάνει τα εμπορεύματα E_5 και E_{12}, καταγράφεται θέτοντας στα πεδία E_5 και E_{12} την τιμή *1* και σε όλα τα υπόλοιπα πεδία την τιμή *0*. Σε μια τέτοια βάση δεδομένων η ανάλυση Κανόνων Συσχέτισης αναζητά συχνά πρότυπα και εξάγει κανόνες από αυτά. Η δυσκολία του προβλήματος έγκειται στο μέγεθος του. Αν θεωρήσουμε μόλις 100 εμπορεύματα (k=100) και κανόνες με δύο εμπορεύματα στο δεξιό και αριστερό μέρος του κανόνα, τότε έχουμε περίπου 25.000.000 κανόνες. Η εξαντλητική επαλήθευση αυτών των κανόνων σε μια

βάση δεδομένων είναι απαγορευτικά χρονοβόρα. Για την αντιμετώπιση αυτού του προβλήματος οι Agrawal, Imielinski and Swami (1993) πρότειναν τους Κανόνες Συσχέτισης.

Ένας εναλλακτικός τρόπος θεώρησης του προβλήματος είναι να ορίσουμε μια πώληση, η οποία περιλαμβάνει τα προϊόντα E_5 και E_{12}, ως ένα σύνολο $T_i=\{E_5, E_{12}\}$. Τα εμπορεύματα του καταστήματος ορίζονται ως ένα σύνολο $M=\{E_1,\ldots,E_k\}$. Κάθε συναλλαγή T θεωρείται ως ένα σύνολο εμπορευμάτων και ως ένα υποσύνολο του M. Αν X είναι ένα σύνολο εμπορευμάτων, λέμε ότι η συναλλαγή T_j περιέχει το X, άν και μόνο αν $X \subseteq T_j$.

Ένας **Κανόνας Συσχέτισης** (Association Rule) έχει τη μορφή $\boldsymbol{X \rightarrow Y}$ όπου $X \subset M$, $Y \subset M$ και $X \cap Y = \varnothing$. Το X είναι ένα υποσύνολο των εμπορευμάτων του καταστήματος, πχ $X=\{E_5,E_6\}$. Το Υ είναι ένα άλλο υποσύνολο των εμπορευμάτων του καταστήματος, το οποίο δεν έχει κανένα κοινό μέλος με το X, πχ $Y=\{E_{12},E_{13}\}$. Ο κανόνας $X \rightarrow Y$ μπορεί να διατυπωθεί ως «όταν κάποιος αγοράζει τα προϊόντα *E5* και *E6* τότε αγοράζει και τα προϊόντα E_{12} και E_{13}».

Δύο ποσοτικά μεγέθη καθορίζουν πόσο ισχυρός είναι ο κανόνας $X \rightarrow Y$. Τα μέτρα αυτά είναι η **υποστήριξη** (support) και η **εμπιστοσύνη** (confidence).

Η **υποστήριξη** του κανόνα $X \rightarrow Y$ είναι το ποσοστό των συναλλαγών (επί του συνόλου των συναλλαγών) που περιέχουν και το X και το Y. Μαθηματικά, αυτό ορίζεται με τη Σχέση 8.1

$$supp(X \rightarrow Y) = P(X \cup Y) \tag{8.1}$$

Η **εμπιστοσύνη** του κανόνα $X \rightarrow Y$ είναι η δεσμευμένη πιθανότητα εμφάνισης του Y, όταν εμφανίζεται το X. Με απλούστερα λόγια, επιλέγονται μόνον οι συναλλαγές που περιέχουν το X και επί αυτών των συναλλαγών υπολογίζεται το ποσοστό εκείνων που περιέχουν το Y. Μαθηματικά, αυτό ορίζεται με τη Σχέση 8.2

$$conf(X \rightarrow Y) = P(Y|X) \tag{8.2}$$

Για την καλύτερη κατανόηση των εννοιών αυτών παραθέτουμε ένα παράδειγμα. Ο Πίνακας 8.1 περιέχει το σύνολο των συναλλαγών ενός καταστήματος. Για κάθε συναλλαγή καταγράφεται ένας αναγνωριστικός κωδικός αριθμός (Transaction ID – TID) και τα εμπορεύματα που πωλήθηκαν σε αυτήν τη συναλλαγή. Τα *Α,Β,Γ,Δ,Ε* είναι διάφορα εμπορεύματα

TID	Εμπορεύματα
101	Α,Β,Γ
102	Γ,Δ
103	Α,Β
104	Α,Β,Δ
105	Α,Δ
106	Β,Γ

Πίνακας 8.1 *Συναλλαγές - εμπορεύματα*

Θεωρούμε τον κανόνα $A \rightarrow B$, ο οποίος σημαίνει ότι όταν κάποιος αγοράζει το προϊόν A, τότε αγοράζει και το προϊόν B. Παρατηρούμε ότι σε τρεις από τις συνολικά έξι συναλλαγές (101, 103, 104) πωλούνται ταυτόχρονα τα προϊόντα A και B. Η υποστήριξη του κανόνα είναι 3/6, δηλαδή 50%. Επίσης παρατηρούμε ότι το προϊόν A εμφανίζεται σε τέσσερεις συναλλαγές (101, 103,104,105) και ότι σε τρεις από αυτές (101, 103, 104) εμφανίζεται και το προϊόν B. Η εμπιστοσύνη του κανόνα είναι 3/4, δηλαδή 75%.

Για την ανακάλυψη Κανόνων Συσχέτισης, ο χρήστης προκαθορίζει ελάχιστες τιμές για την υποστήριξη και την εμπιστοσύνη. Στη συνέχεια, ο αλγόριθμος διατρέχει τη βάση δεδομένων, αναλύει τα δεδομένα και εντοπίζει όλους τους κανόνες που έχουν υποστήριξη και εμπιστοσύνη ίση ή μεγαλύτερη από τις προκαθορισμένες τιμές. Οι κανόνες αυτοί θεωρούνται ισχυροί.

Ορισμένοι πρόσθετοι όροι των Κανόνων Συσχέτισης είναι οι ακόλουθοι:

- **Στοιχειοσύνολο** (Itemset) ονομάζεται ένα σύνολο από στοιχεία (items). Στο παράδειγμα μας ένα

σύνολο *I*, που περιέχει τα εμπορεύματα *Α,Δ,Ε (I={Α,Δ,Ε})*, είναι ένα στοιχειοσύνολο.

- **k-Στοιχειοσύνολο** (*k*-Itemset) είναι ένα στοιχειοσύνολο που περιέχει *k* στοιχεία. Το στοιχειοσύνολο *I={Α,Δ,Ε}* είναι ένα 3-Στοιχειοσύνολο.
- **Συχνότητα Εμφάνισης** (frequency ή support count ή count) ενός στοιχειοσυνόλου είναι το πλήθος των συναλλαγών που περιέχουν το στοιχειοσύνολο. Η συχνότητα εμφάνισης του στοιχειοσυνόλου *{Β,Γ}* είναι ίση με 2.
- **Υποστήριξη** (support) ενός στοιχειοσυνόλου είναι το ποσοστό των συναλλαγών που περιέχουν το στοιχειοσύνολο. Η υποστήριξη του στοιχειοσυνόλου *{Β,Γ}* είναι ίση με 2/6=33%
- **Συχνό στοιχειοσύνολο** είναι εκείνο το στοιχειοσύνολο του οποίου η υποστήριξη είναι ίση ή μεγαλύτερη από την ελάχιστη υποστήριξη, που ορίζει ο χρήστης.

Η υποστήριξη και η εμπιστοσύνη είναι δύο ισχυρά μέτρα της ισχύος ενός κανόνα. Η υποστήριξη εξασφαλίζει ότι ο κανόνας αφορά ένα ικανοποιητικό ποσοστό των συναλλαγών. Κανόνες με μικρή υποστήριξη μπορεί να θεωρηθεί ότι εκφράζουν ένα τυχαίο γεγονός. Επίσης, η εμπιστοσύνη αποτελεί μέτρο του κατά πόσο η εμφάνιση του αριστερού μέρους πριμοδοτεί την εμφάνιση του δεξιού μέρους του κανόνα. Η υποστήριξη και η εμπιστοσύνη χρησιμοποιούνται για την εύρεση των κανόνων. Ωστόσο, πρέπει να σημειωθεί ότι υψηλά ποσοστά υποστήριξης και εμπιστοσύνης δεν εξασφαλίζουν ότι ο κανόνας αναδεικνύει μια πραγματική σχέση. Έχουν προταθεί και άλλα μέτρα για την αξιολόγηση των Κανόνων Συσχέτισης.

8.3 Εξόρυξη Κανόνων Συσχέτισης

Η ανακάλυψη Κανόνων Συσχέτισης δεν είναι ένα εύκολο καθήκον εξαιτίας του όγκου των δεδομένων. Η απλούστερη εκδοχή εύρεσης κανόνων είναι να δημιουργηθούν όλοι οι δυνατοί κανόνες, στη συνέχεια να υπολογιστεί η υποστήριξη και η εμπιστοσύνη για κάθε έναν από αυτούς και τέλος να διατηρηθούν μόνο όσοι κανόνες έχουν για αυτά τα δύο μέτρα τιμές μεγαλύτερες ή ίσες από τις καθορισμένες τιμές κατωφλίου. Ένας τέτοιος τρόπος επίλυσης του προβλήματος είναι πρακτικά αδύνατος. Για *k* στοιχεία το πλήθος των δυνατών στοιχειοσυνόλων *m* δίνεται από τη σχέση $m=2^k-1$. Αυτό σημαίνει ότι για 20 μόλις στοιχεία, το πλήθος των δυνατών στοιχειοσυνόλων είναι περίπου 1.000.000. Σε ένα πραγματικό πρόβλημα, το *k* μπορεί να ανέρχεται σε εκατοντάδες ή και χιλιάδες στοιχεία. Σε τέτοιες περιπτώσεις ο έλεγχος όλων των δυνατών στοιχειοσυνόλων είναι απλώς αδύνατος.

Έχουν προταθεί πιο αποτελεσματικές μέθοδοι για την ανακάλυψη Κανόνων Συσχέτισης, των οποίων η υποστήριξη και η εμπιστοσύνη υπερβαίνουν ένα προκαθορισμένο κατώφλι. Η διαδικασία ανακάλυψης Κανόνων Συσχέτισης ολοκληρώνεται σε δύο στάδια:

- Στο πρώτο στάδιο **εντοπίζονται τα συχνά στοιχειοσύνολα**. Τα στοιχειοσύνολα αυτά έχουν υποστήριξη μεγαλύτερη ή ίση από την τιμή κατωφλιού.
- Στο δεύτερο στάδιο **δημιουργούνται οι κανόνες σχέσης από τα συχνά στοιχειοσύνολα**. Οι κανόνες ικανοποιούν τη συνθήκη της εμπιστοσύνης.

8.3.1 Εντοπισμός συχνών στοιχειοσυνόλων – Ο αλγόριθμος Apriori

Η εύρεση συχνών στοιχειοσυνόλων είναι ένα ενδιαφέρον πρόβλημα. Ο πρώτος σχετικός αλγόριθμος που προτάθηκε ονομάζεται Apriori (Agrawal & Srikant, 1994). Το όνομα του οφείλεται στο γεγονός ότι χρησιμοποιεί προηγούμενη (prior) γνώση σχετικά με τη συχνότητα *k*-Στοιχειοσυνόλων, για να βρει συχνά (*k+1*)-Στοιχειοσύνολα.

Η προηγούμενη γνώση που χρησιμοποιείται αφορά την **αντιμονότονη ιδιότητα της υποστήριξης**. Η ιδιότητα αυτή ορίζει ότι η υποστήριξη ενός στοιχειοσυνόλου είναι ίση ή μικρότερη από την υποστήριξη κάθε δυνατού υποσυνόλου του. Η αντιμονότονη ιδιότητα της υποστήριξης μαθηματικά ορίζεται με τη Σχέση 8.3

$$\forall X, Y: (X \subseteq Y) \Longrightarrow supp(X) \geq supp(Y) \tag{8.3}$$

Από τη Σχέση 8.3 προκύπτει ότι για να είναι ένα στοιχειοσύνολο συχνό πρέπει όλα τα μη κενά υποσύνολα του να είναι επίσης συχνά. Αντιστρόφως, εάν ένα στοιχειοσύνολο είναι μη συχνό, τότε η πρόσθεση σε αυτό

ενός νέου στοιχείου δεν μπορεί να δημιουργήσει ένα νέο συχνό στοιχειοσύνολο. Τα υπερσύνολα ενός μη συχνού στοιχειοσυνόλου είναι μη συχνά. Οι επιπτώσεις της αντιμονότονης ιδιότητας της υποστήριξης παρουσιάζονται στο Σχήμα 8.1. Το πλέγμα απεικονίζει τα δυνατά υποσύνολα του 5-στοιχειοσυνόλου *{Α,Β,Γ,Δ,Ε}*. Συνολικά υπάρχουν 30 υποσύνολα, εξαιρουμένου του κενού. Αν γνωρίζουμε ότι το στοιχειοσύνολο *{Α,Β}* είναι μη συχνό τότε γνωρίζουμε ότι και τα υπερσύνολα του, *{Α,Β,Γ}, {Α,Β,Δ}, {Α,Β,Ε}, {Α,Β,Γ,Δ}, {Α,Β,Γ,Ε}* και *{Α,Β,Δ,Ε}* είναι επίσης μη συχνά. Επίσης, το στοιχειοσύνολο *{Γ,Ε}* είναι μη συχνό, οπότε και τα υπερσύνολα του είναι μη συχνά. Όπως φαίνεται στο Σχήμα 8.1, το μέγεθος του προβλήματος περιορίζεται αισθητά. Όλοι οι κόμβοι οι οποίοι είναι χρωματισμένοι με κόκκινο χρώμα είναι υπερσύνολα μη συχνών στοιχειοσυνόλων. Για τους κόμβους αυτούς δεν χρειάζεται να καταμετρηθεί η συχνότητα εμφάνισης τους στη βάση δεδομένων.

Ο αλγόριθμος Apriori περιλαμβάνει μια επαναλαμβανόμενη διαδικασία. Κατά τη διάρκεια των επαναλήψεων πραγματοποιούνται δυο βήματα:

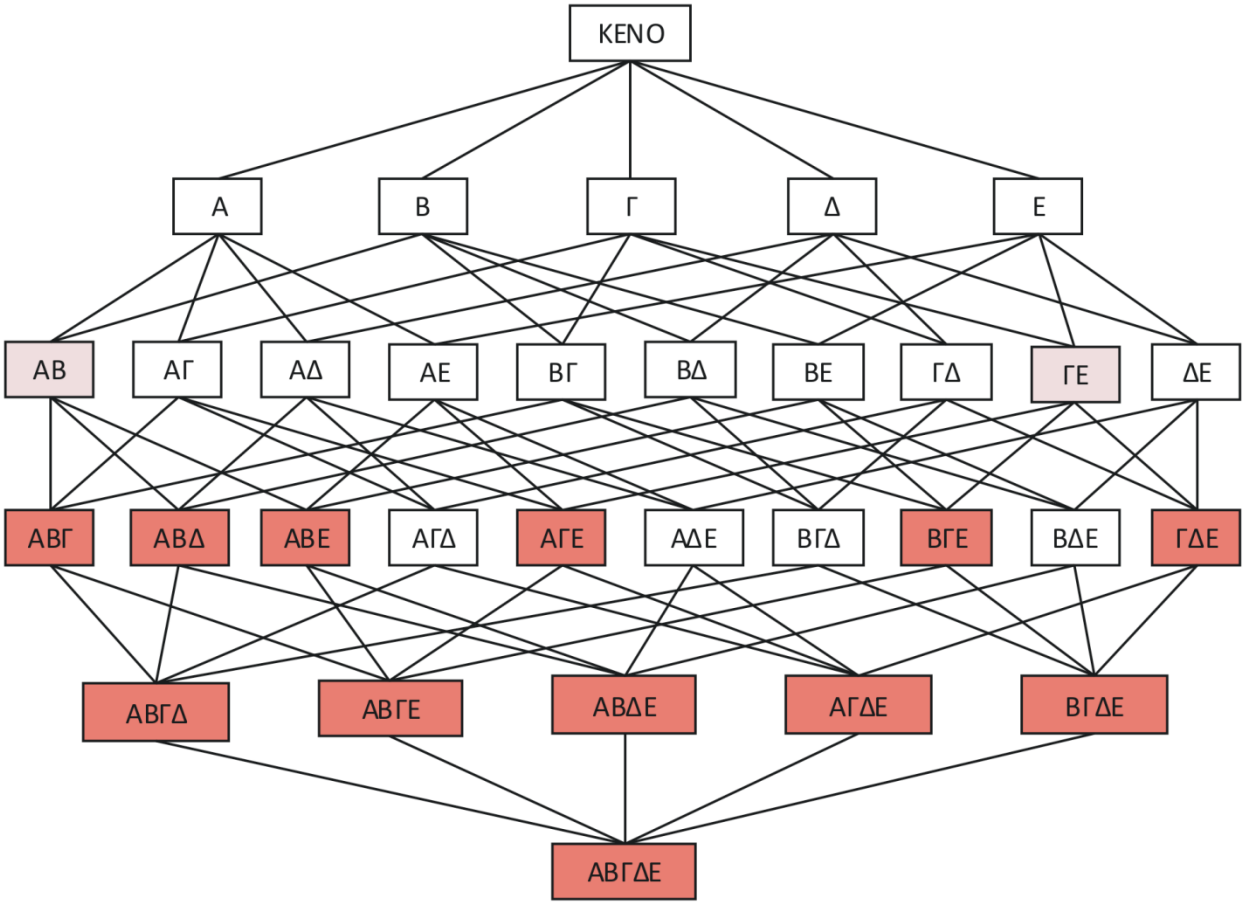
- Δημιουργούνται *k*-στοιχειοσύνολα από τα συχνά (k-1)-στοιχειοσύνολα. Αν L_{k-1} είναι το σύνολο των συχνών (*k-1*)στοιχειοσυνόλων, τότε δημιουργείται ένα σύνολο C_k υποψήφιων *k*-στοιχειοσυνόλων, σύμφωνα με τη Σχέση 8.4.

$$C_k = L_{k-1} \bowtie L_{k-1} \tag{8.4}$$

- Θεωρούμε ότι τα στοιχεία των στοιχειοσυνόλων είναι διατεταγμένα με αλφαβητική σειρά, πχ *{Α,Γ,Δ}*. Δύο μέλη του L_{k-1} μπορούν να συνδεθούν όταν τα πρώτα *k-2* μέλη τους είναι κοινά. Το *{Α,Β,Γ}* μπορεί να συνδεθεί με το *{Α.Β,Δ}* γιατί τα δύο πρώτα στοιχεία (*Α* και *Β*) είναι κοινά. Το αποτέλεσμα είναι το σύνολο *{Α,Β,Γ,Δ}*. Αντιθέτως τα *{Α,Β,Γ}* και *{Α,Γ,Δ}* δεν μπορούν να συνδεθούν γιατί τα δύο πρώτα στοιχεία τους (*Α,Β* και *Α,Γ*) δεν είναι κοινά. Το C_k περιέχει *k*-στοιχειοσύνολα που μπορεί να είναι ή να μη είναι συχνά, όμως όλα τα συχνά *k*-στοιχειοσύνολα ανήκουν στο C_k.
- Στο δεύτερο βήμα πραγματοποιείται κλάδεμα στο C_k. Όλα τα μη συχνά *k*-στοιχειοσύνολα απομακρύνονται και λαμβάνουμε το σύνολο L_k των συχνών *k*-στοιχειοσυνόλων. Για το κλάδεμα του C_k χρησιμοποιείται η αντιμονότονη ιδιότητα της υποστήριξης. Κάθε *k*-στοιχειοσύνολο ελέγχεται για το εάν κάθε (*k-1*)-υποσύνολο του είναι συχνό, αν δηλαδή ανήκει στο L_{k-1}. Εάν κάποιο (*k-1*)-υποσύνολο δεν είναι συχνό τότε και το *k*-στοιχειοσύνολο δεν είναι συχνό.

Αναλυτικότερα, ο αλγόριθμος Apriori περιλαμβάνει τα παρακάτω βήματα:

1. Αρχικά ελέγχεται η βάση δεδομένων και γίνεται καταμέτρηση των εμφανίσεων του κάθε στοιχείου.
2. Δημιουργείται ένα σύνολο με εκείνα τα στοιχεία που η συχνότητα εμφάνισης τους ισούται με ή υπερβαίνει το καθορισμένο κατώφλι. Το σύνολο αυτό μπορεί να θεωρηθεί ως το σύνολο L_k των συχνών *k*-στοιχειοσυνόλων όπου *k*=1.
3. Από τα συχνά *k*-στοιχειοσύνολα του L_k δημιουργείται ένα σύνολο με (*k+1*)-στοιχειοσύνολα. Το νέο σύνολο C_{k+1} δημιουργείται με τη συνένωση του L_k με τον εαυτό του (C_{k+1}=$L_k \bowtie L_k$).
4. Το C_{k+1} ελέγχεται για ύπαρξη μη συχνών στοιχειοσυνόλων σύμφωνα με την αντιμονότονη ιδιότητα της υποστήριξης. Όσα (*k+1*)-στοιχειοσύνολα βρεθούν να περιέχουν τουλάχιστον ένα μη συχνό *k*-στοιχειοσύνολο διαγράφονται από το C_{k+1}.
5. Πραγματοποιείται έλεγχος στη βάση δεδομένων και υπολογίζεται η υποστήριξη για τα εναπομείναντα μέλη του C_{k+1}. Όσα μέλη του C_{k+1} έχουν υποστήριξη μικρότερη από την τιμή κατωφλιού διαγράφονται από το C_{k+1}. Με την ολοκλήρωση αυτού του βήματος έχουμε βρει το σύνολο των συχνών (*k+1*)-στοιχειοσυνόλων L_{k+1}.
6. Αν το L_{k+1} δεν είναι κενό επανερχόμαστε στο βήμα 3.

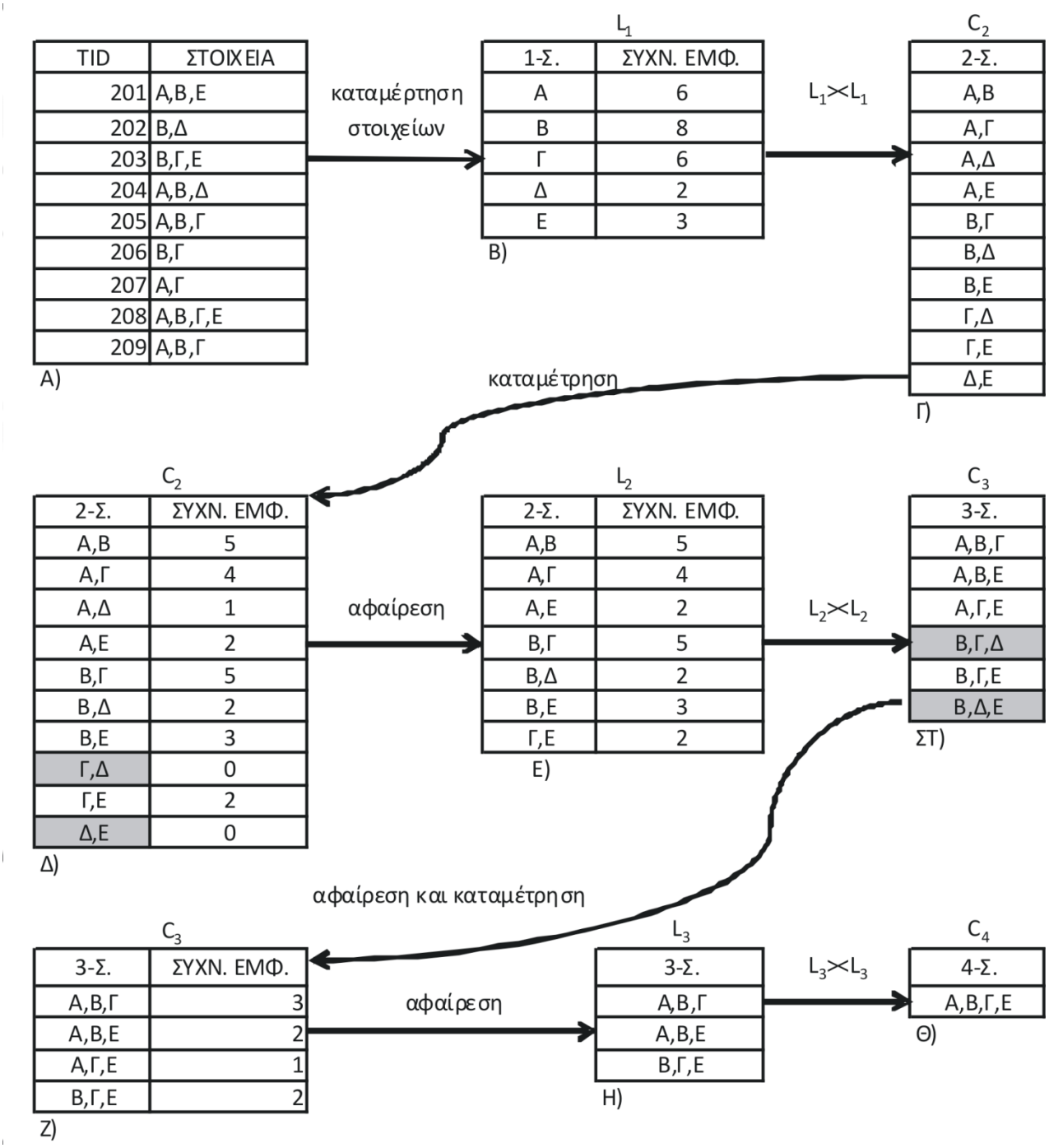


Σχήμα 8.1 *Μείωση πλήθους υποψήφιων στοιχειοσυνόλων λόγω αντιμονότονης ιδιότητας της υποστήριξης*

Για την καλύτερη κατανόηση του αλγορίθμου Apriori παραθέτουμε το παράδειγμα του Σχήματος 8.2.

Στο τμήμα Α) του Σχήματος 8.2 απεικονίζονται τα δεδομένα. Η βάση δεδομένων περιέχει εννέα συνολικά συναλλαγές. Για κάθε συναλλαγή καταγράφονται τα εμπορεύματα που πωλήθηκαν ως *Α, Β* κλπ. Ορίζουμε ότι θέλουμε να βρούμε συχνά στοιχειοσύνολα με συχνότητα εμφάνισης ίση ή μεγαλύτερη από δύο. Η βάση δεδομένων σαρώνεται και υπολογίζεται η συχνότητα εμφάνισης κάθε εμπορεύματος. Τα αποτελέσματα παρουσιάζονται στο τμήμα Β). Μπορούμε να θεωρήσουμε το σύνολο των εμπορευμάτων ως το σύνολο *1*-στοιχειοσυνόλων C_1.Παρατηρούμε ότι η συχνότητα εμφάνισης κάθε εμπορεύματος είναι ίση ή μεγαλύτερη από δύο. Αυτό σημαίνει ότι όλα τα *1*-στοιχειοσύνολα είναι συχνά και κανένα δεν πρέπει να διαγραφεί. Σε αυτήν την περίπτωση C_1=L_1. Στη συνέχεια, το L_1 συνενώνεται με τον εαυτό του και προκύπτει το σύνολο *2*-στοιχειοσυνόλων C_2 το οποίο απεικονίζεται στο τμήμα Γ).

Πραγματοποιείται καταμέτρηση των εμφανίσεων των μελών του C_2 στη βάση δεδομένων, και τα αποτελέσματα απεικονίζονται στο τμήμα Δ). Παρατηρούμε ότι τρία στοιχειοσύνολα, το *{Α,Δ}*, το *{Γ,Δ}* και το *{Δ,Ε}*, έχουν συχνότητα εμφάνισης μικρότερη από δύο. Τα τρία αυτά στοιχειοσύνολα απομακρύνονται από το C_1 και προκύπτει το σύνολο των συχνών *2*-στοιχειοσυνόλων L_2, που παρουσιάζεται στο τμήμα Ε). Στη συνέχεια, το L_2 συνενώνεται με τον εαυτό του. Για να μπορούν να συνδυαστούν δύο *2*-στοιχειοσύνολα, μέλη του L_2, πρέπει το πρώτο μέλος τους να είναι κοινό. Σύμφωνα με αυτόν τον κανόνα, το *{Α,Β}* συνδυάζεται με το *{Α,Γ}* και το *{Α,Ε}*. Επίσης, το *{Α,Γ}* συνδυάζεται με το *{Α,Ε}*. Με τον ίδιο τρόπο συνδυάζονται και τα μέλη του L_2 των οποίων το πρώτο στοιχείο είναι το *Β*. Το *{Γ,Ε}* δεν μπορεί να συνδυαστεί με κανένα μέλος του L_2. Το αποτέλεσμα της συνένωσης είναι το σύνολο C_3 και απεικονίζεται στο τμήμα ΣΤ).



Σχήμα 8.2 *Apriori - Παράδειγμα*

Πραγματοποιείται έλεγχος στο C_3 για να βρεθούν μέλη που έχουν μη συχνά υποσύνολα. Δύο μέλη του C_3 έχουν μη συχνά υποσύνολα. Ειδικότερα, το *{Β,Γ,Δ}* έχει υποσύνολο το *{Γ,Δ}* το οποίο είναι μη συχνό καθώς δεν ανήκει στο L_2. Επίσης, το *{Β,Δ,Ε}* έχει υποσύνολο το *{Δ,Ε}* το οποίο είναι μη συχνό. Από την αντιμονότονη ιδιότητα της υποστήριξης συμπεραίνουμε ότι και τα *{Β,Γ,Δ}*, *{Β,Δ,Ε}* είναι μη συχνά. Για τον λόγο αυτό, τα δύο στοιχειοσύνολα απομακρύνονται από το C_3, και για τα υπόλοιπα στοιχειοσύνολα του πραγματοποιείται καταμέτρηση στη βάση δεδομένων. Τα αποτελέσματα παρουσιάζονται στο τμήμα Ζ). Παρατηρούμε ότι ένα μέλος του C_3, το *{Α,Γ,Ε}*, έχει συχνότητα εμφάνισης ίση με ένα. Το στοιχειοσύνολο αυτό είναι μη συχνό και απομακρύνεται από το C_3. Το αποτέλεσμα είναι το σύνολο συχνών *3*-στοιχειοσυνόλων L_3, το οποίο παρουσιάζεται στο τμήμα Η). Στη συνέχεια, το L_3 συνενώνεται με τον εαυτό του. Τα μέλη του που μπορούν να συνδυαστούν είναι το *{Α,Β,Γ}* και το *{Α,Β,Ε}*, γιατί αυτά τα δύο στοιχειοσύνολα έχουν τα δυο πρώτα μέλη τους κοινά. Το αποτέλεσμα της συνένωσης είναι το C_4, το οποίο απεικονίζεται στο τμήμα Θ). Το C_4 έχει ένα μόνο μέλος, το *{Α,Β,Γ,Ε}*. Παρατηρούμε ότι το *{Α,Β,Γ,Ε}* έχει υποσύνολο το *{Α,Γ,Ε}* το οποίο είναι μη συχνό, οπότε συμπεραίνουμε ότι και το *{Α,Β,Γ,Ε}* είναι μη συχνό. Με την ολοκλήρωση της εκτέλεσης του αλγορίθμου έχουμε εντοπίσει τα συχνά στοιχειοσύνολα, τα οποία είναι τα μέλη του L_2 και του L_3.

8.3.2 Δημιουργία Κανόνων Συσχέτισης από τα συχνά στοιχειοσύνολα

Με την ολοκλήρωση του πρώτου σταδίου έχουν βρεθεί τα συχνά στοιχειοσύνολα της βάσης δεδομένων. Η υποστήριξη αυτών των στοιχειοσυνόλων είναι μεγαλύτερη από την τιμή κατωφλιού υποστήριξης. Τα συχνά στοιχειοσύνολα χρησιμοποιούνται για την εύρεση των κανόνων. Ο αλγόριθμος δημιουργίας κανόνων είναι ο ακόλουθος:

- Για κάθε συχνό στοιχειοσύνολο *Z* δημιουργούνται όλα τα μη κενά υποσύνολα του.
- Για κάθε μη κενό υποσύνολο *X* δημιουργείται κανόνας της μορφής $X \rightarrow Z\text{-}X$.
- Κάθε κανόνας που δημιουργήθηκε στο προηγούμενο βήμα ελέγχεται ως προς την τιμή της εμπιστοσύνης. Αν η εμπιστοσύνη είναι μεγαλύτερη από την τιμή κατωφλιού εμπιστοσύνης, τότε ο κανόνας θεωρείται ισχυρός.

Μετά την εκτέλεση του αλγορίθμου έχουν εντοπιστεί όλοι οι ισχυροί κανόνες, οι οποίοι έχουν υποστήριξη και εμπιστοσύνη μεγαλύτερη από τις ελάχιστες καθορισμένες τιμές. Σημειωτέον ότι οι κανόνες δημιουργήθηκαν από συχνά στοιχειοσύνολα, οπότε είναι εξασφαλισμένο ότι η υποστήριξη τους ισούται ή υπερβαίνει την τιμή κατωφλιού.

- Η εμπιστοσύνη ενός κανόνα της μορφής $X \rightarrow Y$ όπου $Y=Z\text{-}X$ υπολογίζεται σύμφωνα με τη Σχέση 8.5

$$conf(X \rightarrow Y) = P(Y|X) = \frac{\sup\_count(X \cup Y)}{\sup\_count(X)} = \frac{\sup\_count(Z)}{\sup\_count(X)} \tag{8.5}$$

Σημειωτέον ότι και το *{X}* και το *{Z-X}* είναι συχνά στοιχειοσύνολα, οπότε η συχνότητα εμφάνισης τους είναι γνωστή ήδη από το πρώτο στάδιο, και η υποστήριξη του κανόνα $X \rightarrow Z\text{-}X$ μπορεί να υπολογιστεί εύκολα. Στο παράδειγμα του Σχήματος 8.3 παρουσιάζεται η δημιουργία κανόνων για το συχνό στοιχειοσύνολο *{Α,Β,Γ}* του προηγούμενου παραδείγματος.

Στοιχειοσύνολο	sup_count
Α	6
Β	8
Γ	6
Δ	2
Ε	3
Α,Β	5
Α,Γ	4
Α,Ε	2
Β,Γ	5
Β,Δ	2
Β,Ε	3
Γ,Ε	2
Α,Β,Γ	3

Α,Β,Γ

Κανόνας	Εμπιστοσύνη	
Α→Β,Γ	sup_count(Α,Β,Γ)/sup_count(Α)	50%
Β→Α,Γ	sup_count(Α,Β,Γ)/sup_count(Β)	38%
Γ→Α,Β	sup_count(Α,Β,Γ)/sup_count(Γ)	50%
Α,Β→Γ	sup_count(Α,Β,Γ)/sup_count(Α,Β)	60%
Α,Γ→Β	sup_count(Α,Β,Γ)/sup_count(Α,Γ)	75%
Β,Γ→Α	sup_count(Α,Β,Γ)/sup_count(Β,Γ)	60%

Σχήμα 8.3 *Δημιουργία κανόνων και υπολογισμός εμπιστοσύνης*

Από το συχνό στοιχειοσύνολο *{Α,Β,Γ}* δημιουργούνται οι κανόνες *(Α→Β,Γ)*, *(Β→Α,Γ)*, *(Γ→Α,Β)*, (*Α,-Β→Γ),(Α,Γ→Β)* και *(Β,Γ→Α)*. Οι συχνότητες εμφάνισης των στοιχειοσυνόλων που συμμετέχουν στους κανόνες είναι γνωστές, οπότε υπολογίζονται εύκολα οι τιμές εμπιστοσύνης των κανόνων. Αν έχουμε ορίσει ως κατώφλι εμπιστοσύνης την τιμή 60%, τότε μόνον οι κανόνες *(Α,Β→Γ)*, *(Α,Γ→Β)* και *(Β,Γ→Α)* είναι ισχυροί.

Από ένα στοιχειοσύνολο μεγέθους *k* δημιουργούνται συνολικά 2^k-2 κανόνες. Αυτό σημαίνει ότι για ένα στοιχειοσύνολο με 100 μέλη δημιουργούνται περίπου 10^{30} υποψήφιοι κανόνες. Το πλήθος των υποψήφιων κανόνων μπορεί να είναι πολύ μεγάλο. Για τον λόγο αυτό είναι χρήσιμο να εφαρμοστούν τεχνικές περιορισμού του. Όπως φαίνεται από τη Σχέση 8.5, για κανόνες της μορφής $X \rightarrow Z\text{-}X$ που δημιουργούνται από ένα στοιχειοσύνολο *Z*, η εμπιστοσύνη επηρεάζεται από τη συχνότητα εμφάνισης του *X*, αφού η συχνότητα εμφάνισης

του Z είναι η ίδια για όλους τους κανόνες. Η συχνότητα εμφάνισης του X βρίσκεται στον παρονομαστή της Εξίσωσης 8.5, οπότε και είναι αντιστρόφως ανάλογη με την εμπιστοσύνη. Από την αντιμονότονη ιδιότητα της υποστήριξης γνωρίζουμε ότι εάν $X_1 \subseteq X_2$ τότε ισχύει ότι $sup\_count(X_1) >= sup\_count(X_2)$ και κατά συνέπεια θα ισχύει ότι $conf(X_2 \rightarrow Z\text{-}X_2) >= conf(X_1 \rightarrow Z\text{-}X_1)$. Σε ελεύθερη διατύπωση αυτό σημαίνει ότι ένας κανόνας της μορφής $X \rightarrow Z\text{-}X$ έχει εμπιστοσύνη ίση ή μεγαλύτερη από κάθε κανόνα που στο αριστερό του σκέλος έχει ένα υποσύνολο του X. Τη διαπίστωση αυτή μπορούμε να την επαληθεύσουμε και από τα στοιχεία του Σχήματος 8.3. Το πρακτικό συμπέρασμα που προκύπτει είναι ότι εάν η εμπιστοσύνη ενός κανόνα $X \rightarrow Z\text{-}X$ είναι μικρότερη από την τιμή κατωφλιού, τότε και όλοι ο κανόνες που στο αριστερό τους σκέλος έχουν ένα υποσύνολο του X θα είναι και αυτοί μη ισχυροί.

8.4 Πρόσθετα κριτήρια αποτίμησης των κανόνων.

Η ανάλυση Κανόνων Συσχέτισης, όπως έχει περιγραφτεί μέχρι τώρα, αποδίδει κανόνες, οι οποίοι ικανοποιούν το κριτήριο των τιμών της ελάχιστης υποστήριξης και εμπιστοσύνης και επομένως είναι ισχυροί. Αυτό όμως δεν σημαίνει και ότι οι κανόνες είναι ενδιαφέροντες ή ότι συνιστούν θετικές συσχετίσεις μεταξύ του αριστερού και του δεξιού σκέλους του κανόνα. Καταρχάς, ορισμένοι κανόνες μπορεί να αναδεικνύουν πληροφορίες οι οποίες είναι αδιάφορες στον αναλυτή. Επίσης, ορισμένοι κανόνες μπορεί να αναδεικνύουν τετριμμένες πληροφορίες, οι οποίες είναι ήδη γνωστές. Μια περίπτωση ανακάλυψης τετριμμένης γνώσης είναι η ακόλουθη. Σε έρευνα σχετικά με την πολιτική δανειοδοτήσεων σε μια τράπεζα αναλύθηκαν τα δεδομένα με χρήση Κανόνων Συσχέτισης. Η βάση δεδομένων περιείχε την απόφαση έγκρισης ή απόρριψης του δανείου, καθώς και τα ατομικά στοιχεία των πελατών που έκαναν αίτηση, όπως το επάγγελμα, τα ετήσια εισοδήματα, την περιουσιακή κατάσταση, την οικογενειακή κατάσταση κλπ. Ένας από τους ισχυρότερους κανόνες που ανακαλύφθηκαν με υποστήριξη 30% και εμπιστοσύνη 94% ήταν ο κανόνας *οικογενειακή κατάσταση(ανύπανδρος) → αριθμός παιδιών(κανένα)*. Ο κανόνας αυτός, αν και ισχυρός, ήταν και αδιάφορος για τον αναλυτή και προφανής. Η ανάλυση Κανόνων Συσχέτισης αποδίδει πολλούς κανόνες, που αναδεικνύουν διαφόρων ειδών πληροφορίες. Τελικός αρμόδιος για την αξιολόγηση τους είναι ο αναλυτής. Ωστόσο, μπορούν να εφαρμοστούν πρόσθετα αντικειμενικά κριτήρια, τα οποία θα περιορίσουν τον αριθμό των κανόνων και θα καταδείξουν πρόσθετα στοιχεία σχετικά με το είδος της συσχέτισης μεταξύ του αριστερού και του δεξιού σκέλους των κανόνων.

Ένα αντικειμενικό κριτήριο αποτίμησης των κανόνων είναι το στατιστικό μέτρο **Lift**. Το Lift είναι ένα ποσοτικό μέτρο, το οποίο δείχνει πόσο καλύτερη ή χειρότερη είναι η επίδοση ενός κανόνα σε σχέση με έναν κανόνα τυχαίας επιλογής, αποδίδει δηλαδή τον βαθμό και το είδος συσχέτισης μεταξύ των γεγονότων του κανόνα. Από τη Στατιστική γνωρίζουμε ότι αν δύο γεγονότα X κα Y είναι μεταξύ τους ανεξάρτητα, τότε η πιθανότητα ταυτόχρονης εμφάνισης τους ισούται με το γινόμενο των πιθανοτήτων της εμφάνισης του κάθε γεγονότος ξεχωριστά, όπως φαίνεται και από την Εξίσωση 8.6.

$$P(X \wedge Y) = P(X) * P(Y) \tag{8.6}$$

Αν δύο γεγονότα είναι θετικά συσχετισμένα, αν δηλαδή η εμφάνιση του ενός ενισχύει την πιθανότητα εμφάνισης του άλλου, τότε η πιθανότητα ταυτόχρονης εμφάνισης τους είναι μεγαλύτερη από το γινόμενο των πιθανοτήτων εμφάνισης του κάθε γεγονότος. Αντιστρόφως, αν δύο γεγονότα είναι αρνητικά συσχετισμένα και η εμφάνιση του ενός αποτελεί αντικίνητρο για την εμφάνιση του άλλου, τότε η πιθανότητα ταυτόχρονης εμφάνισης τους είναι μικρότερη από το γινόμενο των πιθανοτήτων εμφάνισης του καθενός ξεχωριστά.

Σε ότι αφορά τους Κανόνες Συσχέτισης, αν θεωρήσουμε έναν κανόνα της μορφής $X \rightarrow Y$ και η πιθανότητα εμφάνισης του κανόνα είναι ίση με το γινόμενο των πιθανοτήτων εμφάνισης του αριστερού και δεξιού σκέλους, τότε τα δύο σκέλη είναι μεταξύ τους ανεξάρτητα. Ένας κανόνας είναι πραγματικά ενδιαφέρων όταν η πιθανότητα εμφάνισης του διαφέρει από το γινόμενο $P(X)*P(Y)$ (Piatetsky-Shapiro, 1991). Το στατιστικό μέτρο Lift χρησιμοποιείται για την αποτίμηση του βαθμού ενδιαφέροντος ενός κανόνα. Το Lift ορίζεται από την Εξίσωση 8.7

$$Lift(X \rightarrow Y) = \frac{supp(X \rightarrow Y)}{supp(X) * supp(Y)} \tag{8.7}$$

Αν η τιμή του Lift ισούται με ένα, τότε δεν υπάρχει συσχέτιση μεταξύ των *X* και *Y*, και ο κανόνας εκφράζει ένα τυχαίο γεγονός. Αν η τιμή του Lift είναι μεγαλύτερη από ένα, τότε υπάρχει θετική συσχέτιση μεταξύ των *X* και *Y* και η εμφάνιση του *X* ενισχύει την πιθανότητα εμφάνισης του *Y*. Τέλος, αν η τιμή του Lift είναι μικρότερη από ένα, τότε τα *X* και *Y* είναι αρνητικά συσχετισμένα, και η εμφάνιση του *X* μειώνει την πιθανότητα εμφάνισης του *Y*.

Η σημασία του Lift καταδεικνύεται καλύτερα με το ακόλουθο παράδειγμα. Θεωρούμε στοιχεία πωλήσεων από ένα σούπερ μάρκετ. Το σύνολο των πωλήσεων είναι 1000. Σε 500 από αυτές πωλήθηκε κρασί, σε 750 ξηροί καρποί και σε 300 πωλήθηκαν ταυτόχρονα κρασί και ξηροί καρποί. Ο κανόνας αγοράζει(κρασί) → αγοράζει(ξηροί καρποί) έχει υποστήριξη 30% και εμπιστοσύνη 60%. Σύμφωνα με τα κριτήρια της υποστήριξης και της εμπιστοσύνης είναι ένας ισχυρός κανόνας. Κάποιος θα μπορούσε να συμπεράνει ότι η αγορά κρασιού ενισχύει την πιθανότητα αγοράς ξηρών καρπών. Όμως η πιθανότητα αγοράς ξηρών καρπών στο σύνολο των πωλήσεων είναι 75%, ενώ η πιθανότητα αγοράς ξηρών καρπών με δεδομένο ότι έχει αγοραστεί κρασί είναι 60%. Στην πραγματικότητα η αγορά κρασιού μειώνει την πιθανότητα αγοράς ξηρών καρπών. Η αρνητική συσχέτιση των δύο αυτών προϊόντων φαίνεται από την τιμή του Lift, η οποία είναι 0,8, δηλαδή μικρότερη από 1. Αν οι ξηροί καρποί είχαν πωληθεί σε 500 πωλήσεις και τα υπόλοιπα στοιχεία παρέμεναν αμετάβλητα, τότε το Lift του κανόνα θα ήταν 1,2 και η αγορά του κρασιού θα ενίσχυε την πιθανότητα αγοράς ξηρών καρπών. Τα στοιχεία αυτά παρουσιάζονται στο Σχήμα 8.4

κρασί -> ξηροί καρποί

συναλλαγές	1000
κρασί	500
ξηροί καρποί	750
κρασί+ξηροί καρποί	300

support	0,3
confidence	0,6
lift	0,8

συναλλαγές	1000
κρασί	500
ξηροί καρποί	500
κρασί+ξηροί καρποί	300

support	0,3
confidence	0,6
lift	1,2

Σχήμα 8.4 *Lift Κανόνων Συσχέτισης*

Το Lift είναι ένα χρήσιμο μέτρο για την κατανόηση της συσχέτισης μεταξύ του αριστερού και του δεξιού μέλους του κανόνα. Ουσιαστικά δείχνει τον παράγοντα αύξησης ή μείωσης της πιθανότητας πώλησης του προϊόντος *Y*, με δεδομένο ότι έχει πωληθεί το προϊόν *X*. Κανόνες των οποίων η τιμή Lift είναι μεγαλύτερη από 1 είναι πραγματικά ενδιαφέροντες, καθώς καταδεικνύουν περιπτώσεις προϊόντων που η συχνότητα ταυτόχρονης πώλησης τους υπερβαίνει το τυχαίο.

Εκτός από το Lift, έχουν προταθεί και άλλα μέτρα αποτίμησης των κανόνων, όπως το J-measure (Smyth & Goodman, 1992), το κατευθυνόμενο ασύμμετρο Lift (Brin, Motwani, Ullman & Tsur, 1997), στατιστικοί συντελεστές συσχέτισης (Tan & Kumar, 2002). Για περισσότερες πληροφορίες ο αναγνώστης παραπέμπεται στο Hilderman and Hamilton (2001).

8.5 Εξόρυξη Πολυδιάστατων Κανόνων Συσχέτισης

Όλοι οι κανόνες, οι οποίοι μας απασχόλησαν μέχρι στιγμής, έχουν τη μορφή *αγορά προϊόντος (X)→ αγορά προϊόντος (Y)*. Οι κανόνες αυτοί περιλαμβάνουν ένα μόνο κατηγόρημα, την αγορά προϊόντος και πολλές εμφανίσεις αυτού του κατηγορήματος. Το κατηγόρημα που συμμετέχει στον κανόνα καλείται διάσταση και οι κανόνες που περιέχουν ένα μόνο κατηγόρημα καλούνται **μονοδιάστατοι**.

Τα δεδομένα, που τηρούνται σε σχεσιακές βάσεις δεδομένων, είναι οργανωμένα σε μορφή πινάκων, οι οποίοι έχουν στήλες και γραμμές. Κάθε στήλη αναφέρεται σε μια ιδιότητα των δεδομένων και καλείται πεδίο. Οι πληροφορίες που τηρούνται δεν αναφέρονται μόνο στα προϊόντα που πωλήθηκαν σε κάθε συναλλαγή, αλλά αναφέρονται και σε πολλά άλλα στοιχεία. Τέτοια στοιχεία μπορεί να είναι η ποσότητα του προϊόντος που πωλήθηκε, ατομικά στοιχεία πελάτη, όπως το εισόδημα και η ηλικία, χαρακτηριστικά του προϊόντος κλπ. Ο πρωτότυπος αλγόριθμος Apriori είναι ικανός να εντοπίζει συχνά στοιχειοσύνολα, που θα χρησιμοποιηθούν για την εξόρυξη μονοδιάστατων κανόνων. Ωστόσο, παρουσιάζει ενδιαφέρον η εξόρυξη κανόνων συσχέτισης από

σχεσιακές βάσεις δεδομένων. Οι κανόνες αυτοί εμπλέκουν περισσότερα από ένα πεδία, έτσι ώστε να μελετούν τις σχέσεις μεταξύ αυτών των πεδίων.

Θεωρώντας κάθε πεδίο της βάσης δεδομένων ως ένα κατηγόρημα, οι κανόνες που θα δημιουργηθούν περιέχουν πολλά κατηγορήματα και καλούνται **πολυδιάστατοι**. Παράδειγμα πολυδιάστατου κανόνα είναι ο ακόλουθος:

- *ηλικία(νέος) ΚΑΙ εισόδημα(υψηλό)* → *αγοράζει(tablet).*

Ο παραπάνω κανόνας περιέχει τρία κατηγορήματα, την ηλικία, το εισόδημα και την αγορά. Κάθε ένα από αυτά τα κατηγορήματα συμμετέχει στον κανόνα μόνο μια φορά. Είναι όμως δυνατό σε έναν κανόνα να συμμετέχει το ίδιο κατηγόρημα περισσότερες φορές με διαφορετικές τιμές. Κανόνες αυτής της μορφής καλούνται **υβριδικών διαστάσεων**. Παράδειγμα υβριδικού κανόνα είναι ο ακόλουθος:

- *ηλικία(νέος) ΚΑΙ αγοράζει(tablet)* → *αγοράζει(smartphone).*

Η μέθοδος εξόρυξης Κανόνων Συσχέτισης, η οποία παρουσιάστηκε στο υποκεφάλαιο 8.3.1 και που χρησιμοποιεί τον αλγόριθμο Apriori, είναι σχεδιασμένη για την ανακάλυψη μονοδιάστατων κανόνων. Για την εξόρυξη πολυδιάστατων κανόνων πρέπει να γίνουν τροποποιήσεις στη μέθοδο. Ο Intan (2006) επιχειρεί μια γενίκευση του εννοιολογικού πλαισίου των Κανόνων Συσχέτισης. Αρχικά επαναπροσδιορίζονται οι έννοιες της υποστήριξης και της εμπιστοσύνης, έτσι ώστε να είναι σε συμφωνία με έννοιες του σχεσιακού μοντέλου. Σε μια σχεσιακή βάση δεδομένων ο κανόνας $X \rightarrow Y$ αναφέρεται σε γραμμές ενός πίνακα όπου τα X και Y εμφανίζονται ταυτόχρονα. Σε αυτήν την περίπτωση, η υποστήριξη και η εμπιστοσύνη μπορούν να οριστούν ως εξής:

$$supp = \frac{\gamma\rho\alpha\mu\mu\acute{\epsilon}\varsigma(X\ \kappa\alpha\iota\ Y)}{\gamma\rho\alpha\mu\mu\acute{\epsilon}\varsigma(\acute{o}\lambda\epsilon\varsigma)} \tag{8.8}$$

$$conf = \frac{\gamma\rho\alpha\mu\mu\acute{\epsilon}\varsigma(X\ \kappa\alpha\iota\ Y)}{\gamma\rho\alpha\mu\mu\acute{\epsilon}\varsigma(X)} \tag{8.9}$$

όπου *γραμμές*(X *και* Y) το πλήθος των γραμμών του πίνακα οι οποίες περιέχουν το X και το Y, *γραμμές*(X) το πλήθος των γραμμών οι οποίες περιέχουν το X και *γραμμές(όλες)* το συνολικό πλήθος των γραμμών του πίνακα. Στη συνέχεια, ο συγγραφέας διατυπώνει διαδοχικούς ορισμούς της υποστήριξης και της εμπιστοσύνης αυξάνοντας τον βαθμό γενίκευσης. Οι τελικοί ορισμοί κάνουν χρήση εννοιών των ασαφών συνόλων και παρέχουν μια θεωρητική βάση για εξόρυξη ασαφών κανόνων συσχέτισης.

Η ταχεία εξόρυξη πολυδιάστατων κανόνων συσχέτισης αποτελεί αντικείμενο μελέτης διαφόρων ερευνητών. Οι Khare, Adlakha and Pardasani (2010) προτείνουν έναν αλγόριθμο για την εξόρυξη πολυδιάστατων κανόνων συσχέτισης, ο οποίος χρειάζεται να σαρώσει μόνον μια φορά τη βάση δεδομένων. Αρχικά η βάση δεδομένων μετασχηματίζεται σε έναν πίνακα δυαδικών τιμών. Για κάθε πεδίο δημιουργούνται τόσες στήλες όσες είναι και οι τιμές που υπάρχουν στο πεδίο. Για παράδειγμα, αν στο πεδίο M υπάρχουν δύο τιμές, $M1$ και $M2$, σε όλη τη βάση δεδομένων, τότε δημιουργούνται δύο πεδία $M1$ και $M2$. Αν σε μια γραμμή της βάσης δεδομένων υπάρχει στο πεδίο M η τιμή $M1$, τότε στον πίνακα δυαδικών τιμών θα μπει η τιμή 1 στο πεδίο $M1$ και η τιμή 0 στο πεδίο $M2$. Για κάθε στήλη καταμετρώνται οι εμφανίσεις του αριθμού 1. Αν το πλήθος των αριθμών 1 είναι μικρότερο από την ελάχιστη υποστήριξη, τότε η στήλη απομακρύνεται από τον πίνακα. Στη συνέχεια, συνδυάζοντας συχνά στοιχειοσύνολα, δημιουργούνται νέα υποψήφια στοιχειοσύνολα και καταμετράται η υποστήριξη τους. Όλοι οι υπολογισμοί γίνονται στον πίνακα δυαδικών τιμών με τη χρήση του λογικού τελεστή *AND* και δεν χρειάζεται νέα σάρωση της βάσης δεδομένων.

8.6 Εξόρυξη Κανόνων Συσχέτισης με διαφορετικά επίπεδα γενίκευσης

Στο Κεφάλαιο 4 αναφερθήκαμε στις ιεραρχίες εννοιών. Μια ιεραρχία εννοιών είναι μια διάταξη εννοιών σύμφωνα με τον βαθμό γενίκευσης, από το ειδικότερο προς το γενικότερο. Για παράδειγμα, τα «σανδάλια» υπάγονται στην ευρύτερη κατηγορία «υποδήματα», η οποία με τη σειρά της υπάγεται στην κατηγορία «είδη

ένδυσης και υπόδησης». Τα «σανδάλια», τα «υποδήματα» και τα «είδη ένδυσης και υπόδησης» συγκροτούν μια ιεραρχία εννοιών.

Είναι αρκετά πιθανό, αντικείμενα τα οποία ανήκουν στο κατώτερο επίπεδο μιας ιεραρχίας εννοιών να παρουσιάζονται σπάνια στις συναλλαγές και να έχουν χαμηλή τιμή υποστήριξης. Κατά την εξόρυξη κανόνων τα αντικείμενα αυτά θα απορριφθούν ως μη συχνά. Αν όμως χρησιμοποιηθούν στην εξόρυξη ευρύτερες έννοιες, τότε είναι πιθανό να προκύψουν ενδιαφέροντες κανόνες. Για παράδειγμα, οι κανόνας *αγοράζει(σανδάλια) → αγοράζει(μπουφάν)* μπορεί να έχει χαμηλή υποστήριξη, όμως ο κανόνας *αγοράζει(υποδήματα) → αγοράζει(μπουφάν)* να έχει υψηλή υποστήριξη και να είναι ισχυρός. Επίσης, διαφορετικοί χρήστες είναι πιθανό να ενδιαφέρονται να πραγματοποιήσουν αναλύσεις σε διαφορετικά επίπεδα γενίκευσης. Για τους λόγους αυτούς δημιουργήθηκαν αλγόριθμοι για την εξόρυξη κανόνων συσχέτισης με διαφορετικά επίπεδα γενίκευσης. Οι αλγόριθμοι αυτοί κάνουν χρήση των ιεραρχιών εννοιών και οι κανόνες που δημιουργούνται καλούνται κανόνες πολλαπλών επιπέδων (multilevel association rules).

Όταν ο αναλυτής εξορύσσει κανόνες πολλαπλών επιπέδων, μπορεί να ορίσει την ίδια τιμή ελάχιστης υποστήριξης για όλα τα επίπεδα. Η μέθοδος αυτή είναι απλή αλλά παρουσιάζει προβλήματα. Αν η τιμή ελάχιστης υποστήριξης είναι μεγάλη, τότε θα χαθούν ενδιαφέροντες κανόνες στα χαμηλότερα επίπεδα γενίκευσης. Ο κανόνας *αγοράζει(σανδάλια) → αγοράζει(μπουφάν)* μπορεί να είναι ενδιαφέρων αλλά να μην εντοπιστεί. Αν πάλι η τιμή υποστήριξης είναι μικρή, τότε θα παραχθούν πάρα πολλοί, μη ενδιαφέροντες κανόνες στα ανώτερα επίπεδα. Η λύση σε αυτό το πρόβλημα, είναι να καθοριστούν διαφορετικές τιμές ελάχιστης υποστήριξης για διαφορετικά επίπεδα, μικρότερες για τα κατώτερα επίπεδα και μεγαλύτερες για τα ανώτερα.

Κατά την εξόρυξη κανόνων πολλαπλών επιπέδων με διαφορετικές τιμές υποστήριξης ανά επίπεδο, μπορεί να γίνει έλεγχος της υποστήριξης όλων των στοιχειοσυνόλων όλων των επιπέδων. Ωστόσο, η διαδικασία αυτή είναι αργή. Εναλλακτικά, μπορεί να χρησιμοποιηθεί γνώση που αφορά ένα επίπεδο για τον έλεγχο στοιχειοσυνόλων χαμηλότερου επιπέδου. Ειδικότερα, αν ένα στοιχείο υψηλού επιπέδου βρεθεί να έχει υποστήριξη μικρότερη της τιμής κατωφλιού που έχει οριστεί γι' αυτό το επίπεδο, τότε τα στοιχειοσύνολα χαμηλότερου επιπέδου που περιέχουν τέκνα αυτού του στοιχείου μπορούν να εξαιρεθούν. Αν για παράδειγμα βρεθεί ότι τα «υποδήματα» έχουν χαμηλή υποστήριξη, τότε και τα στοιχειοσύνολα που περιέχουν τα «σανδάλια» θεωρούνται ότι είναι και αυτά μη συχνά, απορρίπτονται και εξαιρούνται από τον υπολογισμό της υποστήριξης. Η τεχνική αυτή επιταχύνει την εξόρυξη των κανόνων, όμως δεν έχει απόλυτη ισχύ. Ας υποθέσουμε ότι ορίζουμε την τιμή 5 ως κατώφλι υποστήριξης για το επίπεδο των «υποδημάτων» και τιμή κατωφλιού ίση με 2 για το επίπεδο των «σανδαλιών». Είναι δυνατόν τα υποδήματα να έχουν υποστήριξη ίση με 4 και τα σανδάλια υποστήριξη ίση με 2. Σε μια τέτοια περίπτωση οι κανόνες για τα σανδάλια θα είχαν αποκλειστεί εκ των προτέρων, ενώ θα έπρεπε να έχουν διατηρηθεί.

Μια τεχνική για τη σχετική αντιμετώπιση του παραπάνω προβλήματος είναι να καθοριστεί μια νέα τιμή *διάβασης επιπέδου*, η οποία είναι μικρότερη από το κατώφλι υποστήριξης του ανώτερου επιπέδου και μεγαλύτερη από το κατώφλι του κατώτερου επιπέδου. Στοιχειοσύνολα του κατώτερου επιπέδου ελέγχονται, όταν η υποστήριξη του αντικειμένου του ανώτερου επιπέδου είναι ίση ή μεγαλύτερη από την τιμή διάβασης του ανώτερου επιπέδου. Επανερχόμενοι στο προηγούμενο παράδειγμα, μπορούμε να ορίσουμε τιμή διάβασης επιπέδου την τιμή 4. Τα «υποδήματα», τα οποία έχουν τιμή υποστήριξης ίση με 4, θα αποκλειστούν από τους κανόνες του ανώτερου επιπέδου. Όμως η υποστήριξη των «υποδημάτων» είναι ίση με την τιμή διάβασης επιπέδου, οπότε θα γίνει ο έλεγχος υποστήριξης των «σανδαλιών» καθώς και όλων των άλλων στοιχείων, τα οποία υπάγονται στα «υποδήματα». Εφόσον η υποστήριξη των «σανδαλιών» είναι ίση με 2 και η τιμή κατωφλιού υποστήριξης γι' αυτό το επίπεδο είναι επίσης ίση με 2, θα παραχθούν οι ισχυροί κανόνες που περιλαμβάνουν τα σανδάλια.

8.7 Κανόνες Συσχέτισης με πεδία συνεχών τιμών

Η εξόρυξη Κανόνων Συσχέτισης αφορά την ανακάλυψη κανόνων που προέρχονται από τη συχνή ταυτόχρονη εμφάνιση τιμών σε πεδία της βάσης δεδομένων. Τα πεδία, τα οποία χειρίστηκαν οι αλγόριθμοι που μέχρι στιγμής παρουσιάστηκαν, ήταν δυαδικά ή ονομαστικά (nominal). Τα δυαδικά πεδία μπορούν να πάρουν δύο δυνατές τιμές, δηλαδή 0 και 1. Τα ονομαστικά πεδία παίρνουν ονομαστικές τιμές, δηλαδή ακολουθίες χαρακτήρων (λέξεις). Οι ονομαστικές τιμές είναι διακριτές, περιορισμένες σε αριθμό και δεν υπάρχει κάποια ιεράρχηση μεταξύ των τιμών. Τα ονομαστικά δεδομένα καλούνται και ποιοτικά (quantitative) ή κατηγορικά (categorical). Παράδειγμα ονομαστικού πεδίου είναι η ονομασία ενός προϊόντος ή το επάγγελμα. Σε μια βάση δεδομένων όμως περιλαμβάνονται και αριθμητικά πεδία, πεδία δηλαδή των οποίων οι τιμές είναι αριθμοί. Οι αριθμητικές τιμές είναι διατεταγμένες και συνεχόμενες. Για τον λόγο αυτό, τα αριθμητικά πεδία καλούνται και συνεχή

(continuous) ή ποσοτικά (qualitative). Παράδειγμα αριθμητικού πεδίου είναι το εισόδημα.

Η εξόρυξη Κανόνων Συσχέτισης σε αριθμητικά πεδία παρουσιάζει ιδιαιτερότητες. Ας θεωρήσουμε τον κανόνα ηλικία*(X)→ αγοράζει(φορητό υπολογιστή)*. Θεωρητικά θα έπρεπε να δημιουργηθούν ξεχωριστοί κανόνες για τις ηλικίες 30, 31, 32 κλπ. Κανόνες αυτής της μορφής έχουν πολύ μικρή υποστήριξη, γιατί αφορούν ένα πολύ μικρό τμήμα των καταναλωτών. Επιπλέον, δεν καταγράφουν πραγματική πληροφορία, καθώς δεν υπάρχει ουσιαστική ηλικιακή διαφορά μεταξύ αυτών των καταναλωτών. Σε περίπτωση δε, που το αριθμητικό πεδίο περιέχει πραγματικές και όχι ακέραιες τιμές, η εξόρυξη κανόνων είναι πρακτικά αδύνατη.

Για την αντιμετώπιση του προβλήματος εξόρυξης Κανόνων Συσχέτισης από πεδία αριθμητικών τιμών έχουν προταθεί διάφορες τεχνικές, οι περισσότερες από τις οποίες στηρίζονται στη διακριτοποίηση. Διακριτοποίηση είναι η εργασία αντιστοίχισης των αριθμητικών τιμών σε ονομαστικές τιμές, οι οποίες καθορίζουν περιοχές τιμών. Με τη χρήση της διακριτοποίησης, η εξόρυξη των κανόνων μετατρέπεται σε πρόβλημα εύρεσης συχνών συνδυασμών ονομαστικών τιμών. Οι ονομαστικές τιμές δεν περιγράφουν κάποιο εμπόρευμα ή αντικείμενο, αλλά διαστήματα αριθμητικών τιμών. Για τον λόγο αυτό, στη θέση του όρου «σύνολα στοιχείων» (item sets) χρησιμοποιείται ο όρος «σύνολα κατηγορημάτων» (predicate sets).

Αναλυτική αναφορά στη διακριτοποίηση, καθώς και σε συγκεκριμένες σχετικές τεχνικές, γίνεται στο Κεφάλαιο 7. Στο σημείο αυτό θα υπενθυμίσουμε δύο πολύ συνηθισμένες και απλές τεχνικές, τη διακριτοποίηση ίσου πλάτους (equal width (EW)) και τη διακριτοποίηση ίσης συχνότητας (equal frequency (EF)). Στη διακριτοποίηση ίσου πλάτους ορίζονται διαστήματα ίσου μεγέθους. Στη διακριτοποίηση ίσης συχνότητας τα διαστήματα ορίζονται με τέτοιο τρόπο, ώστε κάθε διάστημα να περιέχει ίσο πλήθος τιμών. Και οι δυο αυτές τεχνικές έχουν δεχτεί κριτική για την αποτελεσματικότητα τους να ορίζουν κατάλληλα διαστήματα. Στη διακριτοποίηση ίσου πλάτους και στην περίπτωση ύπαρξης εξαιρέσεων, είναι δυνατό να οριστούν πολλά διαστήματα, τα οποία δεν περιέχουν καμία τιμή. Στη διακριτοποίηση ίσης συχνότητας είναι δυνατό πολύ διαφορετικές τιμές να βρεθούν στο ίδιο διάστημα, ενώ πολύ κοντινές τιμές να βρεθούν σε διαφορετικά διαστήματα. Μεταξύ των δύο μεθόδων προτιμότερη είναι η πρώτη, καθώς τα προβλήματα που μπορεί να δημιουργήσει η δεύτερη είναι σοβαρότερα. Πρόσθετες και πιο περίτεχνες μέθοδοι διακριτοποίησης έχουν προταθεί από ερευνητές.

Για τη δημιουργία Κανόνων Συσχέτισης σε δεδομένα με αριθμητικά πεδία και με τη χρήση διακριτοποίησης, μια δυνατή προσέγγιση είναι να διακριτοποιηθούν τα αριθμητικά πεδία πριν από την καθεαυτό διαδικασία εξόρυξης, ως ένα τυπικό στάδιο της προεπεξεργασίας των δεδομένων. Για κάθε αριθμητικό πεδίο προστίθεται ένα επιπλέον πεδίο που περιέχει τις διακριτοποιημένες τιμές. Ο αλγόριθμος αναζήτησης συχνών συνόλων κατηγορημάτων τροποποιείται κατάλληλα, έτσι ώστε να αναζητά συχνή ταυτόχρονη εμφάνιση τιμών σε περισσότερα πεδία και όχι σε ένα μόνο πεδίο. Για τη διακριτοποίηση μπορεί να εφαρμοστεί κάποια από τις μεθόδους που περιγράφονται στο Κεφάλαιο 7.

Μια εναλλακτική μέθοδος δημιουργίας Κανόνων Συσχέτισης από δεδομένα με πεδία αριθμητικών τιμών προτάθηκε από τους Vannucci and Colla, (2004). Οι συγγραφείς τονίζουν τις αδυναμίες των μεθόδων διακριτοποίησης ίσου πλάτους και ίσης συχνότητας και διερευνούν τη δυνατότητα εφαρμογής των μεθόδων συσταδοποίησης k-Means και Self Organizing Maps για τον καθορισμό των διαστημάτων. Οι δύο αυτές μέθοδοι παρουσιάζονται αναλυτικά στο Κεφάλαιο 11. Η μέθοδος k-Means απέφερε ενθαρρυντικά αποτελέσματα, αποδείχτηκε όμως ευπαθής στην προκαθορισμένη τιμή των συστάδων k. Η μέθοδος SOM δημιουργεί αυτόματα τις συστάδες, χωρίς προκαθορισμό του πλήθους τους, και διατηρεί την κατανομή των παρατηρήσεων εκπαίδευσης, που στην προκείμενη περίπτωση ήταν μονοδιάστατα διανύσματα, τα οποία περιείχαν τις τιμές του αριθμητικού πεδίου. Σύμφωνα με τα αποτελέσματα, η μέθοδος SOM απέδωσε καλύτερα και προτείνεται από τους συγγραφείς ως κατάλληλη για τη διακριτοποίηση των αριθμητικών πεδίων.

Μια διαφορετική προσέγγιση είναι η δυναμική διακριτοποίηση των αριθμητικών πεδίων κατά τη διάρκεια της διαδικασίας εξόρυξης των κανόνων. Οι Lent, Swami and Widom (1997) πρότειναν μια γεωμετρική μέθοδο συσταδοποίησης δισδιάστατων κανόνων συσχέτισης, την οποία ονόμασαν ARCS (Association Rule Clustering System). Η μέθοδος είναι κατάλληλη για τη δημιουργία κανόνων, που στο αριστερό σκέλος τους βρίσκονται δύο αριθμητικά πεδία. Παράδειγμα τέτοιου κανόνα είναι το ακόλουθο:

- *ηλικία(35) ΚΑΙ εισόδημα(30000) → πιστοληπτική ικανότητα(καλή).*

Σύμφωνα με τη μέθοδο αυτή, τα δύο αριθμητικά πεδία διαμορφώνουν ένα σύστημα αξόνων, και οι παρατηρήσεις τοποθετούνται στον χώρο αυτόν ανάλογα με τις τιμές τους. Οι δύο άξονες διακριτοποιούνται έτσι ώστε να σχηματίζουν ένα πλέγμα αποτελούμενο από κελιά. Για κάθε κελί καταγράφεται ο αριθμός των εμφανίσεων της κάθε κατηγορίας. Εφαρμόζοντας μεθόδους συσταδοποίησης, αναζητούνται γειτονικά κελιά τα οποία συγκροτούν τετράγωνες συστάδες και που εκφράζουν γενικότερους κανόνες. Οι κανόνες αυτοί ενοποιούνται σε

άλλους, ευρύτερους κανόνες. Για παράδειγμα οι κανόνες:

- *ηλικία(35) ΚΑΙ εισόδημα(25.000..30.000) → πιστοληπτική ικανότητα(καλή),*
- *ηλικία(36) ΚΑΙ εισόδημα(25.000..30.000) → πιστοληπτική ικανότητα(καλή),*
- *ηλικία (35) ΚΑΙ εισόδημα(20000..25.000) → πιστοληπτική ικανότητα(καλή),*
- *ηλικία (36) ΚΑΙ εισόδημα(20000..25.000) → πιστοληπτική ικανότητα(καλή),*

μπορούν να ενοποιηθούν στον κανόνα:

- *ηλικία(35..36) ΚΑΙ εισόδημα(20.000..30.000) → πιστοληπτική ικανότητα(καλή).*

Με τον τρόπο αυτόν δημιουργούνται κανόνες, οι οποίοι χρησιμοποιούν διαστήματα που εκφράζουν συγκεντρώσεις τιμών δεδομένων.

Η εξόρυξη Κανόνων Συσχέτισης από δεδομένα αριθμητικών τιμών με χρήση δυναμικής διακριτοποίησης είναι ένα ενεργό πεδίο έρευνας. Οι Taboada et al. (2007) προτείνουν μια μέθοδο Γενετικού Προγραμματισμού, η οποία ονομάζεται Genetic Network Programming (GNP). Ο Γενετικός Προγραμματισμός είναι μια εξέλιξη των Γενετικών Αλγορίθμων (παρουσίαση των Γενετικών Αλγορίθμων γίνεται στο Κεφάλαιο 3), όπου στη θέση των χρωμοσωμάτων βρίσκονται προγράμματα υπολογιστών. Τα προγράμματα εξελίσσονται με χρήση της διασταύρωσης και της μετάλλαξης, έτσι ώστε να επιτευχθεί ένας σκοπός. Ο Γενετικός Προγραμματισμός προέκυψε από την επιθυμία να αναπαραχθεί η ευφυΐα των μηχανών, συχνά όμως χρησιμοποιείται ως μέθοδος βελτιστοποίησης. Οι συγγραφείς εφαρμόζουν τη μέθοδο GNP για να χειριστούν τις αριθμητικές τιμές άμεσα, δηλαδή χωρίς να εφαρμοστεί προηγούμενη διακριτοποίηση ως ένα ξεχωριστό βήμα προεπεξεργασίας. Τα αποτελέσματα πειραμάτων όπου χρησιμοποιήθηκαν πραγματικά δεδομένα, έδειξαν ότι η μέθοδος είναι αποτελεσματική για την εξόρυξη Κανόνων Συσχέτισης από δεδομένα με αριθμητικές τιμές.

Ένα άλλο παράδειγμα περίτεχνης μεθόδου, η οποία εξορύσσει Κανόνες Συσχέτισης από δεδομένα με αριθμητικά πεδία με χρήση δυναμικής διακριτοποίησης, είναι η μέθοδος που πρότεινε ο Yang (2013). Στη μέθοδο αυτή συνδυάζεται ο αλγόριθμος Cultural Algorithm, ο οποίος με εξελικτικό τρόπο προσομοιάζει τη διάδοση πεποιθήσεων σε έναν πληθυσμό, με τον αλγόριθμο Immune Algorithm, ο οποίος υπάγεται στην κατηγορία των Τεχνητών Ανοσοποιητικών Συστημάτων (Artificial Immune Systems). Τα Τεχνητά Ανοσοποιητικά Συστήματα είναι ένας κλάδος της Τεχνητής Νοημοσύνης που αντιγράφει το βιολογικό ανοσοποιητικό σύστημα. Ο αναγνώστης, που ενδιαφέρεται για το αντικείμενο, μπορεί να αναζητήσει πρόσθετες πληροφορίες στο βιβλίο των De Castro and Timmis (2002). Οι αλγόριθμοι είναι εξελικτικοί και περιλαμβάνουν τη δημιουργία πληθυσμών και την αξιολόγηση τους. Η κωδικοποίηση των γονιδίων περιλαμβάνει τα σημεία αποκοπής, τα οποία καθορίζουν τα διαστήματα τιμών. Η συνάρτηση συγγένειας είναι μια αξιολόγηση μεταξύ αντιγόνων και αντισωμάτων και σχετίζεται με την υποστήριξη ενός υποψήφιου αντικειμένου. Ο ανοσοποιητικός αλγόριθμος χρησιμοποιείται για τη διακριτοποίηση των συνεχόμενων πεδίων και για τον εντοπισμό Κανόνων Συσχέτισης. Η μέθοδος δοκιμάστηκε με γνωστά σύνολα δεδομένων και αποδείχτηκε αποτελεσματική και αποδοτική, εφόσον επέτυχε να εντοπίσει κανόνες αυξημένης ακρίβειας σε μικρότερο χρόνο. Επίσης, η μέθοδος επιτυγχάνει να ολοκληρώσει σε μια ενιαία διαδικασία τη διακριτοποίηση των αριθμητικών πεδίων, τη μείωση του αριθμού των γνωρισμάτων και την εξόρυξη Κανόνων Συσχέτισης.

Όπως αναφέρθηκε και προηγουμένως, η ανακάλυψη Κανόνων Συσχέτισης από γνωρίσματα αριθμητικών τιμών είναι ένα ενεργό πεδίο έρευνας. Αναμένεται ότι μελλοντικές επιστημονικές εργασίες θα προτείνουν νέες, ακόμα πιο εξελιγμένες τεχνικές.

8.8 Εξόρυξη Κανόνων Συσχέτισης βασισμένη σε περιορισμούς

Η εξόρυξη κανόνων συσχέτισης, όπως έχει περιγραφτεί μέχρι αυτό το σημείο, ανακαλύπτει ενδιαφέρουσες συσχετίσεις που υπάρχουν σε μεγάλα σύνολα δεδομένων και αποφέρει ένα σύνολο κανόνων της μορφής $X \rightarrow Y$, οι οποίοι ικανοποιούν τα κριτήρια της ελάχιστης υποστήριξης και εμπιστοσύνης. Η διαδικασία εξόρυξης Κανόνων Συσχέτισης μπορεί να αναδείξει ενδιαφέροντες κανόνες, ωστόσο συχνά παρουσιάζονται τα παρακάτω προβλήματα:

- Το πλήθος των κανόνων που προκύπτουν είναι υπερβολικά μεγάλο.
- Πολλοί από τους κανόνες δεν έχουν σχέση με το θέμα που μελετά ο αναλυτής και γι' αυτόν τον λόγο τού είναι αδιάφοροι.
- Ο χρόνος εξόρυξης των κανόνων είναι υπερβολικά μεγάλος.

Με άλλα λόγια, ο χρήστης επιβαρύνεται με μεγάλο υπολογιστικό κόστος, δυσανάλογο με το ωφέλιμο αποτέλεσμα. Απάντηση σε αυτά τα προβλήματα δίνει η εξόρυξη Κανόνων Συσχέτισης η οποία είναι βασισμένη σε περιορισμούς. Με τη χρήση περιορισμών επιτυγχάνεται επικέντρωση στο ζήτημα που ενδιαφέρει τον αναλυτή και ταυτόχρονα μειώνεται ο χώρος αναζήτησης λύσεων.

Ένας τρόπος επιβολής περιορισμών στη διαδικασία εξόρυξης Κανόνων Συσχέτισης είναι με τη χρήση μετακανόνων. Οι μετακανόνες είναι πρότυποι κανόνες, οι οποίοι ορίζουν τη συντακτική δομή των κανόνων που θα δημιουργηθούν. Ο μετακανόνας ορίζει το πλήθος των κατηγορημάτων στο αριστερό και δεξιό σκέλος των κανόνων. Επίσης, μπορεί να συγκεκριμενοποιεί ορισμένα κατηγορήματα ή και τιμές. Ένα παράδειγμα μετακανόνα είναι το ακόλουθο:

- $P_1(X)$ *ΚΑΙ* $P_2(Y) \rightarrow P_3(Z)$.

Τα P_1, P_2, P_3 συμβολίζουν μεταβλητές κατηγορημάτων, στη θέση τους δηλαδή μπορεί να βρεθεί οποιοδήποτε κατηγόρημα. Τα *X, Y, Z* συμβολίζουν μεταβλητές τιμών, πχ στη θέση του *X* μπορεί να βρεθεί οποιαδήποτε τιμή υπάρχει στο κατηγόρημα P_1. Με απλά λόγια, ο παραπάνω μετακανόνας σημαίνει «εξόρυξε μόνο κανόνες, οι οποίοι στο αριστερό σκέλος έχουν δύο κατηγορήματα και στο δεξιό ένα κατηγόρημα». Η εφαρμογή του μετακανόνα μπορεί να αποφέρει κανόνες σαν τους ακόλουθους:

- *ηλικία(30..40) ΚΑΙ εισόδημα(25.000..30.000)* → *πιστοληπτική ικανότητα(καλή),*
- *ηλικία(30..40) ΚΑΙ εισόδημα(25.000..30.000)* → *αγοράζει(smartphone).*

Όπως ήδη αναφέρθηκε, ένας μετακανόνας μπορεί να συγκεκριμενοποιεί κατηγορήματα ή και τιμές. Ο μετακανόνας

- *P1(X) ΚΑΙ P2(Y)* → *αγοράζει(smartphone)*

συγκεκριμενοποιεί το κατηγόρημα και την τιμή του στο δεξιό σκέλος των κανόνων που θα εξορυχτούν. Ο μετακανόνας ορίζει ότι πρέπει να εξορυχτούν κανόνες, στους οποίους ο συνδυασμός δύο συνθηκών οδηγεί στην αγορά smartphone. Η εφαρμογή αυτού του μετακανόνα μπορεί να αποφέρει κανόνες σαν τους ακόλουθους:

- *ηλικία(30..40) ΚΑΙ εισόδημα(25.000..30.000)* → *αγοράζει(smartphone),*
- *ηλικία(30..40) ΚΑΙ αγοράζει(tablet)* → *αγοράζει(smartphone).*

Εκτός των μετακανόνων, μπορούν να επιβληθούν πρόσθετοι περιορισμοί. Οι περιορισμοί αυτοί μπορούν να αφορούν σχέσεις μεταξύ συνόλων και υποσυνόλων, καθορισμό σταθερών τιμών για μεταβλητές, καθώς και συναρτήσεις συναθροίσεων. Παράδειγμα εξόρυξης Κανόνων Συσχέτισης με χρήση περιορισμών είναι το εξής: «Ανακάλυψε κανόνες όπου άρρενες πελάτες ηλικίας από 30 έως 40 χρονών πραγματοποιούν αγορές έως 100 ευρώ».

Οι περιορισμοί οι οποίοι επιβάλλονται στην εξόρυξη Κανόνων Συσχέτισης εντάσσονται στις παρακάτω κατηγορίες:

- **Αντιμονότονοι**. Ένας περιορισμός χαρακτηρίζεται αντιμονότονος εάν ισχύει ότι η παραβίαση του περιορισμού από ένα στοιχειοσύνολο συνεπάγεται και την παραβίαση του περιορισμού από κάθε υπερσύνολο του. Ο περιορισμός «το σύνολο της αξίας των αγορών να είναι μικρότερο από 100 ευρώ» είναι αντιμονότονος, γιατί εάν η αξία ενός συνόλου προϊόντων υπερβαίνει τα 100 ευρώ κάθε σύνολο που θα περιέχει επιπλέον προϊόντα θα αξίζει και αυτό περισσότερο από 100 ευρώ.
- **Μονότονοι**. Ένας περιορισμός χαρακτηρίζεται μονότονος εάν ισχύει ότι η ικανοποίηση του περιορισμού από ένα στοιχειοσύνολο συνεπάγεται και την ικανοποίηση του περιορισμού από κάθε υπερσύνολο του. Ο περιορισμός «το σύνολο της αξίας των αγορών να είναι μεγαλύτερο από 100 ευρώ» είναι μονότονος, γιατί εάν η αξία ενός συνόλου προϊόντων υπερβαίνει τα 100 ευρώ τότε και η αξία κάθε υπερσυνόλου θα υπερβαίνει και αυτή τα 100 ευρώ.
- **Περιληπτικοί**. Ένας περιορισμός χαρακτηρίζεται περιληπτικός εάν μπορούμε να δημιουργήσουμε όλα τα στοιχειοσύνολα, τα οποία τον ικανοποιούν, χωρίς να χρειαστεί να υπολογίσουμε την υποστήριξη τους. Ο περιορισμός «αγορές συνόλου προϊόντων, η τιμή του ενός από τα οποία να υπερβαίνει τα 100 ευρώ» είναι περιληπτικός, αφού μπορούμε να δημιουργήσουμε τα σύνολα προϊόντων που θα περιέχουν τουλάχιστον ένα με αξία μεγαλύτερη των 100 ευρώ.

- **Μετατρέψιμοι**. Ορισμένοι περιορισμοί δεν είναι ούτε μονότονοι ούτε αντιμονότονοι, μπορούν να γίνουν όμως τέτοιοι εάν τα αντικείμενα διαταχτούν σε αύξουσα ή φθίνουσα σειρά. Για παράδειγμα, ο κανόνας «αγορά προϊόντων, των οποίων η μέση τιμή είναι μεγαλύτερη από 100 ευρώ» δεν είναι ούτε μονότονος ούτε αντιμονότονος, αφού η προσθήκη ενός πολύ φθηνού ή πολύ ακριβού προϊόντος σε ένα σύνολο προϊόντων που ικανοποιεί τον περιορισμό, μπορεί να αλλάξει σημαντικά τη μέση τιμή και να παραβιάσει τον περιορισμό. Αν όμως τα προϊόντα διαταχτούν σε αύξουσα σειρά τιμής, τότε ο κανόνας μετατρέπεται σε μονότονο. Σε ένα σύνολο προϊόντων που ικανοποιούν τον περιορισμό, η προσθήκη ακόμα ακριβότερων προϊόντων θα δημιουργήσει στοιχειοσύνολα τα οποία επίσης θα ικανοποιούν τον περιορισμό.
- **Μη μετατρέψιμοι**. Πρόκειται για περιορισμούς οι οποίοι δεν μπορούν να μετατραπούν σε μονότονους ή αντιμονότονους με διάταξη των στοιχείων.

Η χρήση περιορισμών επιτρέπει στον χρήστη να επικεντρώσει την έρευνα του σε ζητήματα που τον ενδιαφέρουν. Για την εξόρυξη Κανόνων Συσχέτισης με χρήση περιορισμών μπορούν να υιοθετηθούν δύο προσεγγίσεις. Σύμφωνα με την απλοϊκή προσέγγιση, εντοπίζονται όλα τα στοιχειοσύνολα, και σε επόμενο στάδιο διαγράφονται όσα παραβιάζουν τους περιορισμούς. Η προσέγγιση αυτή επιβαρύνει τη διαδικασία της εξόρυξης με περιττό υπολογιστικό κόστος και προκαλεί ανώφελες καθυστερήσεις. Μια πολύ πιο αποδοτική προσέγγιση είναι η ενσωμάτωση των περιορισμών στη διαδικασία εύρεσης στοιχειοσυνόλων. Με τον τρόπο αυτό, περιορίζεται εκ των προτέρων το πλήθος των στοιχειοσυνόλων, των οποίων η υποστήριξη πρέπει να υπολογιστεί, και η όλη διαδικασία επιταχύνεται σημαντικά. Ένα απλό παράδειγμα είναι η επιβολή περιορισμού με τη χρήση του μετακανόνα $P1(X)\ KAI\ P2(Y) \rightarrow P3(Z)$. Στην περίπτωση αυτή όλα τα στοιχειοσύνολα με μέγεθος τέσσερα ή μεγαλύτερο απορρίπτονται εκ των προτέρων.

Η εξόρυξη κανόνων συσχέτισης με χρήση περιορισμών είναι ένα εκτεταμένο πεδίο έρευνας. Οι Ng, Lakshmanan, Han and Pang (1998) εισήγαγαν τις έννοιες των αντιμονότονων και των περιληπτικών περιορισμών και πρότειναν τον αλγόριθμο CAP, ο οποίος χρησιμοποιεί περιορισμούς σε συνδυασμό με τον Apriori. Οι Pei and Han (2000) ανέπτυξαν τη μέθοδο CFG, η οποία συνδυάζει την εφαρμογή περιορισμών με τη μέθοδο FP Growth. Οι Grahne, Lakshmanan and Wang (2000) εισήγαγαν την έννοια των μονότονων περιορισμών και την αξιοποίησαν για την ανακάλυψη συσχετισμένων στοιχειοσυνόλων. Οι μετατρέψιμοι περιορισμοί προτάθηκαν από τους Pei, Han and Lakshmanan (2001). Πρόσθετες επιστημονικές εργασίες που δημοσιεύτηκαν αργότερα, πρότειναν ακόμα πιο βελτιωμένους αλγορίθμους για την εξόρυξη κανόνων συσχέτισης με χρήση περιορισμών. Οι Kifer, Gehrke, Bucila and White (2003) χρησιμοποιούν περιορισμούς σχετικά με τη διασπορά των κανόνων. Για περισσότερες λεπτομέρειες σχετικά με την εξόρυξη Κανόνων Συσχέτισης με χρήση περιορισμών παραπέμπουμε τον αναγνώστη στο Boulicaut and Jeudy (2005).

8.9 Μελέτη Περίπτωσης – Μοντελοποίηση Αποφάσεων Εξωτερικών Ελεγκτών με χρήση Κανόνων Συσχέτισης

Στον σύγχρονο επιχειρηματικό κόσμο οι ελεγκτικοί μηχανισμοί των επιχειρήσεων γενικότερα και ο εξωτερικός έλεγχος ειδικότερα, αποκτούν όλο και μεγαλύτερη σημασία. Οι λόγοι γι' αυτό το φαινόμενο είναι πολλοί. Καταρχάς, στις σύγχρονες μεγάλες επιχειρήσεις υπάρχει διαχωρισμός μεταξύ της ιδιοκτησίας και της διοίκησης. Η ιδιοκτησία των επιχειρήσεων ανήκει όχι μόνο σε ιδιώτες, αλλά και σε θεσμικούς επενδυτές, όπως ασφαλιστικά ταμεία, εταιρείες συμμετοχών, επενδυτικά funds και άλλους φορείς. Η διοίκηση ανατίθεται σε μάνατζερς, οι οποίοι προσλαμβάνονται γι' αυτόν τον σκοπό. Ερευνητικές εργασίες έχουν δείξει ότι ο διαχωρισμός ιδιοκτησίας και διοίκησης δημιουργεί κίνητρα στους μάνατζερς να λειτουργήσουν προς ίδιο όφελος και σε βάρος των μετόχων (Kane & Velury, 2004). Ένας παράγοντας που συμβάλλει σε αυτήν την κατεύθυνση είναι η ασυμμετρία της πληροφόρησης μεταξύ των ιδιοκτητών και της διοίκησης. Ο εξωτερικός έλεγχος μειώνει αυτήν την ασυμμετρία και διασφαλίζει τα συμφέροντα των ιδιοκτητών. Μια άλλη ευεργετική επίπτωση του εξωτερικού ελέγχου είναι ότι αυξάνει την αξιοπιστία των δημοσιευμένων χρηματοοικονομικών καταστάσεων και με τον τρόπο αυτόν μειώνει τον επενδυτικό κίνδυνο. Η μείωση του επενδυτικού κινδύνου μπορεί να προσελκύσει το ενδιαφέρον των επενδυτών, να αυξήσει τις τιμές των μετοχών και να μειώσει το κόστος του χρήματος. Πέρα όμως από τη διαχρονική αξία του εξωτερικού ελέγχου, στη σημερινή εποχή της οικονομικής κρίσης, με το ρευστό οικονομικό περιβάλλον και τις αποτυχίες μεγάλων επιχειρήσεων, η σημασία του εξωτερικού ελέγχου αναβαθμίζεται ακόμα περισσότερο.

Οι εξωτερικοί ελεγκτές καλούνται να φέρουν σε πέρας ένα έργο πολύ σημαντικό, αλλά ταυτόχρονα δύσκολο και απαιτητικό. Οι αποφάσεις που λαμβάνονται είναι αδόμητες και οι συνθήκες χαρακτηρίζονται από

μεγάλο βαθμό αβεβαιότητας. Επαγγελματικές ενώσεις, όπως η AICPA (American Institute of Certified Public Accountants) εκδίδουν τα ελεγκτικά πρότυπα (Statements of Auditing Standards), τα οποία περιέχουν οδηγίες για την αποτελεσματικότερη διεξαγωγή του ελέγχου. Ταυτόχρονα, μια πολύ πλούσια ερευνητική βιβλιογραφία επανειλημμένως επισημαίνει την ανάγκη εμπλουτισμού των ασκούμενων ελεγκτικών πρακτικών με σύγχρονες και εξελιγμένες τεχνικές ανάλυσης δεδομένων (Calderon & Cheh, 2002; Koskivaara, 2004). Ανταποκρινόμενοι σε αυτήν την ανάγκη, πλήθος ερευνητών εφάρμοσαν μεθόδους εξόρυξης δεδομένων και ανέπτυξαν μοντέλα ικανά να προβλέπουν τις περιπτώσεις κατά τις οποίες οι εξωτερικοί ελεγκτές εκδίδουν δυσμενή σχόλια. Ενδεικτικά αναφέρουμε τις σχετικές μελέτες των Lenard, Alam and Madey (1995) και των Gaganis, Pasiouras and Doumpos (2007). Στις περισσότερες από αυτές τις εργασίες εφαρμόζονται μέθοδοι όπως τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης. Τέτοιες μέθοδοι, αν και επιτυγχάνουν πολύ υψηλά ποσοστά ορθών προβλέψεων, δεν προσφέρουν στους ελεγκτές ένα σύνολο απλών και κατανοητών κανόνων που να περιγράφουν τον τρόπο με τον οποίο λαμβάνονται οι αποφάσεις κατηγοριοποίησης. Γενικώς, οι ερευνητές τείνουν να επικεντρώνουν περισσότερο στην ανάπτυξη μοντέλων υψηλών επιδόσεων και λιγότερο στην ερμηνεία των μοντέλων. Για τους εξωτερικούς ελεγκτές όμως, η γνώση του μηχανισμού λήψης απόφασης είναι ιδιαίτερης σημασίας. Επίσης, η ρύθμιση των παραμέτρων μεθόδων όπως τα Νευρωνικά Δίκτυα απαιτεί εξειδικευμένες γνώσεις, τις οποίες πιθανώς στερούνται αρκετοί ελεγκτές.

Η μέθοδος των Κανόνων Συσχέτισης δεν είναι ένας καθαρόαιμος κατηγοριοποιητής. Όμως, με την εφαρμογή περιορισμών κατά την εξόρυξη, μπορούν να παραχθούν κανόνες, οι οποίοι στο δεξιό τους σκέλος να καταλήγουν σε μια απόφαση κατηγοριοποίησης. Επιπλέον, οι κανόνες αυτοί είναι πλήρως κατανοητοί και έχουν τη μορφή IF (συνθήκη 1) AND (συνθήκη 2) AND … AND (συνθήκη n) THEN (κατηγορία = X). Για την περίπτωση των εξωτερικών ελεγκτών, κανόνες αυτής της μορφής, που περιγράφουν το αποτέλεσμα του ελέγχου, μπορούν να θεωρηθούν ισχυρές ενδείξεις και να χρησιμοποιηθούν για τον σχεδιασμό κατάλληλων ελεγκτικών διαδικασιών. Στην εργασία του Kirkos (2010) επιχειρείται η μοντελοποίηση των αποφάσεων των εξωτερικών ελεγκτών με χρήση των Κανόνων Συσχέτισης. Η εργασία αυτή επεκτείνει προηγούμενη έρευνα (Kirkos, Spathis, Nanopoulos & Manolopoulos, 2007), στην οποία αναπτύχθηκαν και συγκρίθηκαν μοντέλα Δένδρων Αποφάσεων, Νευρωνικών Δικτύων και Μπαϋεσιανών Δικτύων. Τα μοντέλα αυτά επέτυχαν υψηλή ακρίβεια πρόβλεψης έναντι άγνωστων παρατηρήσεων (περίπου 80%), δεν απέφεραν όμως ένα σύνολο απλών κανόνων που να περιγράφουν περιπτώσεις έκδοσης ή μη έκδοσης δυσμενών σχολίων. Το Δένδρο Αποφάσεων, το οποίο ήταν και το μοναδικό ερμηνεύσιμο μοντέλο, περιείχε 55 κόμβους και 28 φύλα, ορισμένοι δε κανόνες περιείχαν μέχρι και 12 λογικές συνθήκες συνδυασμένες με τον τελεστή AND.

Τα δεδομένα της έρευνας προήλθαν από τη βάση οικονομικών δεδομένων FAME (Financial Analysis Made Easy), η οποία περιλαμβάνει στοιχεία για περίπου 3.000.000 βρετανικές και ιρλανδικές επιχειρήσεις. Επιλέχτηκαν επιχειρήσεις οι οποίες ήταν εισηγμένες στο χρηματιστήριου του Λονδίνου και του Δουβλίνου και οι οποίες πήραν δυσμενή σχόλια από τους εξωτερικούς ελεγκτές τη δεκαετία 1995-2004. Συνολικά εντοπίστηκαν 225 περιπτώσεις. Οι επιχειρήσεις αυτές ταιριάστηκαν με ίσο αριθμό επιχειρήσεων οι οποίες δεν πήραν σχόλια. Τα κριτήρια επιλογής ήταν ο τετραψήφιος κωδικός δραστηριότητας (Standard Industry Code), καθώς και το οικονομικό έτος για να εξαλειφθούν μακροοικονομικές επιρροές. Το τελικό σύνολο δεδομένων περιείχε στοιχεία για 450 επιχειρήσεις.

Η αρχική επιλογή μεταβλητών στηρίχθηκε σε προηγούμενες ερευνητικές εργασίες. Υπάρχει πλουσιότατη ερευνητική βιβλιογραφία που αναφέρεται στους παράγοντες, οι οποίοι παίζουν σημαντικό ρόλο στην έκδοση σχολίων. Ενδεικτικά αναφέρονται οι εργασίες των Spathis (2003), Reynolds and Francis (2001) και των Bell and Tabor (1991). Συνολικά επιλέχτηκαν 24 μεταβλητές προερχόμενες από τις χρηματοοικονομικές καταστάσεις. Ακολούθως πραγματοποιήθηκε επιλογή σημαντικών μεταβλητών. Η μέθοδος που εφαρμόστηκε ήταν η Correlation Based Feature Subset Evaluator – CFS, η οποία περιγράφεται στο Κεφάλαιο 7. Η εφαρμογή της μεθόδου απέφερε ένα σύνολο 17 σημαντικών και ανεξάρτητων μεταξύ τους μεταβλητών.

Όπως έχει ήδη αναφερθεί, η εξόρυξη Κανόνων Συσχέτισης από δεδομένα αριθμητικών τιμών απαιτεί ειδικούς χειρισμούς. Το σύνολο δεδομένων που χρησιμοποιήθηκε στη συγκεκριμένη έρευνα περιέχει, με την εξαίρεση του πεδίου της κλάσης, μόνο αριθμητικά πεδία. Για την αντιμετώπιση του προβλήματος εφαρμόστηκε διακριτοποίηση των δεδομένων, ως ένα ξεχωριστό στάδιο προεπεξεργασίας. Η μέθοδος διακριτοποίησης που επιλέχτηκε ήταν η επιβλεπόμενη και βασισμένη στην εντροπία διακριτοποίηση. Η συγκεκριμένη μέθοδος (περιγράφεται στο Κεφάλαιο 7) χρησιμοποιεί τις τιμές της κλάσης για τον καθορισμό των διαστημάτων. Για τον λόγο αυτό, τα διαστήματα τιμών που ορίζονται είναι καλύτερα προσαρμοσμένα στις ανάγκες της κατηγοριοποίησης. Πρέπει να σημειωθεί ότι τα ίδια δεδομένα χρησιμοποιήθηκαν στην εργασία των Kirkos et al. (2007). Στην εργασία αυτή δημιουργήθηκαν ένα μοντέλο Δένδρου Αποφάσεων και ένα μοντέλο Νευρωνικού Δικτύου χρησιμοποιώντας τα αριθμητικά δεδομένα, καθώς επίσης και ένα μοντέλο Μπαϋεσιανού Δικτύου, το

οποίο χρησιμοποιούσε τα διακριτοποιημένα δεδομένα. Σύμφωνα με τα αποτελέσματα των πειραμάτων που διεξήχθηκαν, το Μπαϋεσιανό Δίκτυο επέτυχε υψηλότερο ποσοστό ορθών κατηγοριοποιήσεων έναντι άγνωστων παρατηρήσεων. Το γεγονός αυτό αποδεικνύει ότι οι καθορισμένες περιοχές τιμών περιέχουν ουσιαστική πληροφορία, που σχετίζεται με τις τιμές της κλάσης.

Στο επόμενο στάδιο έγινε εξόρυξη Κανόνων Συσχέτισης από τα διακριτοποιημένα δεδομένα. Εξορύχτηκαν κανόνες οι οποίοι περιέχουν μέχρι και τέσσερα κατηγορήματα συνθηκών και οι οποίοι συνεπάγονται μια απόφαση κατηγοριοποίησης, δηλαδή την απόφαση έκδοσης ή μη έκδοσης δυσμενών σχολίων. Σύμφωνα με την αγγλική ορολογία, οι επιχειρήσεις οι οποίες έλαβαν δυσμενή σχόλια χαρακτηρίζονται «Qualified», ενώ οι επιχειρήσεις που δεν έλαβαν σχόλια χαρακτηρίζονται «Unqualified». Στόχος ήταν η εξόρυξη ενός διαχειρίσιμου και σχετικά μικρού αριθμού κανόνων, με ικανοποιητικές τιμές υποστήριξης, εμπιστοσύνης και Lift. Μετά από αρκετούς πειραματισμούς με τις τιμές της υποστήριξης και της εμπιστοσύνης, εξορύχτηκαν 17 κανόνες για την έκδοση σχολίων και 15 κανόνες για τη μη έκδοση σχολίων Οι κανόνες αυτοί παρουσιάζονται στους Πίνακες 8.2 και 8.3 αντιστοίχως

Η ανάλυση των κανόνων που ανακαλύφθηκαν αποφέρει ενδιαφέροντα συμπεράσματα. Για την περίπτωση των εταιρειών που πήραν σχόλια (Qualified), παρατηρούμε ότι μόνον έξι από τις 17 μεταβλητές συμμετέχουν στους κανόνες. Οι μεταβλητές αυτές είναι οι Κέρδη Προ Φόρων (Profit Before Taxation),Κεφάλαιο Κίνησης (Working Capital),Τάση Τρεχουσών Υποχρεώσεων (Current Liabilities Trend), Κέρδη προς Σύνολο Ενεργητικού (Return on Total Assets), Μη Διανεμόμενα Κέρδη (Retained Profit) και Μετοχικό Κεφάλαιο (Shareholder's Funds). Με μοναδική εξαίρεση την Τάση Τρεχουσών Υποχρεώσεων, όλες οι άλλες μεταβλητές συμμετέχουν στους κανόνες με τις χαμηλότερες τιμές τους. Τρεις από τις μεταβλητές των κανόνων αναφέρονται στην κερδοφορία και συμμετέχουν συνολικά 27 φορές, ενώ οι άλλες μεταβλητές συμμετέχουν σημαντικά λιγότερες φορές. Οι μεταβλητές κερδοφορίας επικρατούν στις συνθήκες των κανόνων. Τα αποτελέσματα αυτά συνιστούν ότι η χαμηλή κερδοφορία σχετίζεται πολύ έντονα με την έκδοση δυσμενών σχολίων.

Σε ότι αφορά τους κανόνες που αναφέρονται στις εταιρείες οι οποίες δεν πήραν σχόλια (Unqualified), είναι ενδιαφέρον ότι δεν περιλαμβάνεται καμία μεταβλητή κερδοφορίας. Το εύρημα αυτό αποτελεί ισχυρή ένδειξη ότι η υψηλή κερδοφορία δεν σχετίζεται με την μη έκδοση σχολίων. Το ZScore του Altman αποτελεί τη μοναδική συνθήκη σε έναν κανόνα και συμμετέχει σε τρεις επιπλέον κανόνες. Μπορούμε να συμπεράνουμε ότι μια υψηλή τιμή ZScore αποτελεί ισχυρή ένδειξη για τη μη έκδοση σχολίων. Επίσης παρατηρούμε ότι ο αριθμοδείκτης Κεφάλαιο Κίνησης προς Σύνολο Ενεργητικού (Working Capital to Total Assets (WCTA)) συμμετέχει σε δέκα κανόνες και μάλιστα με την ενδιάμεση τιμή του. Ο αριθμοδείκτης αυτός συνδυάζεται με μεσαίες ή υψηλές τιμές του δείκτη πιστοληπτικής ικανότητας Quiscore. Μπορούμε να συμπεράνουμε ότι μεσαίου ύψους τιμές του δείκτη WCTA, συνδυασμένες με μεσαίο ή υψηλό σκορ πιστοληπτικής ικανότητας, συνιστούν τη μη έκδοση σχολίων.

Conditions	Consequence	Support	Confidence	Lift
Return_on_Total_Assets=<-33.14	Qualified	0.28	0.869	1.7379
Return_on_Total_Assets=<-33.14 AND Profit_L_before_ Taxation=<72.5	Qualified	0.28	0.869	1.7379
Return_on_Total_Assets=<-33.14 AND Retained_Profit_Loss=<43	Qualified	0.2778	0.8681	1.7361
Return_on_Total_Assets=<-33.14 AND Current_Liabilities_Trend=>-69.315	Qualified	0.2644	0.8623	1.7246
Profit_L_before_Taxation=<72,5 AND Shareholders_Funds=<2481.5	Qualified	0.2867	0.8323	1.6645
Retained_Profit_Loss=<43 AND Shareholders_Funds=<2481.5	Qualified	0.2844	0.8205	1.6410
Profit_L_before_Taxation=<72.5 AND Working_Capital=<216.5	Qualified	0.2756	0.7898	1.5795
Working_Capital=<216.5 AND Retained_Profit_Loss=<43	Qualified	0.2756	0.7750	1.5500
Profit_L_before_Taxation=<72.5 AND Return_on_Total_Assets=<-33.14 AND Retained_Profit_Loss=<43	Qualified	0.2778	0.8681	1.7361

Profit_L_before_Taxation=<72.5 AND Current_Liabilities_Trend=>-69.315 AND Return_on_Total_Assets=<-33.14	Qualified	0.2644	0.8623	1.7246
Current_Liabilities_Trend=>-69.315 AND Return_on_Total_Assets=<-33.14 AND Retained_Profit_Loss=<43	Qualified	0.2622	0.8613	1.7226
Profit_L_before_Taxation=<72.5 AND Retained_Profit_Loss=<43 AND Shareholders_Funds=<2481.5	Qualified	0.2822	0.8301	1.6601
Profit_L_before_Taxation=<72.5 AND Current_Liabilities_Trend=>-69.315 AND Shareholders_Funds=<2481,5	Qualified	0.2689	0.8231	1.6463
Current_Liabilities_Trend=>-69.315 AND Retained_Profit_Loss=<43 AND Shareholders_Funds=<2481.5	Qualified	0.2667	0.8108	1.6215
Profit_L_before_Taxation=<72.5 AND Working_Capital=<216.5 AND Retained_Profit_Loss=<43	Qualified	0.2733	0.7885	1.5769
Profit_L_before_Taxation=<72.5 AND Working_Capital=<216.5 AND Current_Liabilities_Trend=>-69.315	Qualified	0.2622	0.7815	1.5629
Working_Capital=<216,5 AND Current_Liabilities_Trend=>-69.315 AND Retained_Profit_Loss=<43	Qualified	0.2622	0.7662	1.5325

Πίνακας 8.2 *Κανόνες Συσχέτισης για τις εταιρείες που πήραν σχόλια*

Conditions	Consequence	Support	Confidence	Lift
Zscore=>0.4893796	Unqualified	0.3822	0.7713	1.5426
Zscore=>0.4893796 AND Current_Liabilities_Trend=>-69.315	Unqualified	0.3822	0.7748	1.5495
Zscore=>0.4893796 AND Current_Assets=>-48.87	Unqualified	0.3822	0.7748	1.5495
QuiScore=>26.5 AND WCTA=>-1.271049E-02 <0.3884639	Unqualified	0.4000	0.7692	1.5385
QuiScore=>26.5 AND WCTA=>-1.271049E-02 <0.3884639 AND AFTA=<7.724846E-03	Unqualified	0.3867	0.8018	1.6037
Total_Assets_Trend=>-26.59 AND QuiScore=>26.5 AND WCTA=>-1.271049E-02 <0.3884639	Unqualified	0.3800	0.7953	1.5907
Current_Assets=>-48.87 AND QuiScore=>26.5 AND WCTA=>-1.271049E-02 <0.3884639	Unqualified	0.3911	0.7892	1.5785
Current_Liabilities_Trend=>-69.315 AND QuiScore=>26.5 AND WCTA=>-1.271049E-02 <0.3884639	Unqualified	0.4000	0.7759	1.5517
Solvency_Ratio=>17.2 AND WCTA=>-1.271049E-02 <0.3884639 AND AFTA=<7.724846E-03	Unqualified	0.3978	0.7749	1.5498
Zscore=>0.4893796 AND Current_Assets=>-48.87 AND Current_Liabilities_Trend=>-69.315	Unqualified	0.3822	0.7748	1.5485

Solvency_Ratio=>17.2 AND QuiScore=>26.5 AND WCTA=>-1.271049E-02 <0.3884639	Unqualified	0.3911	0.7719	1.6539
Current_Assets=>-48.87 AND WCTA=>-1.271049E-02 <0.3884639 AND AFTA=<7.724846E-03	Unqualified	0.4067	0.7689	1.5378
Total_Assets_Trend=>-26.59 AND WCTA=>-1.271049E-02<0.3884639 AND AFTA=<7.724846E-03	Unqualified	0.3933	0.7662	1.5325
Solvency_Ratio=>17.2 AND Total_Assets_Trend=>-26.59 AND WCTA=>-1.271049E-02 <0.3884639	Unqualified	0.3867	0.7598	1.5197
Current_Assets=>-48.87 AND Total_Assets_Trend=>-26.59 AND WCTA=>-1.271049E-02 <0.3884639	Unqualified	0.4044	0.7552	1.5104

Πίνακας 8.3 *Κανόνες Συσχέτισης για τις εταιρείες που δεν πήραν σχόλια*

Βιβλιογραφία/Αναφορές

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the 1993 ACM SIGMOD I8th International Conference on Management of Data,* 207-216. NewYork, NY: ACM. doi: 10.1145/170035.170072

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the 20th International Conference on Very Large Databases,* 487-499. San Francisco, CA: Morgan Kaufmann.

Bell, T., & Tabor, R. (1991). Empirical Analysis of Audit Uncertainty Qualifications. *Journal of Accounting Research*, *29*(2), 350-370. doi: 10.2307/2491053

Brin, S., Motwani, R., Ullman, J., & Tsur, S. (1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. *ACM SIGMOD Record*, *26*(2), 255-264. doi: 10.1145/253262.253325

Boulicaut, J. F., & Jeudy, B. (2005). Constraint-Based Data Mining. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers* (pp. 399-416). New York, NY: Springer Science + Business Media Inc.

Calderon, T., & Cheh, J. (2002). A Roadmap for Future Neural Networks Rsearch in Auditing and Risk Assessment. *International Journal of Accounting Information Systems, 3*(4), 203-236. doi: 10.1016/s1467-0895(02)00068-4

De Castro, L., & Timmis, J. (2002). *Artificial Immune Systems: A new Computational Intelligence Approach*. London, UK: Springer-Verlag.

Gaganis, C., Pasiouras, F., & Doumpos, M. (2007). Probabilistic Neural Networks for the Identification of Qualified Audit Opinions. *Expert Systems with Applications, 32*(1), 114-124. doi: 10.1016/j.eswa.2005.11.003

Gkoulalas-Divanis, A., & Verykios, V. (2010). *Association Rules Hiding for Data Mining*. Dordrecht: Springer.

Goethals, B. (2005). Frequent Set Mining. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers* (pp. 377-397). New York, NY: Springer Science + Business Media Inc.

Grahne, G., Lakshmanan, L., & Wang, X. (2000). Efficient Mining of Constrained Correlated Sets. *Proceedings of the 16th International Conference on Data Engineering,* 512–521. San Diego, CA: IEEE. doi: 10.1109/ICDE.2000.839450

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques*. Waltham, MA: Morgan Kaufmann Publishers.

Hilderman, R., & Hamilton, H. (2001). *Knowledge Discovery and Measures of Interest*. Norwell, MA: Kluwer Academic Publishers.

Hoeppner, F. (2005). Association Rules. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers* (pp. 353-376). New York, NY: Springer Science + Business Media Inc.

Intan, R. (2006). A Proposal of Fuzzy Multidimensional Association Rules. *Informatica*, *7*, 85-90.

Kane, G., & Velury, U. (2004). The role of institutional ownership in the market for auditing services: an empirical investigation. *Journal of Business Research*, *57*(9), 976-983. doi: 10.1016/s0148-2963(02)00499-x

Khare, N., Adlakha, N., & Pardasani, K. R. (2010). An Algorithm for Mining Multidimensional Association Rules using Boolean Matrix. *Proceedings of the 2010 IEEE International Conference on Recent Trends in Information, Telecommunication and Computing,* 95-99. Kochi: IEEE. doi: 10.1109/ITC.2010.8

Kifer, D., Gehrke, J., Bucila, C., & White, W. (2003). How to quickly find a witness. *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,* 272– 283. Madison, WI: ACM. doi: 10.1145/773153.773180

Kirkos, E. (2010). Modeling the Auditors' Opinions by Using Association Rules. In L. I. Spendler (Ed.), *Data Mining and Management*. New York, NY: Nova Publishers.

Kirkos, E., Spathis, C., Nanopoulos, A., & Manolopoulos, Y. (2007). Identifying Qualified Auditors' Opinions: A Data Mining Approach. *Journal of Emerging Technologies in Accounting*, *4*(1), 183-197. doi: 10.2308/jeta.2007.4.1.183

Koh, Y. S., & Rountree, N. (2009). *Rare Association Rule Mining and Knowledge Discovery: Technologies*

for Infrequent and Critical Event Detection. Hershey, PA: Information Science Reference.
Koskivaara, E. (2004). Artificial Neural Networks in Analytical Review Procedures. *Managerial Auditing Journal, 19*(2), 191-223. doi: 10.1108/02686900410517821
Lenard, M., Alam, P., & Madey, G. (1995). The Application of Neural Networks and a Qualitative Response Model to the Auditor's Going Concern Uncertainty Decision. *Decision Sciences*, *26*(2), 209-227. doi: 10.1111/j.1540-5915.1995.tb01426.x
Lent, B., Swami, A., & Widom, J. (1997). Clustering Association Rules. *Proceedings of the 13th International Conference on Data Engineering,* 220-231. Birmingham, UK: IEEE. doi: 10.1109/ICDE.1997.581756
Ng, R., Lakshmanan, L., Han, J., & Pang, A. (1998). Exploratory Mining and Pruning Optimizations of Constrained Association Rules. *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data,* 13-24. Seattle, WA: ACM. doi: 10.1145/276304.276307
Pei, J., & Han, J. (2000). Can we Push More Constraints into Frequent Pattern Mining?. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 350–354. Boston, MA: ACM. doi: 10.1145/347090.347166
Pei, J., Han, J., & Lakshmanan, L. V. S. (2001). Mining Frequent Itemsets with Convertible Constraints. *Proceedings of the 17th International Conference on Data Engineering,* 433–442. Heidelberg: IEEE. doi: 10.1109/ICDE.2001.914856
Piatetsky-Shapiro, G. (1991). Discovery, Analysis and Presentation of Strong Rules. In G. Piatetsky-Shapiro & W. J. Frawley (Eds.), *Knowledge Discovery in Databases* (pp. 229-248). Menlo Park, CA: AAAI Press.
Reynolds, J., & Francis, J. (2001). Does Size Matter? The Influence of Large Clients on Office-Level Auditing reporting Decisions. *Journal of Accounting and Economics*, *30*(3), 375-400. doi: 10.1016/s0165-4101(01)00010-6
Smyth, P., & Goodman, R. (1992). An Information Theoretic Approach to Rule Induction from Databases. *IEEE Transactions on Knowledge and Data Engineering*, *4*(4), 301-316. doi: 10.1109/69.149926
Spathis, C. (2003). Audit Qualification, Firm Litigation and Financial information: An Empirical Analysis in Greece. *International Journal of Auditing*, *7*(1), 71-85. doi: 10.1111/1099-1123.00006
Taboada, K., Gonzales, E., Shimada, K., Mabu, S., Hirasawa, K., & Jinglu, H. (2007). Mining Association Rules from Databases with Continuous Attributes using Genetic Network Programming. *Proceedings of the IEEE Congress on Evolutionary Computation*, 1311-1317. Singapore: IEEE. doi: 10.1109/CEC.2007.4424622
Tan, P., & Kumar, V. (2002). Selecting the Right Interestingness Measure for Association Patterns. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 32-41. Edmonton: ACM. doi: 10.1145/775047.775053
Vannucci, M., & Colla, V. (2004). Meaningful Discretization of Continuous Features for Association Rules Mining by means of a SOM. *Proceedings of the European Symposium on Artificial Neural Networks*, 489-494.
Yang, G. (2013). A Novel Method for Mining Association Rules from Continuous Attributes Based on Cultural Immune Algorithm. *Journal of Information & Computational Science*, *10*(9), 2845-2853. doi: 10.12733/jics20102086
Zhang, C., & Zhang, S. (2002). *Association Rule Mining: Models and Algorithms*. Berlin: Springer-Verlag.

Κριτήρια Αξιολόγησης

Άσκηση Υπολογισμών 8.1

Στη βάση δεδομένων όπου τηρούνται στοιχεία πωλήσεων μιας επιχείρησης υπάρχουν οι καταχωρήσεις που παρουσιάζονται στον Πίνακα 8.4

TID	ΕΜΠΟΡΕΥΜΑΤΑ
101	Α,Β,Δ
102	Β,Γ
103	Α,Δ,Ε
104	Β,Δ
105	Α,Δ,Ε
106	Β,Δ
107	Α,Β
108	Α,Β,Δ
109	Β,Γ,Δ,

Πίνακας 8.4 *Δεδομένα Άσκησης 1*

Εφαρμόζοντας τον αλγόριθμο Apriori, βρείτε τα συχνά στοιχειοσύνολα με συχνότητα εμφάνισης ίση ή μεγαλύτερη από 3.

Λύση

Αρχικά γίνεται καταμέτρηση των εμφανίσεων του κάθε στοιχείου. Η συχνότητα εμφάνισης του κάθε στοιχείου είναι η ακόλουθη Α:5, Β:7, Γ:2, Δ:7, Ε:2. Τα στοιχεία Γ και Ε είναι μη συχνά, οπότε αποκλείονται.

Από τα συχνά 1-στοιχειοσύνολα (Α, Β, Δ) δημιουργούνται τα 2-στοιχειοσύνολα. Δημιουργούνται τα στοιχειοσύνολα ΑΒ, ΑΔ, ΒΔ.

Γίνεται καταμέτρηση της συχνότητας εμφάνισης των 2-στοιχιοσυνόλων. Οι συχνότητες εμφάνισης είναι ΑΒ:3, ΑΔ:4, ΒΔ:5. Και τα τρία στοιχειοσύνολα είναι συχνά.

Από τα συχνά 2-στοιχειοσύνολα δημιουργούνται τα 3-στοιχειοσύνολα. Δημιουργείται το στοιχειοσύνολο ΑΒΔ.

Το στοιχειοσύνολο ΑΒΔ εμφανίζεται μόνο 2 φορές και επομένως δεν είναι συχνό.

Άσκηση Υπολογισμών 8.2

Στη βάση δεδομένων με στοιχεία πωλήσεων μιας επιχείρησης υπάρχουν οι καταχωρήσεις που παρουσιάζονται στον Πίνακα 8.5.

TID	ΕΜΠΟΡΕΥΜΑΤΑ
101	Α,Β,Δ,Ε
102	Α,Β,Γ
103	Α,Β,Ε
104	Β,Δ
105	Α,Δ,Ε
106	Α,Β,Γ,Δ
107	Α,Β,Γ,Ε
108	Α,Β,Δ
109	Β,Γ,Δ,

Πίνακας 8.5 *Δεδομένα Άσκησης 2*

Να υπολογιστούν η υποστήριξη και η εμπιστοσύνη του κανόνα Α,Β→Γ.

Λύση

Γίνεται καταμέτρηση των ταυτόχρονων εμφανίσεων των εμπορευμάτων Α,Β και Γ. Τα εμπορεύματα εμφανίζονται σε τρεις, από τις εννέα συνολικά συναλλαγές. Η Υποστήριξη του κανόνα είναι 3/9 = 33%.

Γίνεται καταμέτρηση των ταυτόχρονων εμφανίσεων των εμπορευμάτων Α και Β. Τα εμπορεύματα εμφανίζονται σε έξι, από τις εννέα συνολικά συναλλαγές. Η Εμπιστοσύνη του κανόνα είναι 3/6 = 50%.

Άσκηση Υπολογισμών 8.3

Ο παρακάτω πίνακας δίνει τα στοιχεία πωλήσεων φρούτων και λαχανικών. Η ένδειξη «ΦΡΟΥΤΑ» σημαίνει πώληση φρούτων και η ένδειξη «ΟΧΙ ΦΡΟΥΤΑ» σημαίνει πώληση που δεν περιέχει φρούτα. Τα αντίστοιχα ισχύουν για τα λαχανικά. Συνολικά υπάρχουν 10000 πωλήσεις. Στις 5000 πωλήσεις περιλαμβάνονται φρούτα και από αυτές οι 4000 περιλαμβάνουν και λαχανικά. Στις 7000 πωλήσεις περιλαμβάνονται λαχανικά και από αυτές οι 4000 περιλαμβάνουν και φρούτα.

	ΦΡΟΥΤΑ	ΟΧΙ ΦΡΟΥΤΑ	ΣΥΝΟΛΟ
ΛΑΧΑΝΙΚΑ	4000	3000	7000
ΟΧΙ ΛΑΧΑΝΙΚΑ	1000	2000	3000
ΣΥΝΟΛΟ	5000	5000	10000

Να βρεθεί η υποστήριξη και η εμπιστοσύνη του κανόνα ΦΡΟΥΤΑ→ΛΑΧΑΝΙΚΑ.
Η πώληση φρούτων αυξάνει την πιθανότητα πώλησης λαχανικών;

Λύση

Ταυτόχρονη πώληση φρούτων και λαχανικών υπάρχει στις 4000 πωλήσεις. Η υποστήριξη του κανόνα *ΦΡΟΥΤΑ→ΛΑΧΑΝΙΚΑ* είναι 4000/10000=0,4.

Από τις 5000 πωλήσεις που περιλαμβάνουν φρούτα, οι 4000 περιλαμβάνουν και λαχανικά. Η εμπιστοσύνη του κανόνα είναι 4000/5000=0,8.

Για να βρούμε εάν η πώληση φρούτων αυξάνει την πιθανότητα πώλησης λαχανικών πρέπει να υπολογίσουμε το Lift του κανόνα. Το Lift υπολογίζεται από τη σχέση πιθανότητα(φρούτων+λαχανικών)/(πιθανότητα(φρούτων)*πιθανότητα(λαχανικών)) = 0,4/(0,5*0,7) = 1,14. Το Lift είναι μεγαλύτερο από 1, οπότε η πώληση φρούτων αυξάνει την πιθανότητα πώλησης λαχανικών και ο κανόνας είναι ισχυρός.

Άσκηση Εφαρμογής 8.4

Χρησιμοποιήστε το αρχείο «supermarket.arff». Το αρχείο αυτό μεταφέρεται στον υπολογιστή σας με την εγκατάσταση του WEKA και θα το βρείτε στον υποφάκελο «Data», ο οποίος βρίσκεται στον φάκελο του WEKA. Το αρχείο περιέχει δεδομένα ενός super market της Ν. Ζηλανδίας. Ειδικότερα, περιέχει 217 στήλες και 4.627 γραμμές. Κάθε στήλη αντιστοιχεί σε μια κατηγορία προϊόντος, πχ καφές, λαχανικά, άνθη κλπ. Το όνομα της στήλης είναι η κατηγορία προϊόντος. Στο όνομα ορισμένων στηλών, στη θέση της κατηγορίας προϊόντος, υπάρχει μια κωδική τιμή, η οποία αποτελείται από τη λέξη «department» και έναν αριθμό. Κάθε γραμμή του αρχείου αντιστοιχεί στις αγορές που πραγματοποίησε ένας πελάτης σε μια επίσκεψη του στο κατάστημα. Σε κάθε γραμμή υπάρχει η τιμή «t» στα προϊόντα που αγόρασε ο πελάτης και κενό στα υπόλοιπα προϊόντα. Αν για παράδειγμα ένας πελάτης αγόρασε μόνο καφέ και λαχανικά, τότε σε αυτές τις δύο στήλες θα υπάρχει η τιμή «t», ενώ στις υπόλοιπες στήλες θα υπάρχει κενό. Η τελευταία στήλη ονομάζεται «Total» και αναφέρεται στην αξία της εκάστοτε αγοράς. Οι δυνατές τιμές γι' αυτή τη στήλη είναι «high» και «low».
Χρησιμοποιήστε τον αλγόριθμο Apriori.
Ανακαλύψτε Κανόνες Συσχέτισης με τις προκαθορισμένες τιμές support και confidence του WEKA, προβάλλοντας ταυτόχρονα τα συχνά αντικειμενοσύνολα.
Ανακαλύψτε κανόνες που ισχύουν για τουλάχιστον το 50% των συνολικών πωλήσεων.
Ανακαλύψτε κανόνες όπου το αριστερό μέρος ενισχύει πάρα πολύ την πιθανότητα εμφάνισης του δεξιού μέρους, έχοντας Lift τουλάχιστον 2.
Ανακαλύψτε κανόνες όπου το αριστερό μέρος ενισχύει αρκετά την πιθανότητα εμφάνισης του δεξιού μέρους, έχοντας Lift τουλάχιστον 1.5.

Ανακαλύψτε κανόνες όπου το δεξιό μέλος τους είναι η αγορά κονσερβοποιημένων λαχανικών (canned vegetables).
Ανακαλύψτε κανόνες όπου το δεξιό μέλος τους είναι η αξία αγορών ώστε να συσχετίσετε την αγορά συγκεκριμένων προϊόντων με τη συνολική αξία αγορών.

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «supermarket.arff» πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» μελετήστε τα γνωρίσματα των δεδομένων που περιγράφουν τις κατηγορίες προϊόντων.

Μεταβείτε στο tab «Associate».

Στο πεδίο «Associator» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka\associations\Apriori.

Κάντε κλικ στα περιεχόμενα του πεδίου «Associator» και στο όνομα «Apriori». Ανοίγει το παράθυρο ρύθμισης παραμέτρων του Apriori. Παρατηρήστε ότι η προκαθορισμένη τιμή για την ελάχιστη υποστήριξη είναι 0.1, ότι ως μετρική χρησιμοποιείται η εμπιστοσύνη και ότι η ελάχιστη τιμή μετρικής είναι 0.9. Αυτά σημαίνουν ότι θα εξορυχτούν κανόνες με ελάχιστη υποστήριξη 0.1 και ελάχιστη εμπιστοσύνη 0.9. Στο πεδίο «outputItemSets» επιλέξτε «True», ώστε να εμφανιστούν στα αποτελέσματα τα συχνά αντικειμενοσύνολα. Κάντε κλικ στο κουμπί «OK».

Κάντε κλικ στο κουμπί «Start». Περιμένετε μέχρι να ολοκληρωθεί η εκτέλεση του αλγορίθμου. Κατά τη διάρκεια των υπολογισμών το μικρό πουλί στο κάτω δεξιά μέρος του παράθυρου κινείται.

Στο πεδίο «Associator output» εμφανίζονται τα αποτελέσματα. Εμφανίζονται τα συχνά στοιχειοσύνολα και οι δέκα καλύτεροι κανόνες. Το μέγιστο πλήθος των κανόνων που θα εμφανιστεί καθορίζεται στο παράθυρο παραμέτρων του Apriori. Από τους κανόνες μπορούν να εξαχθούν συμπεράσματα για τις καταναλωτικές συνήθειες των πελατών.

Βήμα 2. Ανοίξτε το παράθυρο ρύθμισης παραμέτρων του Apriori. Ορίστε ελάχιστη υποστήριξη ίση με 0.5 και επαναλάβετε το πείραμα. Θα διαπιστώσετε ότι δεν βρίσκεται κανένας κανόνας. Αυτό συμβαίνει διότι ο συνδυασμός τιμών υποστήριξης και εμπιστοσύνης είναι υπερβολικά μεγάλος και δεν υπάρχουν τέτοιοι κανόνες.

Ανοίξτε το παράθυρο ρύθμισης παραμέτρων του Apriori. Ορίστε ελάχιστη υποστήριξη ίση με 0.5 και ελάχιστη εμπιστοσύνη ίση με 0.7 (στο πεδίο MinMetric). Επαναλάβετε το πείραμα. Θα διαπιστώσετε ότι εξορύσσονται τρεις κανόνες. Οι κανόνες αυτοί ισχύουν τουλάχιστον για το 50% των συνολικών πωλήσεων. Ο πρώτος κανόνας λέει ότι το 80% των πελατών που αγοράζουν κρέμα γάλακτος αγοράζουν επίσης ψωμί και κέικ.

Βήμα 3. Για να ανακαλύψετε κανόνες όπου το αριστερό μέρος ενισχύει πάρα πολύ την πιθανότητα εμφάνισης του δεξιού μέρους, ανοίξτε το παράθυρο ρύθμισης παραμέτρων του Apriori και ορίστε ως μετρική το μέγεθος «Lift» (πεδίο metricType) και ως ελάχιστη τιμή μετρικής την τιμή 2 (πεδίο minMetric). Οι κανόνες αυτοί είναι πάρα πολύ ισχυροί και λογικά δεν μπορεί να ισχύουν για μεγάλο ποσοστό πωλήσεων. Για τον λόγο αυτό, πρέπει να μειώσετε την τιμή υποστήριξης. Ορίστε ελάχιστη τιμή υποστήριξης ίση με 0.12 και εκτελέστε τον αλγόριθμο. Οι κανόνες που θα προκύψουν είναι πολύ ισχυροί. Ο πρώτος κανόνας έχει lift=2.29 και εμπιστοσύνη=0.61, γεγονός που σημαίνει ότι το αριστερό μέρος του κανόνα περίπου τετραπλασιάζει την πιθανότητα εμφάνισης του δεξιού μέρους.

Βήμα 4. Για να ανακαλύψετε κανόνες όπου το αριστερό μέρος ενισχύει αρκετά την πιθανότητα εμφάνισης του δεξιού μέρους, ανοίξτε το παράθυρο ρύθμισης παραμέτρων του Apriori και ορίστε ως μετρική το μέγεθος «Lift» (πεδίο metricType) και ως ελάχιστη τιμή μετρικής την τιμή 1.5 (πεδίο minMetric). Οι κανόνες αυτοί δεν είναι τόσο σπάνιοι, οπότε μπορείτε να αυξήσετε την τιμή της υποστήριξης. Ορίστε ελάχιστη τιμή υποστήριξης ίση με 0.2 και εκτελέστε τον αλγόριθμο. Σύμφωνα με τον πρώτο κανόνα, το αριστερό μέρος του κανόνα πολλαπλασιάζει την πιθανότητα εμφάνισης του δεξιού μέρους κατά δυόμιση περίπου φορές.

Βήμα 5. Για να ανακαλύψετε κανόνες όπου το δεξιό μέλος τους είναι η αγορά κονσερβοποιημένων λαχανικών (canned vegetables), ανοίξτε το παράθυρο ρύθμισης παραμέτρων του Apriori, επιλέξτε την τιμή «True» στο πεδίο «car» και εισάγετε στο πεδίο «classIndex» την τιμή 21. Η τιμή αυτή δηλώνει ότι το δεξιό μέρος των κανόνων που θα εξορυχτούν θα αποτελείται από το εικοστό πρώτο πεδίο, το οποίο είναι τα κονσερβοποιημένα λαχανικά, όπως μπορείτε να διαπιστώσετε στο tab «Preprocess». Για την εξόρυξη κανόνων με προκαθορισμό του δεξιού μέρους, το WEKA επιτρέπει μόνο τη χρήση της εμπιστοσύνης ως μετρικής. Στο πεδίο «metricType» επιλέξτε την τιμή «Confidence». Οι κανόνες που αναζητούμε είναι πολύ συγκεκριμένοι και πιθανώς όχι πολύ συχνοί. Για τον λόγο αυτό, ορίζουμε σχετικά μικρές τιμές υποστήριξης και εμπιστοσύνης. Ορίστε ελάχιστη υποστήριξη 0.1 και ελάχιστη εμπιστοσύνη 0.5. Οι κανόνες που προκύπτουν περιγράφουν τη συσχέτιση της αγοράς κονσερβοποιημένων λαχανικών με την αγορά άλλων συγκεκριμένων προϊόντων.

Βήμα 6. Για να ανακαλύψετε κανόνες όπου το δεξιό μέλος τους είναι η αξία αγορών, ώστε να συσχετίσετε την αγορά συγκεκριμένων προϊόντων με τη συνολική αξία αγορών, ανοίξτε το παράθυρο ρύθμισης παραμέτρων του Apriori, επιλέξτε την τιμή «True» στο πεδίο «car» και εισάγετε στο πεδίο «classIndex» την τιμή 217. Η τιμή αυτή δηλώνει ότι το δεξιό μέρος των κανόνων που θα εξορυχτούν θα αποτελείται από το διακοσιοστό δέκατο έβδομο πεδίο, το οποίο είναι το συνολικό ποσό αγορών. Ορίστε ελάχιστη υποστήριξη 0.1 και ελάχιστη εμπιστοσύνη 0.8 και εκτελέστε τον αλγόριθμο. Προκύπτουν έξι κανόνες, οι οποίοι συσχετίζουν συγκεκριμένα προϊόντα με πωλήσεις υψηλής αξίας.

Επισημαίνεται ότι οι τιμές ρύθμισης των παραμέτρων προέκυψαν μετά από επαναλαμβανόμενες δοκιμές. Στους κανόνες συσχέτισης ο χρήστης πρέπει να πειραματίζεται με τις τιμές των παραμέτρων για να εξορύξει χρήσιμους κανόνες.

9 Κατηγοριοποίηση

Σύνοψη

Το ένατο Κεφάλαιο καλύπτει εν μέρει τη θεματική ενότητα της Κατηγοριοποίησης (Classification). Η κατηγοριοποίηση είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων, με μεγάλο αριθμό εφαρμογών στον χώρο των οικονομικών. Είναι εργασία επιβλεπόμενης μάθησης, που στόχο έχει την ανακάλυψη της σχέσης ανάμεσα σε ένα γνώρισμα στόχο με ονομαστικές τιμές και σε ένα σύνολο άλλων γνωρισμάτων. Μια άλλη εργασία επιβλεπόμενης μάθησης είναι η Παλινδρόμηση, η οποία όμως στοχεύει στην πρόβλεψη αριθμητικών τιμών. Στην κατηγοριοποίηση εφαρμόζεται ένας επαγωγικός αλγόριθμος και κατασκευάζεται ένα μοντέλο. Η διαδικασία της κατηγοριοποίησης περιλαμβάνει τρία στάδια. Στο πρώτο στάδιο ο αλγόριθμος επεξεργάζεται τα δεδομένα του συνόλου εκπαίδευσης και κατασκευάζει ένα μοντέλο. Στο δεύτερο στάδιο ελέγχεται η ικανότητα του μοντέλου να προβλέπει την κλάση άγνωστων παρατηρήσεων. Εάν η επίδοση του μοντέλου κριθεί ικανοποιητική, τότε ακολουθεί το τρίτο στάδιο, το οποίο συνίσταται στη χρήση του μοντέλου για τη διατύπωση προβλέψεων. Κατά την εκπαίδευση πρέπει να αποφευχθεί η υπερπροσαρμογή του μοντέλου, η απομνημόνευση δηλαδή του συγκεκριμένου συνόλου εκπαίδευσης. Αποτέλεσμα της υπερπροσαρμογής είναι η πτώση της επίδοσης έναντι άγνωστων παρατηρήσεων. Κριτήρια για την αξιολόγηση των μεθόδων κατηγοριοποίησης είναι η ακρίβεια πρόβλεψης, η ταχύτητα, η ερμηνευσιμότητα, η επεκτασιμότητα και η ανθεκτικότητα.
Στα πλαίσια του παρόντος κεφαλαίου γίνεται παρουσίαση τριών, πολύ γνωστών μεθόδων κατηγοριοποίησης. Οι μέθοδοι αυτές είναι τα Δένδρα Αποφάσεων, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron και τα Μπαϋεσιανά Δίκτυα. Τα Δένδρα Αποφάσεων βασίζονται στη διαδοχική διάσπαση του συνόλου δεδομένων σε υποσύνολα. Αναπαριστώνται με μια ανεστραμμένη δενδρική δομή, όπου κάθε κόμβος αντιπροσωπεύει έναν έλεγχο στα δεδομένα, κάθε κλάδος αντιπροσωπεύει ένα αποτέλεσμα του ελέγχου και κάθε φύλο αντιπροσωπεύει μια απόφαση κατηγοριοποίησης. Αφού κατασκευαστεί το μοντέλο, μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Μια νέα παρατήρηση κατηγοριοποιείται ακολουθώντας μια διαδρομή από τη ρίζα μέχρι ένα φύλο, σύμφωνα με τους ελέγχους των κόμβων. Έχουν προταθεί διάφοροι αλγόριθμοι για τη δημιουργία Δένδρων Αποφάσεων. Στο παρόν κεφάλαιο παρουσιάζονται αρχικά τα δένδρα τύπου ID3. Τα δένδρα ID3 χρησιμοποιούν ως κριτήριο για τον διαχωρισμό των παρατηρήσεων το Κέρδος Πληροφορίας, δηλαδή τη μείωση της στατιστικής εντροπίας. Τα δένδρα τύπου C4.5 αποτελούν επέκταση – βελτίωση των ID3, χρησιμοποιούν ως κριτήριο διαχωρισμού τον Λόγο Κέρδους και είναι ικανά να χειρίζονται αριθμητικές μεταβλητές εισόδου. Τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron είναι ένα πλέγμα συνδεδεμένων νευρώνων. Κάθε σύνδεση συνοδεύεται από μία αριθμητική τιμή που ονομάζεται βάρος. Ένας νευρώνας μετασχηματίζει το σήμα εισόδου και το μεταβιβάζει σε επόμενους νευρώνες. Οι νευρώνες είναι οργανωμένοι σε επίπεδα, και υπάρχουν ένα επίπεδο εισόδου, ένα επίπεδο εξόδου και ένα ή περισσότερα κρυφά επίπεδα. Ο χρήστης προκαθορίζει τη δομή του δικτύου. Στη συνέχεια ακολουθεί η εκπαίδευση του δικτύου, η οποία συνίσταται στη ρύθμιση των βαρών των συνδέσεων. Ένας πολύ επιτυχημένος αλγόριθμος για την εκπαίδευση του δικτύου είναι η Αντίστροφη Μετάδοση Σφάλματος (Backpropagation). Τα Μπαϋεσιανά Δίκτυα αποτελούνται από έναν κατευθυνόμενο ακυκλικό γράφο και έναν πίνακα κατανομής πιθανοτήτων. Κάθε κόμβος του γράφου συμβολίζει μια στοχαστική μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης ανάμεσα σε δύο μεταβλητές. Οι Αφελείς Μπαϋεσιανοί κατηγοριοποιητές υποθέτουν την υπό συνθήκη ανεξαρτησία των μεταβλητών εισόδου. Αντιθέτως, τα Μπαϋεσιανά Δίκτυα επιτρέπουν την ανεξαρτησία υποσυνόλων των μεταβλητών εισόδου. Η εκπαίδευση ενός Μπαϋεσιανού Δικτύου περιλαμβάνει τον σχεδιασμό του γράφου και τον υπολογισμό του πίνακα πιθανοτήτων. Στο τέλος του κεφαλαίου παρουσιάζεται μια μελέτη περίπτωσης, όπου οι τρεις προαναφερθείσες τεχνικές κατηγοριοποίησης εφαρμόζονται για τον εντοπισμό περιπτώσεων παραποίησης των χρηματοοικονομικών καταστάσεων επιχειρήσεων. Τα τρία μοντέλα επιτυγχάνουν υψηλό βαθμό ακρίβειας έναντι άγνωστων παρατηρήσεων και αποκαλύπτουν σημαντικούς παράγοντες που σχετίζονται με τις περιπτώσεις διοικητικής απάτης

Προηγούμενη γνώση

Η θεματική ενότητα του παρόντος κεφαλαίου είναι αυτόνομη και δεν απαιτούνται ιδιαίτερες προηγούμενες γνώσεις. Ωστόσο, για την καλύτερη κατανόηση των περιεχομένων θα συνιστούσαμε την προηγούμενη ανάγνωση του Κεφαλαίου 6, το οποίο εισάγει τον αναγνώστη στην Εξόρυξη Δεδομένων και την ανάγνωση του Κεφαλαίου 7, το οποίο αναφέρεται στην προεπεξεργασία των δεδομένων. Για τον αναγνώστη που ενδιαφέρεται να αναζητήσει περισσότερες πληροφορίες για την Κατηγοριοποίηση και τις τρεις συγκεκριμένες μεθόδους που παρουσιάζονται,

υπάρχει πληθώρα διαθέσιμων συγγραμμάτων. Ενδεικτικά αναφέρουμε τα βιβλία των Han, Kamber and Pei (2011) και των Maimon and Rokach (2010).

9.1 Εισαγωγή

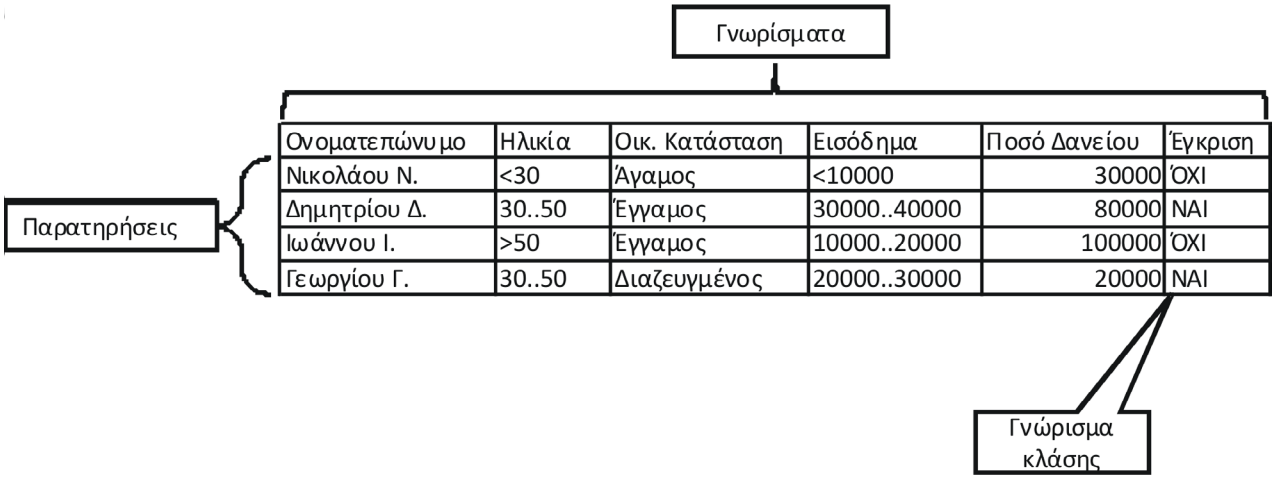
Η κατηγοριοποίηση *(*classification) είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων, με μεγάλο αριθμό εφαρμογών στον χώρο των οικονομικών. Η πρόβλεψη χρεοκοπίας, η έγκριση δανείων, η αναγνώριση απάτης είναι τυπικά προβλήματα κατηγοριοποίησης. Η κατηγοριοποίηση είναι εργασία **επιβλεπόμενης μάθησης**. Στόχος της επιβλεπόμενης μάθησης είναι η ανακάλυψη της σχέσης ανάμεσα σε ένα γνώρισμα στόχο και σε ένα σύνολο άλλων γνωρισμάτων. Το γνώρισμα στόχος αναφέρεται και ως εξαρτημένη μεταβλητή, ενώ τα υπόλοιπα γνωρίσματα αναφέρονται και ως ανεξάρτητες μεταβλητές. Με την επιβλεπόμενη μάθηση επιτυγχάνεται η δημιουργία ενός μηχανισμού λήψης αποφάσεων ή υπολογισμών, ο οποίος είναι ικανός να προβλέπει τις τιμές της εξαρτημένης μεταβλητής χρησιμοποιώντας τις ανεξάρτητες μεταβλητές. Ο μηχανισμός λήψης απόφασης καλείται και μοντέλο και μπορεί να έχει διάφορες μορφές, όπως πχ να είναι ένα σύνολο κανόνων ή μια εξίσωση ή το πλέγμα των νευρώνων και συνδέσεων ενός Νευρωνικού Δικτύου.

Στην επιβλεπόμενη μάθηση ανήκουν η **Κατηγοριοποίηση** (Classification) και η **Παλινδρόμηση** (Regression). Η κατηγοριοποίηση και η παλινδρόμηση έχουν πολλές ομοιότητες. Και στις δύο περιπτώσεις στόχος είναι η πρόβλεψη των τιμών ενός γνωρίσματος, με χρήση άλλων γνωρισμάτων. Επίσης, και στις δύο περιπτώσεις χρησιμοποιείται ένα σύνολο δεδομένων εκπαίδευσης, με την επεξεργασία του οποίου κατασκευάζεται το μοντέλο. Η διαφορά ανάμεσα στην κατηγοριοποίηση και στην παλινδρόμηση έχει σχέση με τον τύπο της εξαρτημένης μεταβλητής. Στόχος της παλινδρόμησης είναι η πρόβλεψη μιας εξαρτημένης μεταβλητής, η οποία περιέχει συνεχόμενες (αριθμητικές) τιμές. Αντιθέτως, κατηγοριοποίηση είναι η πρόβλεψη διακριτών ονομαστικών τιμών. Οι τιμές αυτές είναι συγκεκριμένες, γνωστές εκ των προτέρων και ορίζουν την κλάση (κατηγορία) στην οποία ανήκει κάθε αντικείμενο. Για τον λόγο αυτό, η εξαρτημένη μεταβλητή σε προβλήματα κατηγοριοποίησης καλείται και **γνώρισμα κλάσης**.

Με την ένταξη αντικειμένων σε ομάδες ασχολείται και μια άλλη εργασία Εξόρυξης Δεδομένων, η **Ανάλυση Συστάδων** (Clustering). Οι διαφορές ανάμεσα στην Ανάλυση Συστάδων και στην Κατηγοριοποίηση είναι μεγάλες. Η Ανάλυση Συστάδων επιμερίζει τα αντικείμενα σε ομάδες βάσει της ομοιότητας τους. Οι συστάδες και το πλήθος τους δεν είναι εκ των προτέρων γνωστές. Επίσης, δεν υπάρχει στα δεδομένα κάποιο πεδίο που να καθορίζει την ομάδα στην οποία ανήκει το κάθε αντικείμενο. Αντιθέτως, στην κατηγοριοποίηση οι κατηγορίες είναι εκ των προτέρων γνωστές. Οι τιμές του γνωρίσματος κλάσης ορίζουν την κατηγορία στην οποία ανήκει κάθε αντικείμενο.

Ένα παράδειγμα προβλήματος κατηγοριοποίησης είναι η έγκριση των τραπεζικών δανείων. Το σύνολο δεδομένων περιλαμβάνει στοιχεία για τον υποψήφιο δανειολήπτη, στοιχεία σχετικά με το δάνειο, καθώς επίσης και την τελική απόφαση για την έγκριση ή την απόρριψη του δανείου. Κάθε γραμμή του συνόλου δεδομένων αντιστοιχεί σε μια αίτηση. Οι γραμμές καλούνται και αντικείμενα, παραδείγματα ή παρατηρήσεις. Οι στήλες αναφέρονται σε μια ιδιότητα των αντικειμένων, όπως πχ το επάγγελμα του δανειολήπτη ή το είδος του δανείου (στεγαστικό, καταναλωτικό κλπ.). Οι στήλες καλούνται και πεδία (fields), μεταβλητές (variables), γνωρίσματα (attributes) ή χαρακτηριστικά (features). Το γνώρισμα το οποίο περιέχει την απόφαση της έγκρισης ή απόρριψης του δανείου είναι το γνώρισμα της κλάσης. Η έγκριση του δανείου εξαρτάται από τα στοιχεία της αίτησης, όπως η ηλικία, το επάγγελμα και η οικονομική κατάσταση του δανειολήπτη, το ποσό και ο τύπος του δανείου κλπ. Η δημιουργία ενός μοντέλου, το οποίο θα μπορεί να προβλέπει την έγκριση ή απόρριψη του δανείου χρησιμοποιώντας τα υπόλοιπα στοιχεία της αίτησης, είναι ένα πρόβλημα κατηγοριοποίησης. Στο Σχήμα 9.1 παρουσιάζεται το σύνολο δεδομένων αυτού του παραδείγματος

Ένας Επαγωγικός Αλγόριθμος είναι μια οντότητα, η οποία επεξεργάζεται ένα σύνολο δεδομένων και κατασκευάζει ένα μοντέλο. Το μοντέλο είναι μια τυποποίηση, η οποία περιγράφει τη γενίκευση της σχέσης ανάμεσα σε μια εξαρτημένη μεταβλητή και σε ένα σύνολο ανεξάρτητων μεταβλητών. Με άλλα λόγια, το μοντέλο μπορεί και δέχεται ως είσοδο τις τιμές των ανεξάρτητων μεταβλητών και παράγει ως έξοδο μια τιμή για την εξαρτημένη μεταβλητή. Το μοντέλο, αφού κατασκευαστεί, μπορεί να χρησιμοποιηθεί για την πρόβλεψη της κλάσης νέων παρατηρήσεων.



Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Νικολάου Ν.	<30	Άγαμος	<10000	30000	ΟΧΙ
Δημητρίου Δ.	30..50	Έγγαμος	30000..40000	80000	ΝΑΙ
Ιωάννου Ι.	>50	Έγγαμος	10000..20000	100000	ΟΧΙ
Γεωργίου Γ.	30..50	Διαζευγμένος	20000..30000	20000	ΝΑΙ

Σχήμα 9.1 *Έγκριση τραπεζικών δανείων*

9.2 Επαγωγικοί Αλγόριθμοι και Μοντέλα

Επαγωγικοί αλγόριθμοι ή μέθοδοι κατηγοριοποίησης υπάρχουν πολλές, όπως πχ τα Δένδρα Αποφάσεων, τα Μπαϋεσιανά Δίκτυα, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron κλπ. Τα μοντέλα τα οποία κατασκευάζει η κάθε μέθοδος είναι τελείως διαφορετικά. Για παράδειγμα, ένα μοντέλο Δένδρου Αποφάσεων είναι μια δενδρική δομή, όπου κάθε κόμβος είναι ένας έλεγχος σε κάποιο γνώρισμα, κάθε κλάδος είναι ένα αποτέλεσμα του ελέγχου και κάθε φύλο είναι μια απόφαση κατηγοριοποίησης. Ένα μοντέλο Μπαϋεσιανού Δικτύου είναι ένας κατευθυνόμενος ακυκλικός γράφος και μια κατανομή πιθανοτήτων συσχέτισης μεταξύ των μεταβλητών. Στον γράφο κάθε κόμβος αντιστοιχεί σε μια μεταβλητή.

Όπως αναφέρθηκε και προηγουμένως, ένας επαγωγικός αλγόριθμος επεξεργάζεται ένα σύνολο δεδομένων και παράγει ένα μοντέλο. Αν συμβολίσουμε με το *I* ένα επαγωγικό αλγόριθμο και με *D* ένα σύνολο δεδομένων, τότε το *I(D)* συμβολίζει το μοντέλο που θα παραχθεί από την επεξεργασία του *D* από τον *I*. Αν ο ίδιος αλγόριθμος επεξεργαστεί ένα διαφορετικό σύνολο δεδομένων *D'*, τότε θα παραχθεί ένα διαφορετικό μοντέλο *I(D')*. Εάν το *D'* περιέχει διαφορετικά πεδία από το *D,* είναι προφανές ότι τα δύο μοντέλα θα είναι διαφορετικά. Αν τα δύο σύνολα δεδομένων περιέχουν τα ίδια πεδία, αλλά διαφορετικές παρατηρήσεις, τότε και πάλι θα προκύψουν δύο διαφορετικά μοντέλα. Υπάρχουν μάλιστα μέθοδοι, όπως τα Δένδρα Αποφάσεων, όπου μικρές διαφορές στα σύνολα δεδομένων έχουν σαν αποτέλεσμα τη δημιουργία σημαντικά διαφορετικών μοντέλων.

Μια μέθοδος κατηγοριοποίησης μπορεί να είναι αιτιοκρατική (deterministic) ή στοχαστική (stochastic). Οι στοχαστικές μέθοδοι καλούνται και πιθανολογικές (probabilistic). Μια αιτιοκρατική μέθοδος δημιουργεί μοντέλα, τα οποία εκχωρούν μια παρατήρηση σε μια κλάση. Τα στοχαστικά μοντέλα υπολογίζουν την πιθανότητα να ανήκει η παρατήρηση σε κάθε μια από τις δυνατές κλάσεις. Παράδειγμα αιτιοκρατικής μεθόδου είναι τα Δένδρα αποφάσεων τύπου C4.5, ενώ παράδειγμα στοχαστικής μεθόδου είναι τα Μπαϋεσιανά Δίκτυα.

9.3 Στάδια κατηγοριοποίησης

Η κατηγοριοποίηση περιλαμβάνει τρία στάδια, το στάδιο της επιβλεπόμενης μάθησης, το στάδιο της επικύρωσης του μοντέλου και το στάδιο της χρήσης του μοντέλου. Αναλυτικότερα, οι εργασίες που λαμβάνουν χώρα σε κάθε στάδιο είναι οι ακόλουθες:

- **Επιβλεπόμενη μάθηση**. Στο στάδιο αυτό, μια μέθοδος κατηγοριοποίησης αναλύει ένα σύνολο δεδομένων. Η μέθοδος θα ανακαλύψει σχέσεις μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Το αποτέλεσμα αυτής της επεξεργασίας είναι η κατασκευή ενός μοντέλου. Η κατασκευή ή εκπαίδευση του μοντέλου καθοδηγείται από τις τιμές του γνωρίσματος της κλάσης και για τον λόγο αυτό η διαδικασία ονομάζεται επιβλεπόμενη μάθηση. Το σύνολο δεδομένων, το οποίο χρησιμοποιείται για την εκπαίδευση του μοντέλου, ονομάζεται σύνολο εκπαίδευσης (training data set). Η επιλογή του συνόλου εκπαίδευσης είναι καθοριστικής σημασίας, γιατί το μοντέλο που θα προκύψει θα αποτυπώνει σχέσεις που υπάρχουν στο σύνολο εκπαίδευσης. Μεροληπτικά **σύνολα**

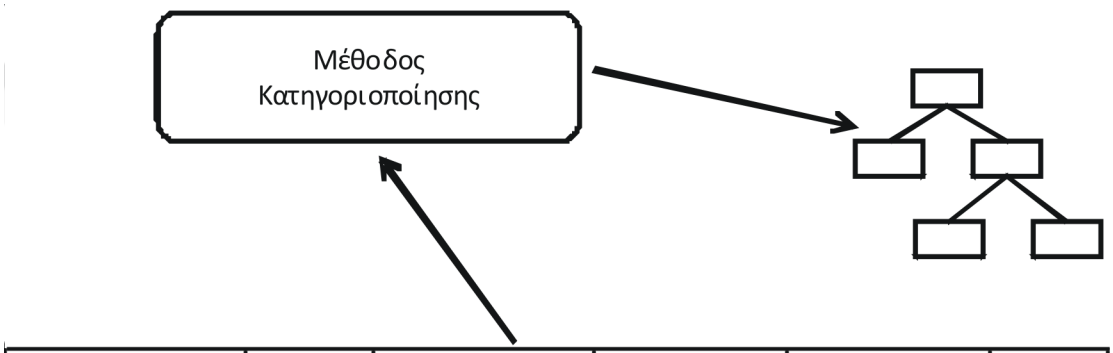
εκπαίδευσης θα οδηγήσουν στην κατασκευή μεροληπτικών μοντέλων.

- **Επικύρωση μοντέλου.** Στο στάδιο αυτό δοκιμάζεται η ακρίβεια του μοντέλου, η ικανότητα του δηλαδή να προβλέπει σωστά την κλάση των παρατηρήσεων. Το μοντέλο τροφοδοτείται με παρατηρήσεις, των οποίων η κλάση είναι γνωστή. Αναλύοντας τα στοιχεία των ανεξάρτητων μεταβλητών κάθε παρατήρησης, το μοντέλο προβλέπει την κλάση της παρατήρησης και στη συνέχεια συγκρίνεται η πρόβλεψη του μοντέλου με την πραγματική τιμή της κλάσης. Αν το μοντέλο επιδείξει ικανοποιητική ακρίβεια προβλέψεων, εάν δηλαδή προβλέψει σωστά την κλάση ενός ικανοποιητικού ποσοστού παρατηρήσεων, τότε θεωρείται επιτυχημένο και μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Η διαδικασία δοκιμής του μοντέλου καλείται **επικύρωση** (validation) και το σύνολο δεδομένων που χρησιμοποιείται για τη δοκιμή καλείται **σύνολο επικύρωσης** (validation set). Σκοπός ενός μοντέλου είναι να χρησιμοποιηθεί για τη διατύπωση προβλέψεων στην πραγματική ζωή και όχι να αναλύσει ένα συγκεκριμένο σύνολο δεδομένων. Το μοντέλο πρέπει να αποδείξει την ικανότητα του να προβλέπει την κλάση άγνωστων παρατηρήσεων, παρατηρήσεων δηλαδή διαφορετικών από αυτές που χρησιμοποιήθηκαν για την εκπαίδευση του. Για τον λόγο αυτό, το σύνολο εκπαίδευσης και το σύνολο επικύρωσης πρέπει να περιέχουν διαφορετικές παρατηρήσεις.
- **Χρήση του μοντέλο**υ. Το μοντέλο, αφού εκπαιδευτεί και επικυρωθεί, χρησιμοποιείται για τη διατύπωση προβλέψεων. Μια νέα παρατήρηση, της οποίας η κλάση είναι άγνωστη, εισάγεται στο μοντέλο. Το μοντέλο χρησιμοποιώντας τις τιμές των ανεξάρτητων μεταβλητών υπολογίζει την τιμή της κλάσης.

Το Σχήμα 9.2 παρουσιάζει τα στάδια της κατηγοριοποίησης με τη βοήθεια ενός παραδείγματος. Στο τμήμα Α) απεικονίζεται η επιβλεπόμενη μάθηση. Μια μέθοδος κατηγοριοποίησης επεξεργάζεται ένα σύνολο εκπαίδευσης, το οποίο περιέχει στοιχεία δανείων και κατασκευάζεται ένα μοντέλο. Το μοντέλο μπορεί να προβλέψει την έγκριση ή απόρριψη του δανείου από τα υπόλοιπα στοιχεία της εκάστοτε αίτησης. Στο τμήμα Β) απεικονίζεται η επικύρωση του μοντέλου. Το μοντέλο τροφοδοτείται με περιπτώσεις δανείων διαφορετικές από αυτές που χρησιμοποιήθηκαν για την εκπαίδευση. Για κάθε δάνειο, το μοντέλο πραγματοποιεί μια πρόβλεψη και η πρόβλεψη αυτή συγκρίνεται με την πραγματική απόφαση έγκρισης ή απόρριψης του δανείου. Υπολογίζεται η ακρίβεια του μοντέλου. Στο τμήμα Γ) το μοντέλο χρησιμοποιείται για την πρόβλεψη έγκρισης νέων δανείων.

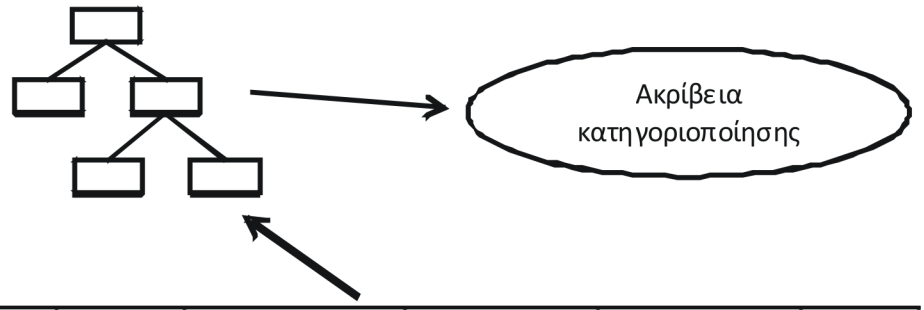
9.4 Υπερπροσαρμογή μοντέλων

Αναφέρθηκε προηγουμένως ότι για την εκτίμηση της ακρίβειας του μοντέλου πρέπει να χρησιμοποιηθούν παρατηρήσεις διαφορετικές από αυτές που χρησιμοποιήθηκαν για την εκπαίδευση. «Γιατί το σύνολο εκπαίδευσης δεν είναι αρκετά καλό για τον έλεγχο της ακρίβειας;», θα μπορούσε να αναρωτηθεί κανείς. Η απάντηση έχει σχέση με το πρόβλημα της υπερπροσαρμογής των μοντέλων στο σύνολο δεδομένων εκπαίδευσης. Με τον όρο **υπερπροσαρμογή** στα δεδομένα εκπαίδευσης (data overfitting) ορίζουμε το φαινόμενο όπου το μοντέλο «απομνημονεύει» τις περιπτώσεις οι οποίες υπάρχουν στο σύνολο εκπαίδευσης, αντί να εκπαιδεύεται ουσιαστικά, ενσωματώνοντας «κανόνες» γενικότερης ισχύος. Ένα υπερβολικά προσαρμοσμένο μοντέλο ενσωματώνει και τον θόρυβο των δεδομένων. Ακόμα όμως και όταν δεν υπάρχει θόρυβος, η υπερβολική προσαρμογή του μοντέλου στα συγκεκριμένα δεδομένα θα το εμποδίσει να προβλέψει σωστά την κλάση νέων παρατηρήσεων. Η υπερπροσαρμογή παρουσιάζεται όταν ένα μοντέλο είναι υπερβολικά περίπλοκο. Το μοντέλο αυτό είναι ικανό να αφομοιώσει τις ιδιαιτερότητες των δεδομένων εκπαίδευσης, αντί να καταγράψει σχέσεις γενικότερης ισχύος. Στο Σχήμα 9.3 στο τμήμα Α) απεικονίζεται ένα σύνολο δεδομένων εκπαίδευσης με δύο μεταβλητές. Επίσης παρουσιάζονται δύο μοντέλα, το ένα από τα οποία συμβολίζεται με τη συνεχόμενη γραμμή, ενώ το δεύτερο με τη διακεκομμένη γραμμή. Το πρώτο μοντέλο επιτυγχάνει ικανοποιητικό βαθμό γενίκευσης, ενώ το δεύτερο είναι πολύ σύνθετο και υπερπροσαρμοσμένο στα δεδομένα.



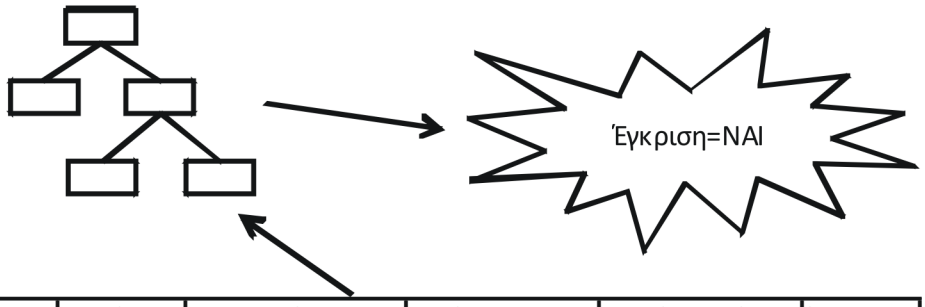
Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Νικολάου Ν.	<30	Άγαμος	<10000	30000	ΟΧΙ
Δημητρίου Δ.	30..50	Έγγαμος	30000..40000	80000	ΝΑΙ
Ιωάννου Ι.	>50	Έγγαμος	10000..20000	100000	ΟΧΙ
Γεωργίου Γ.	30..50	Διαζευγμένος	20000..30000	20000	ΝΑΙ

Α)



Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Κωνσταντίνου Κ.	<30	Άγαμος	<10000	40000	ΟΧΙ
Παναγιώτου Π.	30..50	Έγγαμος	30000..40000	100000	ΝΑΙ

Β)

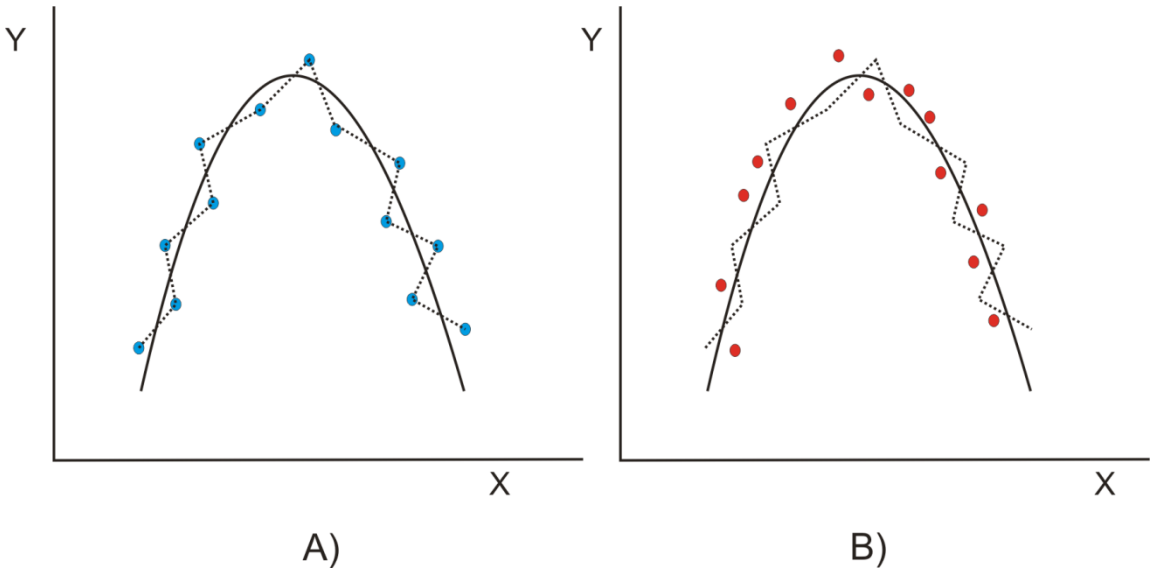


Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Χρήστου Χ.	30..50	Έγγαμος	30000..40000	90000	

Γ)

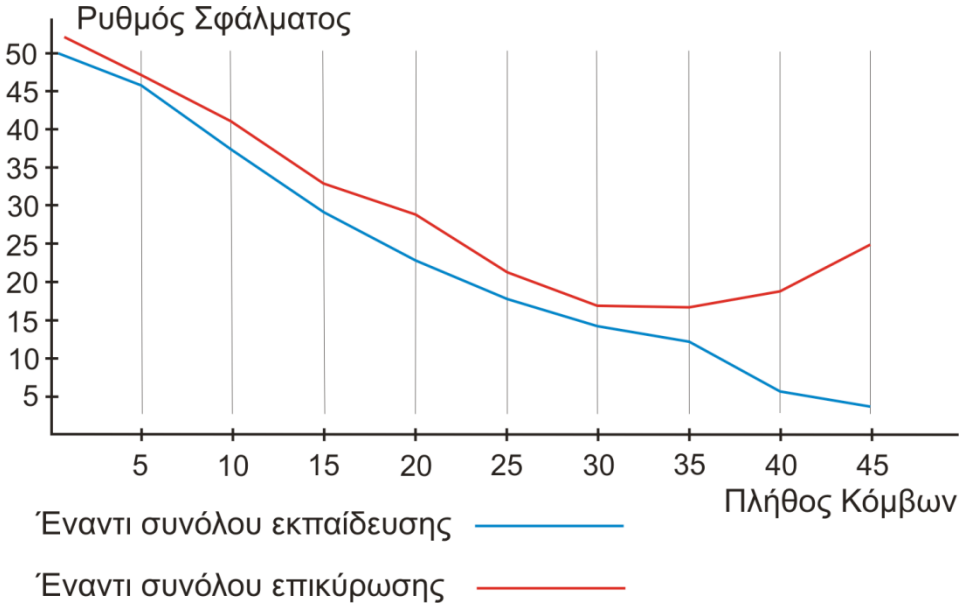
Σχήμα 9.2 *Στάδια Κατηγοριοποίησης*

Ένα υπερπροσαρμοσμένο μοντέλο επιτυγχάνει εξαιρετικά υψηλές επιδόσεις έναντι των δεδομένων εκπαίδευσης, οι επιδόσεις του όμως έναντι άγνωστων παρατηρήσεων δεν είναι ικανοποιητικές. Για τον λόγο αυτό, εξαιρετικά υψηλός ρυθμός ακρίβειας έναντι του συνόλου εκπαίδευσης, όχι μόνον δεν είναι ασφαλές μέτρο της επιτυχίας του μοντέλου, αλλά αποτελεί ένδειξη πιθανής υπερπροσαρμογής του. Στο Σχήμα 9.3 παρουσιάζονται τα δύο μοντέλα και ένα σύνολο νέων παρατηρήσεων, οι οποίες συμβολίζονται με κόκκινες τελείες. Είναι εμφανές ότι το γενικευμένο μοντέλο που συμβολίζεται με τη συνεχόμενη γραμμή προβλέπει καλύτερα τις τιμές του Y από τις τιμές του X.



Σχήμα 9.3 *Υπερπροσαρμογή κατηγοριοποιητή*

Αντίστροφο πρόβλημα της υπερπροσαρμογής είναι η υποπροσαρμογή. Στην περίπτωση της **υποπροσαρμογής**, το μοντέλο είναι υπερβολικά απλό για να ενσωματώσει τις ουσιαστικές σχέσεις, οι οποίες υπάρχουν στα δεδομένα εκπαίδευσης. Αποτέλεσμα της υποπροσαρμογής είναι η χαμηλή ακρίβεια έναντι και των δεδομένων εκπαίδευσης και των άγνωστων παρατηρήσεων. Στο Σχήμα 9.4 παρουσιάζεται η πτώση του ρυθμού σφάλματος σε σχέση με την πολυπλοκότητα του μοντέλου, η οποία εκφράζεται ως πλήθος κόμβων ενός Δένδρου Αποφάσεων. Αρχικά το μοντέλο είναι εξαιρετικά απλό και ο ρυθμός σφάλματος είναι περίπου 50%. Ο ρυθμός σφάλματος μειώνεται με την αύξηση της πολυπλοκότητας, μέχρι τους 30 κόμβους. Πέραν των 30 κόμβων, ο ρυθμός σφάλματος έναντι του συνόλου εκπαίδευσης εξακολουθεί να μειώνεται. Αντιθέτως, ο ρυθμός σφάλματος έναντι του συνόλου επικύρωσης, το οποίο αποτελείται από άγνωστες παρατηρήσεις, αρχικά διατηρείται σταθερός μέχρι τους 35 κόμβους και στη συνέχεια αυξάνεται.



Σχήμα 9.4 *Πτώση Ρυθμού Σφάλματος και πολυπλοκότητα μοντέλου*

9.5 Κριτήρια αξιολόγησης μεθόδων κατηγοριοποίησης

Η έρευνα σχετικά με την κατηγοριοποίηση έχει αποδώσει πλούσιους καρπούς και σήμερα υπάρχουν διαθέσιμες αρκετές και πολύ διαφορετικές μέθοδοι κατηγοριοποίησης. Ορισμένες από αυτές, όπως πχ τα Νευρωνικά

Δίκτυα, θεωρούνται ιδιαίτερα ικανές να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Οι μέθοδοι αυτές μπορούν να θεωρηθούν «καλύτερες» από άλλες, όμως η ακρίβεια δεν είναι το μοναδικό κριτήριο αξιολόγησης των μεθόδων κατηγοριοποίησης. Αναλυτικότερα, οι μέθοδοι κατηγοριοποίησης μπορούν να αξιολογηθούν με βάση τα παρακάτω κριτήρια:

- **Ακρίβεια πρόβλεψης** (accuracy). Είναι η ικανότητα των μοντέλων να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Προφανώς πρόκειται για ένα πολύ σημαντικό κριτήριο και μεγάλο μέρος της έρευνας προσανατολίζεται στην ανακάλυψη μεθόδων υψηλών επιδόσεων.
- **Ταχύτητα** (speed). Σχετίζεται με την πολυπλοκότητα της μεθόδου και το υπολογιστικό κόστος που αυτή συνεπάγεται. Η εκτέλεση περίπλοκων αλγορίθμων, οι οποίοι απαιτούν εκτεταμένους υπολογισμούς, προκαλούν καθυστερήσεις. Καθυστερήσεις μπορεί να υπάρχουν στη διαδικασία κατασκευής, αλλά και στη χρήση των μοντέλων, στην εφαρμογή τους δηλαδή για την κατηγοριοποίηση μιας νέας παρατήρησης. Ορισμένες μέθοδοι, όπως τα Δένδρα Αποφάσεων, διαθέτουν γρήγορους αλγορίθμους και ο χρόνος κατασκευής των μοντέλων είναι μικρός. Άλλες μέθοδοι, όπως τα Νευρωνικά Δίκτυα, χρειάζονται πολύ περισσότερο χρόνο για την εκπαίδευση των μοντέλων. Κατά κανόνα ο χρόνος χρήσης των μοντέλων είναι πολύ μικρός. Ωστόσο, υπάρχουν μέθοδοι, όπως οι k-Πλησιέστεροι Γείτονες, οι οποίες δεν εκπαιδεύουν κάποιο μοντέλο, όμως ο χρόνος για την κατηγοριοποίηση νέων παρατηρήσεων είναι μεγάλος.
- **Ερμηνευσιμότητα** (interpretability). Είναι η ικανότητα της μεθόδου να παράγει μοντέλα, τα οποία είναι κατανοητά από τον άνθρωπο. Για παράδειγμα, στα Δένδρα Αποφάσεων ο τρόπος λήψης της απόφασης κατηγοριοποίησης είναι απολύτως κατανοητός και το μοντέλο μπορεί εύκολα νε μετατραπεί σε ένα σύνολο κανόνων της μορφής ΕΑΝ-ΤΟΤΕ. Αντιθέτως, τα μοντέλα άλλων μεθόδων, όπως τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης, λειτουργούν ως «μαύρα κουτιά». Στα μοντέλα αυτά παρέχονται οι τιμές των μεταβλητών εισόδου και υπολογίζεται η απόφαση κατηγοριοποίησης στην έξοδο. Ο τρόπος λήψης της απόφασης όμως δεν είναι κατανοητός στον άνθρωπο. Η ερμηνευσιμότητα είναι μια σημαντική ιδιότητα των μεθόδων κατηγοριοποίησης. Σε πολλές περιπτώσεις οι χρήστες των μοντέλων επιθυμούν να γνωρίζουν τον τρόπο λήψης της απόφασης, ώστε να είναι πιο σίγουροι για το αποτέλεσμα. Επίσης, στο μοντέλο καταγράφονται σχέσεις μεταξύ των δεδομένων. Ορισμένες από τις σχέσεις αυτές μπορεί να είναι νέες και άγνωστες. Αν το μοντέλο είναι ερμηνεύσιμο θα αποκαλυφθούν οι νέες σχέσεις και η μέθοδος κατηγοριοποίησης θα χρησιμοποιηθεί ως εργαλείο ανάλυσης, ικανό να προσφέρει καινοτόμα γνώση.
- **Επεκτασιμότητα** (scalability). Αναφέρεται στην ικανότητα των μεθόδων να χειριστούν πολύ μεγάλα σύνολα δεδομένων. Η Μηχανική Μάθηση και η Στατιστική προσφέρουν μεθόδους κατηγοριοποίησης. Ωστόσο, η εφαρμογή αυτών των μεθόδων για την επεξεργασία δεδομένων μεγάλου όγκου δεν είναι πάντα εύκολη. Σε αρκετές περιπτώσεις η υπολογιστική πολυπλοκότητα των μεθόδων είναι συνάρτηση του πλήθους των παρατηρήσεων και μάλιστα με σχέση περισσότερο από γραμμική. Επίσης, οι περισσότερες μέθοδοι απαιτούν την εγκατάσταση του συνόλου εκπαίδευσης στην κύρια μνήμη του υπολογιστή. Τα ζητήματα αυτά θέτουν όρια στη δυνατότητα εφαρμογής των μεθόδων. Όμως αντικείμενο της Εξόρυξης Δεδομένων είναι η ανακάλυψη γνώσης από δεδομένα μεγάλου όγκου. Ειδικά στη σημερινή εποχή, η παραγωγή και καταγραφή δεδομένων είναι μαζικότατη. Σε ότι αφορά την εφαρμογή των μεθόδων αυτών για επιχειρηματικούς σκοπούς, η τάση που παρουσιάστηκε στα τέλη της δεκαετίας του 90’ για δημιουργία Αποθηκών Δεδομένων, έχει οδηγήσει στην αποθήκευση δεδομένων, που ο όγκος τους είναι της τάξης μεγέθους terabyte. Για να έχουν πρακτική χρησιμότητα, οι μέθοδοι Εξόρυξης Δεδομένων πρέπει να είναι ικανές να χειριστούν αυτά τα πολύ μεγάλου όγκου δεδομένα. Όπως χαρακτηριστικά επισημαίνουν οι Fayyad, Piatetsky-Shapiro and Smyth (1996), η πρόκληση για την κοινότητα των ερευνητών Εξόρυξης Δεδομένων είναι η κατασκευή μεθόδων που διευκολύνουν τη χρήση αλγορίθμων εξόρυξης δεδομένων σε βάσεις δεδομένων του πραγματικού κόσμου.
- **Ανθεκτικότητα** (robustness). Αναφέρεται στην ικανότητα των μεθόδων να πραγματοποιήσουν ορθές προβλέψεις, όταν τα δεδομένα χαρακτηρίζονται από προβλήματα, όπως ο θόρυβος και οι χαμένες τιμές.

9.6 Προεπεξεργασία δεδομένων για κατηγοριοποίηση

Όπως εξηγείται αναλυτικά στο Κεφάλαιο 6, η προεπεξεργασία των δεδομένων είναι ένα απαραίτητο στάδιο,

το οποίο προηγείται της καθαυτό εξόρυξης δεδομένων. Για την περίπτωση της κατηγοριοποίησης, η προεπεξεργασία μπορεί να βελτιώσει την αποτελεσματικότητα, την αποδοτικότητα και την επεκτασιμότητα των μεθόδων. Στο στάδιο της προεπεξεργασίας αντιμετωπίζεται το πρόβλημα του θορύβου και των χαμένων τιμών. Επίσης, τα δεδομένα μπορούν να αναχθούν σε υψηλότερα επίπεδα γενίκευσης, να διακριτοποιηθούν ώστε να μετατραπούν τα αριθμητικά πεδία σε ονομαστικά και τέλος, να κανονικοποιηθούν, να αντικατασταθούν δηλαδή οι αριθμητικές τιμές με άλλες, πιο «κατάλληλες», αριθμητικές τιμές. Σε πολλές περιπτώσεις η διακριτοποίηση και η κανονικοποίηση είναι απαραίτητες, ώστε να προσαρμοστούν τα δεδομένα σε ιδιαιτερότητες των μεθόδων κατηγοριοποίησης. Για παράδειγμα, η μέθοδος των k-Πλησιέστερων Γειτόνων είναι ιδιαίτερα ευπαθής σε δεδομένα που περιέχουν πεδία με πολύ μεγάλες τιμές και πεδία με πολύ μικρές τιμές.

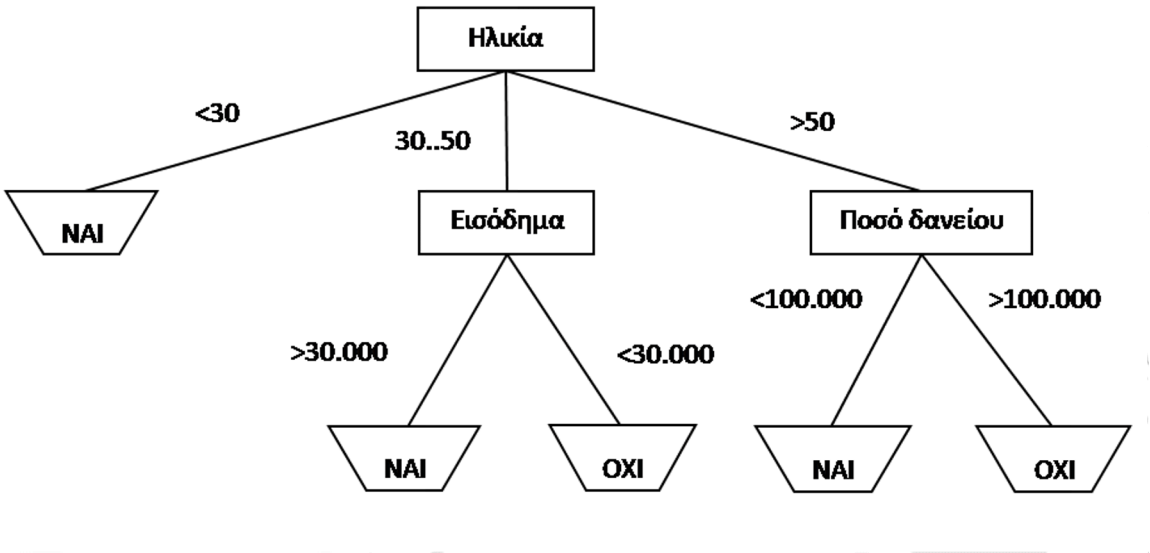
Ιδιαίτερης σημασίας είναι το ζήτημα του πλήθους των διαστάσεων και της επιλογής χαρακτηριστικών. Ερευνητικές εργασίες έχουν αποδείξει ότι το πλήθος των διαστάσεων είναι άμεσα συναρτημένο με το πλήθος των παρατηρήσεων, οι οποίες είναι απαραίτητες για την κατασκευή των μοντέλων. Ανάλογα με το είδος του κατηγοριοποιητή, το πλήθος των παρατηρήσεων μπορεί να είναι γραμμική ή και εκθετική συνάρτηση του πλήθους των διαστάσεων (Fukunaga, 1990; Hwang, Lay & Lippman, 1994). Επίσης, ορισμένες μέθοδοι όπως τα Δένδρα Αποφάσεων είναι ιδιαίτερα ευπαθείς στην ύπαρξη πολλών μεταβλητών εισόδου. Για την αντιμετώπιση του προβλήματος των πολλών διαστάσεων εφαρμόζονται μέθοδοι επιλογής χαρακτηριστικών (feature selection). Αναλυτική παρουσίαση των μεθόδων επιλογής χαρακτηριστικών γίνεται στα πλαίσια του Κεφαλαίου 7. Ωστόσο, η επιλογή χαρακτηριστικών δεν αποτελεί πανάκεια. Σε ορισμένες περιπτώσεις, είναι πιθανόν να επιλεχθεί ένας μεγάλος αριθμός μεταβλητών εισόδου. Αυτό συμβαίνει όταν το γνώρισμα της κλάσης εξαρτάται ουσιαστικά από πολλά άλλα γνωρίσματα. Επίσης, ορισμένες μέθοδοι συναρτούν το πλήθος των επιλεγμένων γνωρισμάτων με το πλήθος των παρατηρήσεων που χρησιμοποιούνται. Αν οι παρατηρήσεις που θα χρησιμοποιηθούν για την επιλογή χαρακτηριστικών είναι λίγες, τότε και τα επιλεγμένα χαρακτηριστικά θα είναι λίγα. Το αποτέλεσμα είναι η απόρριψη σημαντικών χαρακτηριστικών και ο αποκλεισμός τους από τη διαδικασία της κατηγοριοποίησης.

Σε κάθε περίπτωση, ο αναλυτής θα πρέπει να έχει επίγνωση του προβλήματος των διαστάσεων και της επιλογής σημαντικών χαρακτηριστικών σε εργασίες κατηγοριοποίησης. Ο αναλυτής θα πρέπει πιθανώς να πειραματιστεί με διαφορετικές μεθόδους επιλογής χαρακτηριστικών και να μην επαφίεται άκριτα σε μία μόνο μέθοδο.

9.7 Δένδρα Αποφάσεων

9.7.1 Εισαγωγή στα Δένδρα Αποφάσεων

Τα Δένδρα Αποφάσεων είναι μια από τις βασικότερες και πιο δημοφιλείς μεθόδους κατηγοριοποίησης. Βασική λογική της κατασκευής τους είναι η διαδοχική διάσπαση του συνόλου των παρατηρήσεων σε υποσύνολα. Κριτήριο για τη διάσπαση είναι οι τιμές των μεταβλητών. Η διαδικασία των διαδοχικών διασπάσεων αναπαρίσταται με μια ανεστραμμένη δενδρική δομή. Στην κορυφή βρίσκεται ο κόμβος-ρίζα του δένδρου. Σε κατώτερα επίπεδα βρίσκονται επιπλέον κόμβοι, οι οποίοι συνδέονται με ακμές με άλλα στοιχεία του δένδρου. Στο κατώτερο επίπεδο κάθε κλάδου βρίσκονται τα φύλλα του δένδρου. Ο κόμβος - ρίζα έχει μόνο εξερχόμενες ακμές που τον συνδέουν με στοιχεία του κατώτερου επιπέδου. Οι υπόλοιποι κόμβοι έχουν εισερχόμενες ακμές που τους συνδέουν με τους κόμβους του ανώτερου επιπέδου και εξερχόμενες ακμές που τους συνδέουν με στοιχεία του κατώτερου επιπέδου. Τέλος, τα φύλα έχουν μόνο εισερχόμενες ακμές, οι οποίες τα συνδέουν με τους κόμβους του ανώτερου επιπέδου. Κάθε κόμβος αντιπροσωπεύει έναν έλεγχο στα δεδομένα και αντίστοιχη διάσπαση τους σε δύο ή περισσότερα υποσύνολα, ανάλογα με το αποτέλεσμα του ελέγχου. Η συνηθέστερη εκδοχή είναι ο έλεγχος να περιλαμβάνει μία μόνο μεταβλητή, έχουν προταθεί ωστόσο αλγόριθμοι όπου σε έναν κόμβο ελέγχονται περισσότερες μεταβλητές. Κάθε ακμή αντιπροσωπεύει ένα αποτέλεσμα του ελέγχου και το αντίστοιχο υποσύνολο των δεδομένων. Τέλος, κάθε φύλο αντιπροσωπεύει μια απόφαση κατηγοριοποίησης.



Σχήμα 9.5 *Δένδρο Αποφάσεων*

Στο Σχήμα 9.5 απεικονίζεται ένα Δένδρο Αποφάσεων για την έγκριση τραπεζικών δανείων. Ο κόμβος-ρίζα αναφέρεται στο σύνολο των δεδομένων. Στο επίπεδο αυτό οι υποψήφιοι δανειολήπτες χωρίζονται σε τρία υποσύνολα ανάλογα με την ηλικία τους. Στο πρώτο υποσύνολο ανήκουν όσοι έχουν ηλικία μικρότερη των 30 ετών, στο δεύτερο όσοι είναι μεταξύ 30 και 50, ενώ στο τρίτο υποσύνολο ανήκουν όσοι έχουν ηλικία μεγαλύτερη των 50 ετών. Τα τρία υποσύνολα συμβολίζονται με αντίστοιχους κλάδους. Ο πρώτος κλάδος, ο οποίος αντιστοιχεί σε όσους είναι λιγότερο από 30, καταλήγει σε ένα φύλο, δηλαδή σε μια απόφαση κατηγοριοποίησης. Η απόφαση είναι θετική, και αυτό σημαίνει ότι γι' αυτήν την κατηγορία τα δάνεια εγκρίνονται χωρίς περαιτέρω ελέγχους. Ο δεύτερος κλάδος αντιστοιχεί σε όσους είναι μεταξύ 30 και 50 και καταλήγει σε έναν εσωτερικό κόμβο. Στον κόμβο αυτό γίνεται ένας δεύτερος έλεγχος που αφορά το εισόδημα. Αν το εισόδημα είναι μεγαλύτερο των 30.000 τότε το δάνειο εγκρίνεται, διαφορετικά η αίτηση απορρίπτεται. Με τον ίδιο τρόπο, οι υποψήφιοι δανειολήπτες που είναι περισσότερο από 50 χρονών ελέγχονται ως προς το ύψος του δανείου.

Το μοντέλο κατασκευάζεται από έναν αλγόριθμο με επεξεργασία ενός συνόλου δεδομένων εκπαίδευσης. Το μοντέλο, αφού κατασκευαστεί, μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση νέων παρατηρήσεων. Για κάθε νέα παρατήρηση πραγματοποιούνται έλεγχοι τιμών των μεταβλητών της, σύμφωνα με τους κόμβους του δένδρου, και ακολουθείται μια διαδρομή από τη ρίζα μέχρι κάποιο φύλο, όπου λαμβάνεται και η απόφαση κατηγοριοποίησης. Στο παράδειγμα του Σχήματος 9.5, ένας υποψήφιος δανειολήπτης θα ελεγχθεί πρώτα ως προς την ηλικία του. Εάν η ηλικία του είναι από 30 έως 50 χρονών, θα ελεγχθεί το εισόδημα του. Αν το εισόδημα του είναι μεγαλύτερο από 30.000 ευρώ το δάνειο θα εγκριθεί.

9.7.2 Δένδρα Αποφάσεων ID3

Έχουν προταθεί διάφοροι αλγόριθμοι για τη δημιουργία Δένδρων Αποφάσεων. Ένας από τους πιο διαδεδομένους είναι ο ID3, καθώς και οι μετεξελίξεις του, ο C4.5 και η εμπορική του εκδοχή C5.0. Ο ID3 προτάθηκε από τον Quinlan (1986) και υλοποιεί μια καθοδική (top-down) στρατηγική διαίρεσης. Ένα χαρακτηριστικό του είναι ότι απαιτεί την ύπαρξη μόνο ονομαστικών πεδίων. Η βασική ακολουθία βημάτων του ID3 είναι η εξής:

- Δημιουργείται ένας αρχικός κόμβος που αντιπροσωπεύει ολόκληρο το δείγμα.
- Εάν όλες οι παρατηρήσεις του δείγματος ανήκουν στην ίδια κλάση, τότε ο κόμβος μετατρέπεται σε φύλο.
- Διαφορετικά επιλέγεται το γνώρισμα που βέλτιστα διαχωρίζει τις παρατηρήσεις του δείγματος ανάλογα με την κλάση στην οποία ανήκουν.
- Δημιουργούνται κλάδοι που διαχωρίζουν τις παρατηρήσεις του δείγματος. Για κάθε δυνατή τιμή του γνωρίσματος δημιουργείται ένας κλάδος.
- Η διαδικασία επαναλαμβάνεται για κάθε ένα από τα υποσύνολα του δείγματος που δημιουργήθηκαν από τους κλάδους του προηγούμενου βήματος. Η επανάληψη τερματίζεται όταν ικανοποιηθεί του-

λάχιστον μία από τις επόμενες συνθήκες εξόδου:

- Όλες οι παρατηρήσεις ενός κόμβου ανήκουν στην ίδια κλάση.
- Δεν υπάρχουν άλλα γνωρίσματα για τον διαχωρισμό του δείγματος. Σε αυτήν την περίπτωση ο κόμβος μετατρέπεται σε φύλο. Η απόφαση κατηγοριοποίησης στο φύλο είναι η κλάση που πλειοψηφεί στο συγκεκριμένο υποσύνολο παρατηρήσεων.
- Δεν υπάρχουν παρατηρήσεις που να ανήκουν στο υποσύνολο του δείγματος που ορίζει ο κλάδος.

Το βασικότερο πρόβλημα στα Δένδρα Αποφάσεων είναι ο καθορισμός του κριτηρίου, βάση του οποίου θα γίνει ο διαχωρισμός των παρατηρήσεων. Έχουν προταθεί αλγόριθμοι οι οποίοι πραγματοποιούν ελέγχους σε συνδυασμούς μεταβλητών (multivariate). Όμως στους περισσότερους αλγορίθμους, συμπεριλαμβανομένου του ID3, ο έλεγχος πραγματοποιείται σε μία μόνο μεταβλητή (univariate). Διάφοροι ερευνητές έχουν επινοήσει και προτείνει πλήθος μονομεταβλητών κριτηρίων. Σε ένα δεδομένο σημείο του δένδρου, το ερώτημα που τίθεται είναι ποιο γνώρισμα πρέπει να χρησιμοποιηθεί για τον διαχωρισμό των παρατηρήσεων. Στον ID3 το κριτήριο που χρησιμοποιείται ονομάζεται Κέρδος Πληροφορίας (ΚΠ). Για κάθε διαθέσιμο γνώρισμα υπολογίζεται το Κέρδος Πληροφορίας και επιλέγεται το γνώρισμα με τη μεγαλύτερη τιμή ΚΠ.

Το ***Κέρδος Πληροφορίας(S,A)*** *(Information Gain(S,A))* εκφράζει τη μείωση της εντροπίας που θα προκύψει, εάν ένα σύνολο παρατηρήσεων *S* διαχωριστεί σε υποσύνολα με βάση τις τιμές του γνωρίσματος *A*. Η εντροπία μετρά την ανομοιογένεια του συνόλου *S,* ανάλογα με τη διασπορά των παρατηρήσεων ως προς την κλάση στην οποία ανήκουν. Ας θεωρήσουμε ένα σύνολο *S* το οποίο περιέχει s παρατηρήσεις. Εάν η κλάση είναι δυαδική, εάν δηλαδή υπάρχουν δύο δυνατές τιμές για το γνώρισμα της κλάσης, τότε οι παρατηρήσεις της μίας τιμής κλάσης μπορούν να χαρακτηριστούν θετικές, ενώ οι υπόλοιπες μπορούν να χαρακτηριστούν αρνητικές. Το πλήθος των θετικών παρατηρήσεων είναι s_p και το πλήθος των αρνητικών παρατηρήσεων είναι s_n. Η εντροπία του συνόλου *S* ορίζεται από την Εξίσωση 9.1

$$E(S) = -p_p * log_2(p_p) - p_n * log_2(p_n) \tag{9.1}$$

όπου p_p είναι το ποσοστό των θετικών παρατηρήσεων ($p_p=s_p/s$) και p_n είναι το ποσοστό των αρνητικών παρατηρήσεων ($p_n=s_n/s$).

Εάν το γνώρισμα της κλάσης μπορεί να πάρει c διαφορετικές τιμές και το πλήθος των παρατηρήσεων με τιμή κλάσης *i* είναι s_i, τότε η εντροπία του *S* ορίζεται από την Εξίσωση 9.2

$$E(S) = -\sum_{i=1}^{c} p_i * log_2(p_i) \tag{9.2}$$

όπου p_i είναι το ποσοστό των παρατηρήσεων που ανήκουν στην κλάση *i* ($p_i=s_i/s$)

Ας θεωρήσουμε ότι το γνώρισμα *A* μπορεί να πάρει u δυνατές διακριτές τιμές ($a_1,a_2,\ldots,a_u$). Το σύνολο *S* μπορεί να χωριστεί στα υποσύνολα ($S_1,S_2,\ldots,S_u$). Το S_1 αποτελείται από τις παρατηρήσεις οι οποίες έχουν τιμή a_1 στο γνώρισμα *A*. Αντιστοίχως, τα υπόλοιπα υποσύνολα S_j απαρτίζονται από τις παρατηρήσεις που έχουν την εκάστοτε τιμή a_j στο γνώρισμα *A*. Εάν επιλεχθεί ως μεταβλητή διαχωρισμού το γνώρισμα *A*, τότε η εντροπία του διαχωρισμού του συνόλου *S* σε υποσύνολα ανάλογα με τις τιμές του *A* δίνεται από την Εξίσωση 9.3

$$E(S,A) = \sum_{j=1}^{u} \frac{s_j}{s} * E(S_j) \tag{9.3}$$

όπου *u* το πλήθος των δυνατών τιμών του γνωρίσματος *A*, S_j το υποσύνολο των παρατηρήσεων οι οποίες έχουν την τιμή a_j στο γνώρισμα *A*, s_j το πλήθος των μελών του S_j, *s* είναι το πλήθος των μελών του *S* και $E(S_j)$ είναι η εντροπία του S_j, η οποία υπολογίζεται σύμφωνα με την Εξίσωση 9.2 και με το S_j στη θέση του *S*. Ου-

σιαστικά η εντροπία που προκύπτει από τον διαχωρισμό του S ισούται με το άθροισμα των εντροπιών των S_j πολλαπλασιασμένες με έναν συντελεστή βαρύτητας, ο οποίος σχετίζεται με το πλήθος των μελών τους. Όσο μικρότερη είναι η εντροπία τόσο αυξάνει ο βαθμός ομοιογένειας των υποσυνόλων.

Το Κέρδος Πληροφορίας είναι η μείωση της εντροπίας, η οποία προκύπτει από τον διαχωρισμό και ορίζεται από την Εξίσωση 9.4

$$IG(S,A) = E(S) - E(S,A)$$

(9.4)

Ο ID3 υπολογίζει για κάθε γνώρισμα το Κέρδος Πληροφορίας. Το γνώρισμα με το μεγαλύτερο Κέρδος Πληροφορίας επιλέγεται και ο διαχωρισμός των παρατηρήσεων γίνεται με βάση τις τιμές αυτού του γνωρίσματος. Με τον τρόπο αυτόν μεταβαίνουμε σε υποσύνολα μεγαλύτερης ομοιογένειας.

9.7.3 Δένδρα Αποφάσεων C4.5

Ο αλγόριθμος C4.5 αποτελεί επέκταση του ID3 και προτάθηκε από τον ίδιο ερευνητή (Quinlan, 1993). Μια από τις βασικές βελτιώσεις αφορά το κριτήριο διαχωρισμού. Σύμφωνα με τον Quinlan το Κέρδος Πληροφορίας τείνει να ευνοεί γνωρίσματα με μεγάλο πλήθος τιμών. Τα γνωρίσματα αυτά οδηγούν σε μεγάλο αριθμό μικρών και πολύ ομοιογενών υποσυνόλων. Σε πολλές περιπτώσεις όμως, τα γνωρίσματα αυτά δεν περιέχουν ουσιαστική πληροφορία. Αν για παράδειγμα τα δεδομένα περιέχουν πεδίο για κάποιον κωδικό, όπως ο αριθμός ταυτότητας, τότε το πεδίο αυτό θα έχει μεγάλο κέρδος πληροφορίας και θα επιλεγεί. Ωστόσο, δεν περιέχει πληροφορία χρήσιμη για την κατηγοριοποίηση. Για την αντιμετώπιση αυτού του προβλήματος, στον C4.5 χρησιμοποιείται το κριτήριο Λόγος Κέρδους (Gain Ratio), το οποίο ορίζεται με την Εξίσωση 9.5.

$$GainRatio(S,A) = \frac{Information\ Gain\ (S,A)}{Entropy(S,A)}$$

(9.5)

Ο Λόγος Κέρδους κανονικοποιεί το κέρδος πληροφορίας ως προς την εντροπία. Μελέτες έχουν δείξει ότι ο Λόγος Κέρδους βελτιώνει την ακρίβεια και μειώνει την πολυπλοκότητα των δένδρων.

Μια άλλη σημαντική βελτίωση στον C4.5 είναι ότι, σε αντίθεση με τον ID3, μπορεί και χειρίζεται πεδία αριθμητικών τιμών. Για κάθε αριθμητικό πεδίο, ο αλγόριθμος ταξινομεί τις τιμές του, το πλήθος των οποίων είναι πεπερασμένο, σε αύξουσα σειρά, και ορίζει μια τιμή κατωφλιού. Με τον τρόπο αυτόν οι παρατηρήσεις χωρίζονται σε εκείνες των οποίων η τιμή στο συγκεκριμένο πεδίο είναι μικρότερη ή ίση με την τιμή κατωφλιού και σε εκείνες που η τιμή τους είναι μεγαλύτερη. Ακολούθως, το γνώρισμα αντιμετωπίζεται σαν να έχει διακριτές τιμές, όπου οι δύο διακριτές τιμές είναι οι δύο καθορισμένες περιοχές συνεχών τιμών. Επίσης ο C4.5 μπορεί και χειρίζεται δεδομένα με χαμένες τιμές.

9.7.4 Δένδρα CART

Τα δένδρα τύπου CART (Classification And Regression Trees) προτάθηκαν από τους Breiman, Friedman, Olshen and Stone (1984). Ο πρωτότυπος αλγόριθμος τους είναι ενσωματωμένος σε λογισμικά της Salford Systems. Τα δένδρα CART παρουσιάζουν αρκετά ενδιαφέροντα χαρακτηριστικά. Ένα από τα σημαντικότερα χαρακτηριστικά τους είναι ότι μπορούν να χρησιμοποιηθούν και για κατηγοριοποίηση και για παλινδρόμηση, μπορούν δηλαδή να προβλέψουν και ονομαστικές τιμές κλάσης και τιμές αριθμητικών πεδίων. Τα δένδρα CART είναι δυαδικά και κάθε κόμβος μπορεί να έχει μόνο δύο κλάδους. Για τον διαχωρισμό των παρατηρήσεων χρησιμοποιείται το κριτήριο Twoing. Στα πλεονεκτήματα τους περιλαμβάνονται οι υψηλές επιδόσεις σε ταχύτητα και ακρίβεια, καθώς και η ικανότητα τους να χειρίζονται δεδομένα με χαμένες τιμές. Επιπλέον, τα δένδρα CART είναι ικανά να εκτελέσουν κατηγοριοποίηση που λαμβάνει υπόψη το διαφορετικό κόστος σφάλματος (cost sensitive classification). Σε αυτήν την περίπτωση ο αλγόριθμος επιδιώκει να μειώσει τις εσφαλμένες προβλέψεις της πιο ακριβής κλάσης. Το αντικείμενο του διαφορετικού κόστους σφάλματος καλύπτεται στο Κεφάλαιο 10.

Πρόσθετοι αλγόριθμοι δημιουργίας δένδρων αποφάσεων έχουν προταθεί από διάφορους ερευνητές. Ορι-

σμένα δένδρα αποφάσεων είναι τα AID (Sonquist, Baker & Morgan, 1971), CHAID (Kass, 1980) και QUEST (Loh & Shih, 1997).

9.7.5 Κλάδεμα

Κατά τη διαδικασία ανάπτυξης του δένδρου, ο αλγόριθμος έρχεται αντιμέτωπος με παρατηρήσεις που περιέχουν εσφαλμένες ή ακραίες τιμές. Η δημιουργία κλάδων που να αντιστοιχούν σε τέτοιου είδους τιμές, καταγράφει ανωμαλίες των δεδομένων και περιορίζει την ικανότητα του δένδρου να προβλέψει την κλάση νέων παρατηρήσεων. Ένα πρόσθετο ζήτημα είναι οι συνθήκες τερματισμού του βρόχου ανάπτυξης του δένδρου. Αν τα κριτήρια είναι πολύ περιοριστικά, θα δημιουργηθούν μικρά και υποπροσαρμοσμένα δένδρα, ενώ αν τα κριτήρια είναι πολύ χαλαρά, θα δημιουργηθούν υπερβολικά μεγάλα και υπερπροσαρμοσμένα δένδρα. Για την αντιμετώπιση αυτού του προβλήματος προτάθηκε από τους Breiman et al. (1984) μια τεχνική, η οποία προβλέπει την εφαρμογή χαλαρών κριτηρίων, τη δημιουργία υπερπροσαρμοσμένων δένδρων και την ακόλουθη απομάκρυνση των περιττών κλάδων. Η διαδικασία διαγραφής των περιττών κλάδων καλείται **κλάδεμα** (pruning).

Έχουν προταθεί πολλές τεχνικές κλαδέματος. Μια από τις πιο γνωστές είναι η τεχνική Cost-Complexity Pruning. Για κάθε κόμβο του δένδρου υπολογίζεται το αναμενόμενο σφάλμα που θα προκύψει εάν κλαδευτεί το υποδένδρο του κόμβου, καθώς και το αναμενόμενο σφάλμα εάν τα υποδένδρο δεν κλαδευτεί. Εάν το κλάδεμα οδηγήσει σε μεγαλύτερο αναμενόμενο σφάλμα, τότε το υποδένδρο διατηρείται, διαφορετικά κλαδεύεται. Προβλέπεται η δημιουργία πολλών εναλλακτικών δένδρων με κλάδεμα που αυξάνεται προοδευτικά. Τα εναλλακτικά αυτά δένδρα δοκιμάζονται έναντι ενός συνόλου άγνωστων παρατηρήσεων και επιλέγεται το δένδρο το οποίο ελαχιστοποιεί τον αναμενόμενο ρυθμό σφάλματος.

Η προσέγγιση της ανάπτυξης ενός λεπτομερούς δένδρου και του ακόλουθου κλαδέματος του επισύρει πρόσθετο υπολογιστικό κόστος. Εναλλακτικά, μπορεί κατά τη διάρκεια της δημιουργίας του δένδρου να διακοπεί η ανάπτυξη περιττών κλάδων με τον καθορισμό κατάλληλων συνθηκών εξόδου. Η προσέγγιση αυτή έχει το πλεονέκτημα ότι αποφεύγει την άσκοπη εργασία δημιουργίας άχρηστων κλάδων. Ωστόσο, η τεχνική του κλαδέματος αποδίδει καλύτερα αποτελέσματα, ειδικά στην περίπτωση όπου ένας κλάδος δεν είναι σημαντικός, αλλά δύο συνεχόμενοι κλάδοι επιτυγχάνουν ισχυρή κατηγοριοποίηση. Άλλες τεχνικές κλαδέματος είναι η Minimum Description Length Pruning (Quinlan & Rivest, 1989), η Minimum Error Pruning (Olaru & Wehenkel, 2003) και η Pessimistic Pruning (Quinlan, 1993).

9.7.6 Δημιουργία κανόνων από Δένδρα Αποφάσεων

Στα Δένδρα Αποφάσεων η ανακαλυφθείσα γνώση αναπαρίσταται με τέτοιον τρόπο, ώστε είναι εύκολη η εξαγωγή επαγωγικών κανόνων της μορφής IF-THEN. Για κάθε φύλο δημιουργείται ένας κανόνας, ο οποίος περιλαμβάνει τις λογικές συνθήκες όλων των ελέγχων από τη ρίζα έως το φύλο. Οι λογικές αυτές συνθήκες συνδέονται με τον λογικό τελεστή AND. Ωστόσο, μια τέτοιου τύπου εξαγωγή κανόνων, οδηγεί συχνά σε κανόνες περιττά περίπλοκους. Για την απλοποίηση των κανόνων μπορεί να υιοθετηθεί μια προσέγγιση κλαδέματος, η οποία συνίσταται στην απομάκρυνση των μη σημαντικών συνθηκών. Ο καθορισμός των μη σημαντικών συνθηκών επιτυγχάνεται με τη σύγκριση του ρυθμού σφάλματος του απλουστευμένου κανόνα με τον ρυθμό σφάλματος του πλήρους κανόνα.

9.7.7 Πλεονεκτήματα και Μειονεκτήματα των Δένδρων Αποφάσεων

Τα Δένδρα Αποφάσεων προσφέρουν πολλά και σημαντικά **πλεονεκτήματα**:

- Σε αντίθεση με άλλες μεθόδους δεν κάνουν αυθαίρετες υποθέσεις για τη γραμμικότητα της σχέσης μεταξύ των μεταβλητών εισόδου και εξόδου ή για την ανεξαρτησία των μεταβλητών εισόδου.
- Τα Δένδρα Αποφάσεων είναι μη παραμετρικά. Η δημιουργία του δένδρου δεν καθορίζεται από τον καθορισμό πολλών και σύνθετων παραμέτρων.
- Η αναπαράσταση της γνώσης γίνεται με κατανοητό τρόπο και είναι εύκολη η εξαγωγή κατανοητών κανόνων.
- Δέχονται ως μεταβλητές εισόδου και ονομαστικά γνωρίσματα και γνωρίσματα με αριθμητικές τιμές.
- Μπορούν να χειριστούν δεδομένα με χαμένες τιμές.
- Διαθέτουν ένα εξαιρετικά γρήγορο αλγόριθμο εκπαίδευσης.

Μειονεκτήματα των Δένδρων Απόφασης είναι τα εξής:

- Αρκετοί αλγόριθμοι, όπως ο ID3 και ο C4.5, λειτουργούν μόνο για διακριτές και όχι για συνεχόμενες τιμές κλάσης.
- Είναι ιδιαίτερα ευπαθή σε μεταβολές του δείγματος εκπαίδευσης. Οριακές μεταβολές του δείγματος εκπαίδευσης μπορεί να οδηγήσουν στη δημιουργία σημαντικά διαφορετικών δένδρων.
- Αρκετοί αλγόριθμοι Δένδρων Απόφασης, όπως ο ID3 ή ο C4.5, απαιτούν την εγκατάσταση ολόκληρου του δείγματος εκπαίδευσης στην κύρια μνήμη του υπολογιστή, κάτι που προκαλεί προβλήματα στον χειρισμό εξαιρετικά μεγάλων δειγμάτων.

9.8 Κατηγοριοποίηση με Νευρωνικά Δίκτυα

Τα **Νευρωνικά Δίκτυα** (Neural Networks) αποτελούν ένα από τα σημαντικότερα επιτεύγματα της Τεχνητής Νοημοσύνης. Εμπνευσμένα από το βιολογικό νευρικό σύστημα, και ειδικότερα από τον ανθρώπινο εγκέφαλο, διαθέτουν αξιοσημείωτα χαρακτηριστικά, όπως τη δυνατότητα τους να αναπαριστούν σύνθετες εξαρτήσεις ή την ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Χάρη στη στιβαρή θεωρητική τους θεμελίωση και στις ιδιαίτερες δυνατότητες τους έχουν καταστεί ιδιαίτερα δημοφιλή και έχουν εφαρμοστεί σε πολλούς τομείς, όπως η ιατρική, η οικονομία, η άμυνα κλπ. Τα Νευρωνικά Δίκτυα είναι μια τεχνική ισχυρά καθοδηγούμενη από τα δεδομένα. Αυτό σημαίνει ότι δεν επιβάλλουν αυθαίρετες υποθέσεις και ότι τα μοντέλα τους πηγάζουν από την επεξεργασία των δεδομένων. Έχουν προταθεί τύποι Νευρωνικών Δικτύων κατάλληλοι για επιβλεπόμενη, αλλά και για μη επιβλεπόμενη μάθηση. Στο παρόν κεφάλαιο καλύπτονται θέματα Νευρωνικών Δικτύων επιβλεπόμενης μάθησης. Αναφορά στα Νευρωνικά Δίκτυα μη επιβλεπόμενης μάθησης, και ειδικά στους Αυτοοργανούμενους Χάρτες, γίνεται στο Κεφάλαιο 11.

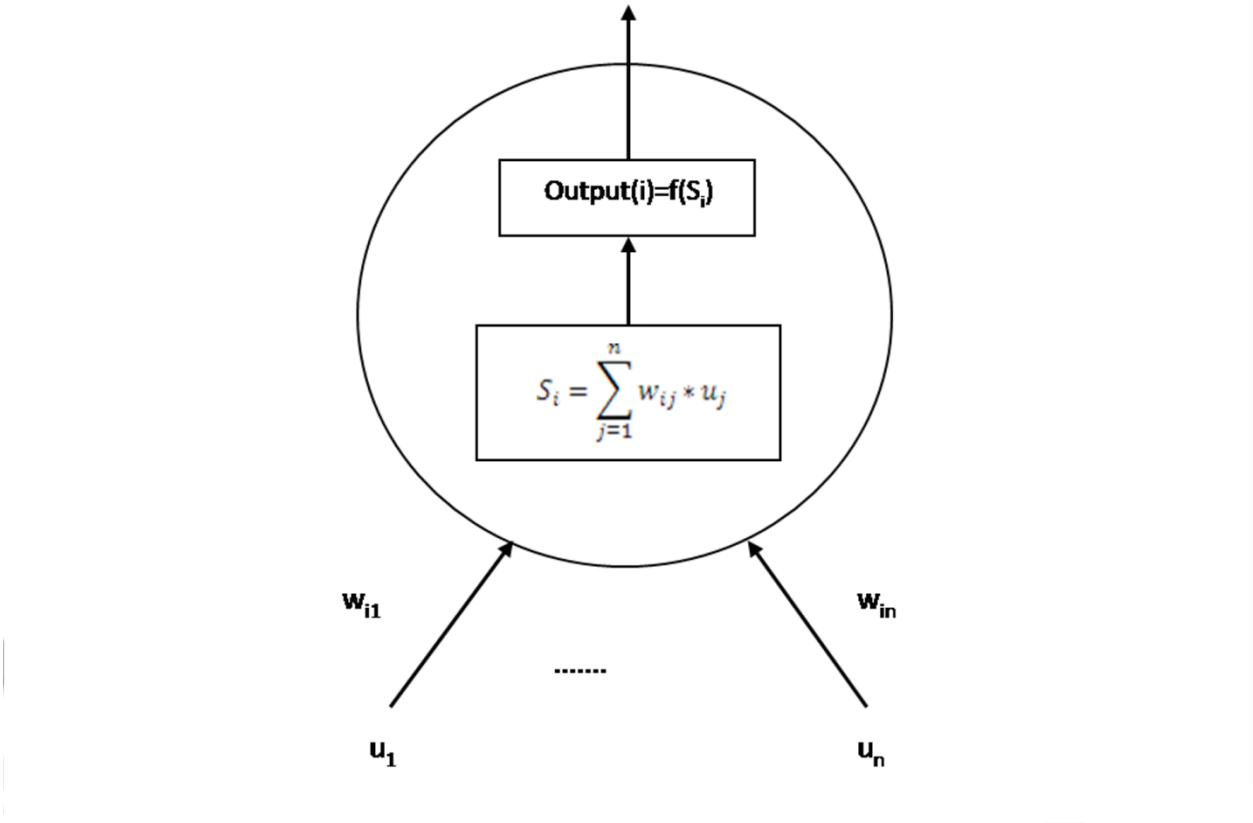
9.8.1 Νευρώνες και Συνδέσεις

Βασική δομική μονάδα των Νευρωνικών Δικτύων είναι οι νευρώνες. Οι νευρώνες ονομάζονται επίσης κόμβοι ή κελιά. Ένας **νευρώνας** είναι μια στοιχειώδης υπολογιστική μονάδα, η οποία δέχεται τιμές εισόδου και υπολογίζει μια τιμή εξόδου. Οι νευρώνες συνδέονται μεταξύ τους με κατευθυνόμενα βέλη ή συνδέσεις. Μέσω των συνδέσεων ένας νευρώνας δέχεται τιμές εισόδου από άλλους νευρώνες. Επίσης, μέσω των συνδέσεων μεταβιβάζει την τιμή εξόδου του σε άλλους νευρώνες. Κάθε σύνδεση συνοδεύεται από μία αριθμητική τιμή που ονομάζεται **βάρος** (weight) w. Το βάρος επηρεάζει την επίδραση μεταξύ των συνδεδεμένων νευρώνων. Εάν u_j είναι η τιμή εξόδου του νευρώνα j, και η τιμή αυτή μεταβιβάζεται στον νευρώνα i, τότε το u_j πολλαπλασιάζεται με το βάρος της σύνδεσης των δύο νευρώνων w_{ij}.

Η επεξεργασία που διενεργεί ένας νευρώνας i ολοκληρώνεται σε δύο στάδια. Στο πρώτο στάδιο αθροίζονται οι τιμές εισόδου. Οι τιμές εισόδου ισούνται με τις τιμές εξόδου των συνδεδεμένων νευρώνων, πολλαπλασιασμένες με τα βάρη των αντίστοιχων συνδέσεων. Για έναν νευρώνα i ο οποίος δέχεται τιμές εισόδου u_j από n νευρώνες, το συνολικό σήμα εισόδου S_i υπολογίζεται σύμφωνα με την Εξίσωση 9.6.

$$S_i = \sum_{j=1}^{n} w_{ij} * u_j \tag{9.6}$$

Στο δεύτερο στάδιο, μετασχηματίζεται το άθροισμα των τιμών εισόδου, με χρήση μιας συνάρτησης γνωστής ως **συνάρτηση ενεργοποίησης** (activation function) ή **συνάρτηση μετασχηματισμού**. Η τιμή που υπολογίζεται είναι η τιμή εξόδου του νευρώνα. Τα παραπάνω απεικονίζονται στο Σχήμα 9.6.



Σχήμα 9.6 *Ενεργοποίηση Νευρώνα*

Διάφορες συναρτήσεις μπορούν να χρησιμοποιηθούν ως συναρτήσεις ενεργοποίησης. Τέτοιες συναρτήσεις είναι η συνάρτηση ημιτόνου, η συνάρτηση συνημίτονου, η συνάρτηση υπερβολικής εφαπτομένης κλπ. Συνήθως όμως χρησιμοποιείται η Σιγμοειδής συνάρτηση, επειδή είναι απλή και μη γραμμική και επειδή μοιάζει με τη συμπεριφορά των πραγματικών νευρώνων. Η Σιγμοειδής συνάρτηση ορίζεται από την Εξίσωση 9.7.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{9.7}$$

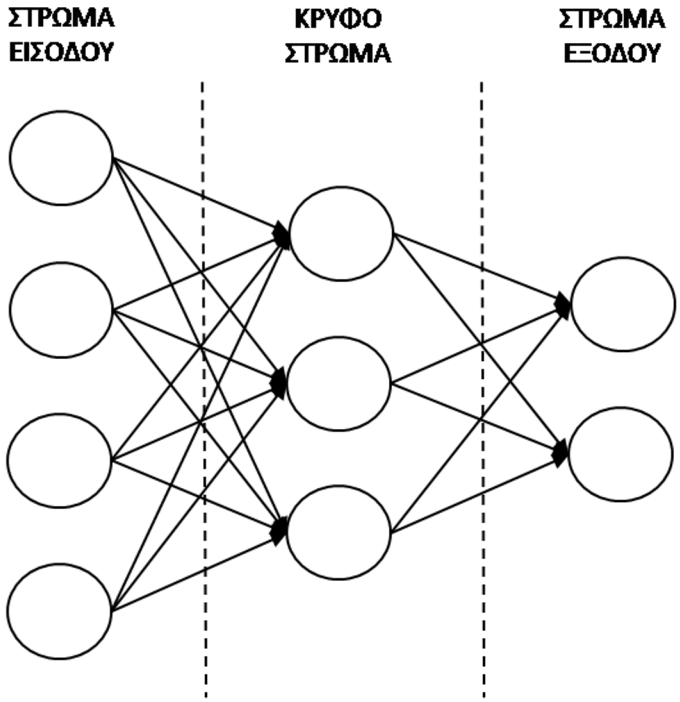
Οι συνδέσεις μπορεί να είναι μονόδρομες ή αμφίδρομες. Όταν ένα δίκτυο δεν περιέχει αμφίδρομες συνδέσεις χαρακτηρίζεται δίκτυο **απλής προώθησης** (feed forward), ενώ όταν περιέχει και αμφίδρομες συνδέσεις χαρακτηρίζεται **αναδρομικό** (recurrent). Τα δίκτυα απλής προώθησης και πολλών επιπέδων είναι ιδιαίτερα αποτελεσματικά για τη μοντελοποίηση σύνθετων, μη γραμμικών σχέσεων ανάμεσα σε μια εξαρτημένη μεταβλητή και πολλές ανεξάρτητες μεταβλητές. Για τον λόγο αυτό, χρησιμοποιούνται συχνά σε προβλήματα κατηγοριοποίησης. Τα νευρωνικά δίκτυα τύπου **Multilayer Perceptron** (**MLP**) είναι δίκτυα απλής προώθησης και ιδιαίτερα δημοφιλή. Σύμφωνα με τους Wong, Bodnovich and Selvi (1997), στο 95% των περιπτώσεων επιχειρηματικών εφαρμογών χρησιμοποιούνται δίκτυα αυτού του τύπου.

9.8.2 Δομή MLP

Σε ένα δίκτυο MLP οι νευρώνες είναι οργανωμένοι σε **στρώματα** (layers) ή επίπεδα. Παράδειγμα Νευρωνικού Δικτύου με επίπεδα παρουσιάζεται στο Σχήμα 9.7. Το πρώτο στρώμα ονομάζεται **στρώμα εισόδου** (input layer). Υπάρχει ένας νευρώνας εισόδου για κάθε ανεξάρτητη μεταβλητή. Χαρακτηριστικό των νευρώνων εισόδου είναι ότι δεν μετασχηματίζουν την τιμή. Απλά δέχονται την τιμή της ανεξάρτητης μεταβλητής και την μεταβιβάζουν στους επόμενους νευρώνες. Το δεύτερο στρώμα ονομάζεται **κρυφό στρώμα** (hidden layer). Οι νευρώνες του κρυφού στρώματος δέχονται τις τιμές των νευρώνων εισόδου πολλαπλασιασμένες με τα βάρη των συνδέσεων, τις αθροίζουν και μετασχηματίζουν το άθροισμα σύμφωνα με τη συνάρτηση μετα-

σχηματισμού. Οι κρυφοί νευρώνες είναι καθοριστικής σημασίας για την καταγραφή των σύνθετων σχέσεων των δεδομένων. Οι τιμές εξόδου των κρυφών νευρώνων, πολλαπλασιασμένες με τα βάρη των συνδέσεων, διαβιβάζονται στους νευρώνες του **στρώματος εξόδου** (output layer). Στους νευρώνες εξόδου υπολογίζεται η τελική πρόβλεψη του δικτύου. Είναι δυνατόν να υπάρχουν περισσότερα κρυφά στρώματα, συνήθως όμως χρησιμοποιείται μόνο ένα κρυφό στρώμα. Για διχότομα προβλήματα κατηγοριοποίησης ένας νευρώνας εξόδου είναι αρκετός. Για προβλήματα με περισσότερες τιμές κλάσης χρειάζεται ένας νευρώνας εξόδου για κάθε δυνατή τιμή της κλάσης. Είναι δυνατόν να υπάρχουν διαφορετικές συναρτήσεις μετασχηματισμού σε νευρώνες διαφορετικών επιπέδων. Οι νευρώνες είναι **πλήρως συνδεδεμένοι** (fully connected), και κάθε νευρώνας διαβιβάζει τιμές σε όλους τους νευρώνες του επόμενου στρώματος και μόνο σε αυτούς. Επίσης, μπορεί να υπάρχει ένας κατά σύμβαση νευρώνας πόλωσης *0*, ο οποίος είναι συνδεδεμένος με όλους τους υπόλοιπους νευρώνες και του οποίου η έξοδος u_0 είναι σταθερά *+1*. Τα βάρη w_{i0} ονομάζονται *πόλωση* (bias).

Ο όρος **αρχιτεκτονική του δικτύου** (network architecture) ή **τοπολογία του δικτύου** (network topology) αναφέρεται στη δομή του δικτύου και περιλαμβάνει ζητήματα όπως το πλήθος των κρυφών στρωμάτων και το πλήθος των νευρώνων σε κάθε στρώμα. Ο χρήστης του νευρωνικού δικτύου οφείλει να προκαθορίσει την αρχιτεκτονική του δικτύου πριν από την εκπαίδευση του. Επίσης, προκαθορίζει τη συνάρτηση μετασχηματισμού των νευρώνων. Η αρχιτεκτονική του δικτύου είναι σημαντική και επηρεάζει την αποτελεσματικότητα του.



Σχήμα 9.7 *Δίκτυο τριών επιπέδων*

9.8.3 Εκπαίδευση Δικτύου

Η εκπαίδευση ενός δικτύου συνίσταται στη ρύθμιση των βαρών των συνδέσεων. Για την εκπαίδευση ενός Νευρωνικού Δικτύου τυπικά απαιτείται ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου. Το σύνολο εκπαίδευσης χρησιμοποιείται για τον καθορισμό των βαρών των συνδέσεων. Το σύνολο ελέγχου χρησιμοποιείται για την εκτίμηση της επίδοσης του μοντέλου. Στην επιβλεπόμενη μάθηση, κάθε παρατήρηση του συνόλου εκπαίδευσης ή ελέγχου περιλαμβάνει και την τιμή της κλάσης της συγκεκριμένης παρατήρησης.

Υπάρχουν διάφοροι αλγόριθμοι για την εκπαίδευση του δικτύου. Ο πιο σημαντικός και διαδεδομένος αλγόριθμος, λόγω της καταλληλότητας του για επιβλεπόμενη μάθηση και για καθήκοντα κατηγοριοποίησης και πρόβλεψης, είναι ο αλγόριθμος **Αντίστροφης Μετάδοσης Σφάλματος** (**Backpropagation**). Συνεισφορά στη διατύπωση του Backpropagation είχαν οι Parker (1985) και Rumelhart, Hinton and Williams (1896).

Για να εκπαιδεύσει το δίκτυο, ο αλγόριθμος Backpropagation εφαρμόζει μια επαναλαμβανόμενη διαδικα-

σία, όπου μια παρατήρηση εκπαίδευσης εφαρμόζεται στο δίκτυο και ακολούθως υπολογίζεται στην έξοδο μια πρόβλεψη για την κλάση της παρατήρησης. Η πρόβλεψη συγκρίνεται με την πραγματική κλάση. Στη συνέχεια τροποποιούνται τα βάρη των συνδέσεων, έτσι ώστε να ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα μεταξύ της πρόβλεψης και της πραγματικής κλάσης. Η τροποποίηση των βαρών γίνεται αρχίζοντας από το επίπεδα εξόδου και συνεχίζοντας προς τα προηγούμενα επίπεδα. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να ικανοποιηθούν οι συνθήκες τερματισμού. Πιο αναλυτικά, τα βήματα του αλγόριθμου Backpropagation είναι τα εξής.

- **Εκχώρηση αρχικών τιμών στα βάρη**. Τα βάρη των συνδέσεων αρχικοποιούνται με τυχαίες τιμές. Οι τιμές αρχικοποίησης είναι μικρές και μπορεί να κυμαίνονται από -1,0 έως 1,0.
- **Διάδοση της εισόδου**. Μια παρατήρηση εκπαίδευσης εφαρμόζεται στην είσοδο του δικτύου. Τα δεδομένα εισόδου διαδίδονται στους νευρώνες των επόμενων στρωμάτων και μετασχηματίζονται χρησιμοποιώντας τα βάρη των συνδέσεων και τις συναρτήσεις μετασχηματισμού. Στο επίπεδο εξόδου υπολογίζεται μια πρόβλεψη.
- **Διάδοση του σφάλματος προς τα πίσω**. Η τιμή κλάσης που υπολόγισε το δίκτυο συγκρίνεται με την πραγματική τιμή. Αν η πρόβλεψη είναι εσφαλμένη, το σφάλμα διαδίδεται προς τα πίσω και επαναϋπολογίζονται τα βάρη των συνδέσεων. Για έναν νευρώνα i που βρίσκεται στο επίπεδο εξόδου το σφάλμα υπολογίζεται ως:

$$Err_i = f(i) * \big(1 - f(i)\big) * \big(T_i - f(i)\big)$$

(9.8)

όπου T_i είναι η πραγματική τιμή που θα έπρεπε να υπολογιστεί με βάση την πραγματική κλάση της συγκεκριμένης παρατήρησης εκπαίδευσης.

Για έναν νευρώνα *i* που βρίσκεται σε κρυφό επίπεδο, το σφάλμα υπολογίζεται σύμφωνα με τη Σχέση 9.9

$$Err_i = f(i) * \big(1 - f(i)\big) * \sum_j Err_j * w_{ij}$$

(9.9)

όπου *j* είναι οι νευρώνες του επόμενου επιπέδου με τους οποίους είναι συνδεδεμένος ο *i*, w_{ij} είναι τα αντίστοιχα βάρη συνδέσεων και Err_j είναι τα σφάλματα των νευρώνων *j*. Τα βάρη των συνδέσεων τροποποιούνται σύμφωνα με την Εξίσωση 9.10

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

(9.10)

όπου το Δw_{ij} ορίζεται από την Εξίσωση 9.11.

$$\Delta w_{ij} = (I) * Err_j * f(i)$$

(9.11)

Η σταθερά *I* στην Εξίσωση 9.12 συμβολίζει τον **ρυθμό εκπαίδευσης** (learning rate) και τυπικά παίρνει τιμές από 0,0 έως και 1,0. Ο ρυθμός εκπαίδευσης επιτρέπει να ρυθμίσουμε τον βαθμό μεταβολής των βαρών σε κάθε επανάληψη. Κατ' επέκταση, ο ρυθμός εκπαίδευσης επηρεάζει την ταχύτητα εκπαίδευσης του δικτύου. Ένας άλλος τρόπος ενίσχυσης του ρυθμού μεταβολής των βαρών είναι με τη χρήση της ορμής (momentum) Η χρήση του momentum συνίσταται στην αύξηση της μεταβολής των βαρών με την πρόσθεση ενός ποσοστού της προηγούμενης μεταβολής. Η προσθήκη του νέου προσθετέου τείνει να διατηρήσει την κατεύθυνση μεταβολής των βαρών. Ο ορμή βοηθά το μοντέλο να μην παγιδευτεί σε τοπικά ελάχιστα.

- **Επανάληψη** της διάδοσης εισόδου με τροφοδοσία μιας νέας παρατήρησης στο δίκτυο **μέχρι την ικανοποίηση της συνθήκης εξόδου**. Συνθήκη εξόδου μπορεί να είναι μια από τις παρακάτω:

    - όλα τα Δw_{ij} έχουν τιμή κάτω από ένα όριο, ή
    - το ποσοστό των εσφαλμένων προβλέψεων είναι κάτω από ένα όριο, ή
    - συμπληρώθηκε ο προκαθορισμένος αριθμός εποχών.

Σύμφωνα με τον αλγόριθμο που περιγράφτηκε προηγουμένως, ο επαναϋπολογισμός των βαρών γίνεται για κάθε παρατήρηση εκπαίδευσης. Σε μια άλλη παραλλαγή του αλγόριθμου οι μεταβολές των βαρών συσσωρεύονται και τα βάρη των συνδέσεων αλλάζουν τιμή όταν ολοκληρωθεί η ανάγνωση όλου του συνόλου εκπαίδευσης. Κάθε επανάληψη του συνόλου εκπαίδευσης καλείται **εποχή**.

9.8.4 Θέματα μοντελοποίησης με νευρωνικά δίκτυα

Η δημιουργία και η εκπαίδευση ενός επιτυχημένου νευρωνικού δικτύου είναι μια απαιτητική και δύσκολη εργασία. Αρχικά πρέπει να καθοριστούν το πλήθος των κρυφών στρωμάτων και το πλήθος των νευρώνων σε κάθε στρώμα. Δυστυχώς, δεν υπάρχουν μαθηματικά θεμελιωμένοι κανόνες γι' αυτά τα ζητήματα. Αξιοποιώντας κυρίως την εμπειρία, έχουν προταθεί ορισμένοι πρακτικοί κανόνες, όπως το πλήθος των κρυφών νευρώνων να είναι το μισό του πλήθους των νευρώνων εισόδου ή να είναι το μισό του αθροίσματος των νευρώνων εισόδου και των δυνατών τιμών της κλάσης. Φυσικά, τέτοιοι πρακτικοί κανόνες δεν έχουν απόλυτη ισχύ και συχνά ο χρήστης είναι υποχρεωμένος να πειραματίζεται με διάφορα μοντέλα, μέχρι να επιλέξει μια αρχιτεκτονική. Επιπλέον, ο χρήστης πρέπει να ρυθμίσει μια σειρά από πρόσθετες παραμέτρους. Ειδικότερα πρέπει να επιλέξει τις συναρτήσεις μετασχηματισμού για κάθε επίπεδο και να ορίσει τιμές για τον ρυθμό εκπαίδευσης, το πλήθος των εποχών και τη ροπή. Οι παράμετροι έχουν επιπτώσεις στην εκπαίδευση του μοντέλου. Μικρός ρυθμός εκπαίδευσης προκαλεί μικρές μεταβολές βαρών και κατ' επέκταση αργή εκπαίδευση του δικτύου. Μεγάλος ρυθμός εκπαίδευσης προκαλεί την ταχεία εκπαίδευση του δικτύου και κίνδυνο υπερπροσαρμογής του μοντέλου. Για τη ρύθμιση όλων αυτών των παραμέτρων δεν υπάρχουν καθορισμένοι κανόνες και οδηγός του χρήστη είναι η εμπειρία.

Άλλα σημαντικά ζητήματα αφορούν τα δεδομένα. Τα νευρωνικά δίκτυα είναι μια μέθοδος ισχυρά καθοδηγούμενη από τα δεδομένα, Για τον λόγο αυτό τα δεδομένα είναι ιδιαίτερα σημαντικά. Στο στάδιο της προεπεξεργασίας πρέπει να έχουν επιλεγεί τα σημαντικά χαρακτηριστικά τα οποία και θα γίνουν οι νευρώνες εισόδου του δικτύου. Τα νευρωνικά δίκτυα λειτουργούν καλύτερα με κανονικοποιημένες τιμές. Ορισμένες υλοποιήσεις επιτρέπουν την αυτοματοποιημένη κανονικοποίηση των τιμών. Στις άλλες περιπτώσεις όμως, ο χρήστης πρέπει να κανονικοποιήσει μόνος του τις τιμές στο στάδιο της προεπεξεργασίας. Το σύνολο δεδομένων του νευρωνικού δικτύου χωρίζεται σε σύνολο εκπαίδευσης (training set) και σύνολο ελέγχου (test set). Το σύνολο εκπαίδευσης χρησιμοποιείται για τη ρύθμιση των βαρών των συνδέσεων. Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος δοκιμάζει το μοντέλο με το σύνολο ελέγχου και διακόπτει την εκπαίδευση εάν θεωρήσει ότι το μοντέλο εκπαιδεύτηκε επαρκώς. Ένα τρίτο σύνολο παρατηρήσεων μπορεί να χρησιμοποιηθεί για τον τελικό έλεγχο του μοντέλου αφού ολοκληρωθεί η εκπαίδευση. Επιπλέον, τα Νευρωνικά Δίκτυα είναι αρκετά σύνθετα μοντέλα και μπορούν να ενσωματώσουν σημαντικό όγκο πληροφορίας. Για τους λόγους αυτούς απαιτείται σημαντικός αριθμός παρατηρήσεων για την επιτυχημένη εκπαίδευση του μοντέλου.

9.8.5 Πλεονεκτήματα και μειονεκτήματα των Νευρωνικών Δικτύων

Τα Νευρωνικά Δίκτυα οφείλουν τη μεγάλη δημοφιλία τους στα αδιαμφισβήτητα **πλεονεκτήματα** τους:

- Τα Νευρωνικά Δίκτυα είναι ιδιαιτέρως κατάλληλα αν δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών εισόδου και εξόδου. Η ύπαρξη των κρυφών στρωμάτων επιτρέπει την ικανοποιητική προσέγγιση σύνθετων συναρτήσεων.
- Είναι ιδιαιτέρως ικανά να κατηγοριοποιήσουν αντικείμενα που δεν περιλαμβάνονταν στο σύνολο εκπαίδευσης και επομένως είναι άγνωστα στο δίκτυο.
- Μπορούν να χειριστούν θορυβώδη και ασυνεπή δεδομένα.

Πέρα από τα πλεονεκτήματά τους, τα Νευρωνικά Δίκτυα δεν στερούνται **μειονεκτημάτων**:

- Το σημαντικότερο ίσως μειονέκτημα τους είναι ότι απαιτείται ο εμπειρικός προσδιορισμός πολλών

παραμέτρων όπως η τοπολογία του δικτύου, ο αριθμός των εποχών εκπαίδευσης, ο καθορισμός του ρυθμού εκπαίδευσης. Για όλες αυτές τις παραμέτρους δεν υπάρχει καθιερωμένη δεοντολογία για τον προσδιορισμό τους.

- Άλλο σημαντικό μειονέκτημα είναι η προβληματική ερμηνευσιμότητα. Ο τρόπος λήψης αποφάσεων των νευρωνικών δικτύων είναι ακατανόητος στους ανθρώπους. Ιδιαίτερα στα χρηματοοικονομικά, ο χρήστης επιθυμεί να διασφαλίζει ότι ο τρόπος λήψης αποφάσεων συνάδει ή έστω δεν αντικρούει με καθιερωμένη γνώση. Επίσης, γενικότερα, ο σκοπός της Εξόρυξης Δεδομένων είναι η ανακάλυψη γνώσης, όχι προβλέψεων.
- Τα Νευρωνικά Δίκτυα απαιτούν μεγάλους χρόνους εκπαίδευσης.

9.9 Μπαϋεσιανοί Κατηγοριοποιητές

Τα Μπαϋεσιανά Δίκτυα (Bayesian Networks) είναι ισχυρά εργαλεία για αναπαράσταση σύνθετων σχέσεων μεταξύ μεταβλητών και για εξαγωγή συμπερασμάτων σε συνθήκες αβεβαιότητας. Ανήκουν στην κατηγορία των γραφικών πιθανοτικών μοντέλων, τα οποία αναπαριστούν σχέσεις με μορφή γράφων. Κάθε κόμβος του γράφου συμβολίζει μια στοχαστική μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης ανάμεσα σε δύο μεταβλητές. Τα Μπαϋεσιανά Δίκτυα αρχικά δεν θεωρήθηκαν εργαλεία κατηγοριοποίησης, αργότερα όμως ανακαλύφθηκε ότι οι Αφελείς Μπαϋεσιανοί κατηγοριοποιητές (Naive Bayesian Classifiers), μια απλουστευμένη εκδοχή των Μπαϋεσιανών Δικτύων, έχουν αυξημένες δυνατότητες κατηγοριοποίησης, συγκρίσιμες με αυτές των Νευρωνικών Δικτύων και των Δένδρων Αποφάσεων. Σήμερα τα Μπαϋεσιανά Δίκτυα αποτελούν μια καταξιωμένη μέθοδο Εξόρυξης Δεδομένων, λόγω της στιβαρής θεωρητικής τους θεμελίωσης, της ικανότητας τους να καταγράφουν περίπλοκες σχέσεις αλληλεξάρτησης, του συμβολικού φορμαλισμού τους και της δυνατότητας τους να εφαρμόζονται σε προβλήματα κατηγοριοποίησης (Heckerman, 1997).

Τα Μπαϋεσιανά Δίκτυα έλκουν το θεωρητικό τους υπόβαθρο από τη στατιστική και πιο συγκεκριμένα από το θεώρημα του Bayes, που υπολογίζει την υπό συνθήκη πιθανότητα $P(H|X)$, δηλαδή την πιθανότητα να επαληθευτεί η υπόθεση H με δεδομένο ότι ισχύει το γεγονός X. Σύμφωνα με το θεώρημα του Bayes, η πιθανότητα $P(H|X)$ δίνεται από την Εξίσωση 9.12

$$P(H|X) = \frac{P(H) * P(X|H)}{P(X)} \tag{9.12}$$

όπου $P(H)$ είναι η εκ των προτέρων πιθανότητα να ισχύει η υπόθεση H, $P(X)$ είναι η εκ των προτέρων πιθανότητα να συμβεί το γεγονός X και $P(X|H)$ είναι η πιθανότητα να συμβεί το γεγονός X με δεδομένο ότι ισχύει η υπόθεση H.

9.9.1 Αφελείς Μπαϋεσιανοί Κατηγοριοποιητές

Ο Αφελής Μπαϋεσιανός κατηγοριοποιητής αποτελεί ευθεία εφαρμογή του θεωρήματος Bayes. Υποθέτουμε ότι X είναι μια παρατήρηση του συνόλου δεδομένων και H είναι η υπόθεση ότι παρατήρηση αυτή ανήκει στην κλάση C_i. Πιο συγκεκριμένα, το X θεωρείται ως ένα άνυσμα n τιμών $X=(x_1,x_2,\ldots,x_n)$. Υποθέτουμε ότι υπάρχουν m κλάσεις $C_1, C_2, \ldots, C_m$. Σύμφωνα με το θεώρημα του Bayes, η πιθανότητα να ανήκει η παρατήρηση X στην κλάση C_i υπολογίζεται από την Εξίσωση 9.13.

$$P(C_i|X) = \frac{P(X|C_i) * P(C_i)}{P(X)} \tag{9.13}$$

Για να προβλέψει την κλάση μιας άγνωστης παρατήρησης, ο Αφελής Μπαϋεσιανός κατηγοριοποιητής υπολογίζει τις πιθανότητες για την κάθε κλάση και εκχωρεί την παρατήρηση στην κλάση με τη μεγαλύτερη πιθανότητα. Εφόσον το $P(X)$ είναι ίδιο για όλες τις κλάσεις και το $P(C_i)$ μπορεί εύκολα να υπολογιστεί (ως το πλήθος των παρατηρήσεων που ανήκουν στην κλάση C_i προς το πλήθος όλων των παρατηρήσεων), το ζητού-

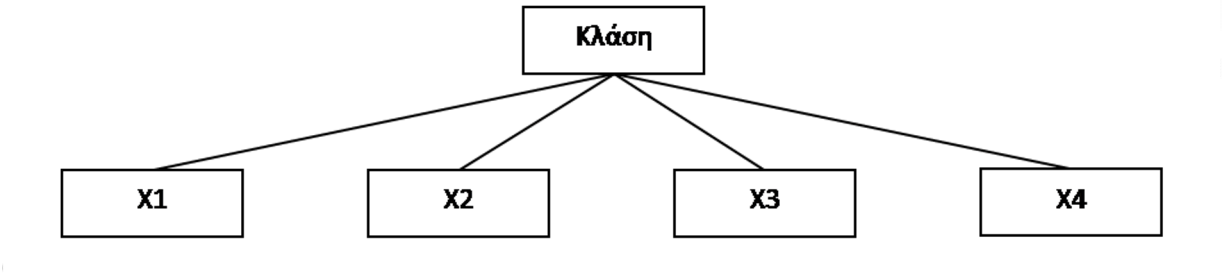
μενο είναι ο υπολογισμός του $P(X|C_i)$. Ο υπολογισμός του $P(X|C_i)$ μπορεί να αποδειχθεί ιδιαίτερα περίπλοκος εάν θεωρηθεί ότι υπάρχει σχέση εξάρτησης μεταξύ των διαστάσεων του ανύσματος Χ, δηλαδή μεταξύ των μεταβλητών εισόδου. Αντιθέτως, αν θεωρηθεί ότι, δοθείσης της κλάσης, οι μεταβλητές εισόδου είναι μεταξύ τους ανεξάρτητες, τότε ο υπολογισμός του $P(X|C_i)$ απλοποιείται και δίνεται από την Εξίσωση 9.14

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

(9.14)

όπου x_k είναι η τιμή της διάστασης k του ανύσματος X.

Ένας Αφελής Μπαϋεσιανός κατηγοριοποιητής με τέσσερεις ανεξάρτητες μεταβλητές παρουσιάζεται με μορφή γραφικού πιθανοτικού μοντέλου στο Σχήμα 9.8.



Σχήμα 9.8 *Αφελής Μπαϋεσιανός Κατηγοριοποιητής*

Ο κατηγοριοποιητής, αφού υπολογίσει τις πιθανότητες $P(C_i|X)$ για όλες τις κλάσεις C_i, εκχωρεί την παρατήρηση στην κλάση με τη μεγαλύτερη πιθανότητα. Εάν ισχύει η υπόθεση ότι δεδομένης της κλάσης είναι ανεξάρτητες οι μεταβλητές εισόδου, ο Αφελής Μπαϋεσιανός κατηγοριοποιητής επιτυγχάνει τους υψηλότερους ρυθμούς ακρίβειας. Ωστόσο, στην πράξη τις περισσότερες φορές η υπόθεση αυτή δεν ισχύει.

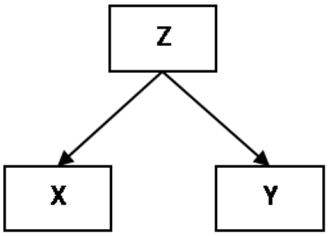
9.9.2 Μπαϋεσιανά Δίκτυα

Τα Μπαϋεσιανά Δίκτυα αποτελούν επέκταση των Αφελών Μπαϋεσιανών κατηγοριοποιητής (ΑΜΚ). Ωστόσο, σε αντίθεση με τους ΑΜΚ δεν υποθέτουν την ανεξαρτησία των μεταβλητών εισόδου. Αντιθέτως, τα Μπαϋεσιανά Δίκτυα επιτρέπουν την ανεξαρτησία υποσυνόλων των μεταβλητών εισόδου. Ένα Μπαϋεσιανό Δίκτυο αναπαριστά τις εξαρτήσεις μεταξύ των μεταβλητών με τη χρήση ενός **Κατευθυνόμενου Ακυκλικού Γράφου** (ΚΑΓ) (Directed Acyclic Graph (DAG)). Κάθε κόμβος του γράφου συμβολίζει μια μεταβλητή και κάθε βέλος συμβολίζει μια σχέση εξάρτησης. Ένα βέλος, το οποίο κατευθύνεται από τη μεταβλητή X προς τη μεταβλητή Y, δηλώνει ότι η Y εξαρτάται από τη X. Η μεταβλητή X καλείται γονέας της Y και η Y καλείται τέκνο της X. Στα Μπαϋεσιανά Δίκτυα μια σημαντική έννοια είναι αυτή της **υπό συνθήκη ανεξαρτησίας** (conditional independence) δύο μεταβλητών. Θεωρούμε τρεις μεταβλητές X,Y,Z οι οποίες συγκροτούν ένα Μπαϋεσιανό Δίκτυο, όπως απεικονίζεται στο Σχήμα 9.9.

Οι μεταβλητές X και Y είναι υπό συνθήκη ανεξάρτητες, εάν οι τιμές της X, με δεδομένες τις τιμές των Y και Z, εξαρτώνται μόνο από τις τιμές της Z. Με μαθηματικό τρόπο η ιδιότητα αυτή αποδίδεται από την Εξίσωση 9.15.

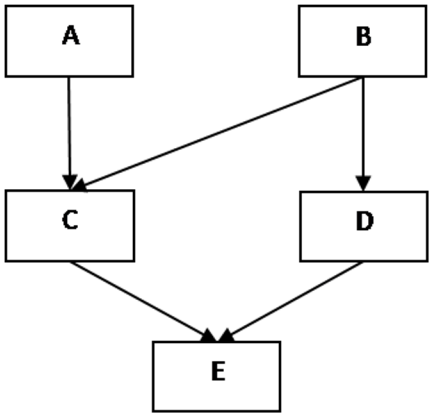
$$P(X|Z,Y) = P(X|Z)$$

(9.15)



Σχήμα 9.9 *Υπό συνθήκη ανεξαρτησία μεταβλητών*

Ο Αφελής Μπαϋεσιανός Κατηγοριοποιητής υποθέτει την υπό συνθήκη ανεξαρτησία των μεταβλητών εισόδου, όταν είναι δεδομένη η τιμή της κλάσης. Στα Μπαϋεσιανά Δίκτυα ισχύει η **τοπική ιδιότητα Markov**, σύμφωνα με την οποία κάθε μεταβλητή είναι υπό συνθήκη ανεξάρτητη από τους μη απογόνους της όταν είναι δεδομένοι οι γονείς της. Το Σχήμα 9.10 απεικονίζει ένα Μπαϋεσιανό Δίκτυο με πέντε μεταβλητές. Η μεταβλητή C είναι ανεξάρτητη από την D εάν είναι γνωστές οι μεταβλητές A και B. Αυτό σημαίνει ότι εάν οι τιμές των μεταβλητών A και B είναι γνωστές, τότε η μεταβλητή D δεν προφέρει πρόσθετη πληροφορία σχετικά με τη μεταβλητή C. Οι μεταβλητές μπορούν να παίρνουν τιμές διακριτές ή συνεχόμενες.



Σχήμα 9.10 *Μπαϋεσιανό Δίκτυο*

Ο γράφος των Μπαϋεσιανών Δικτύων καταγράφει τις σχέσεις μεταξύ των μεταβλητών. Οι σχέσεις αυτές ποσοτικοποιούνται με τον Πίνακα Υπό Συνθήκη Πιθανοτήτων (Conditional Probability Table (CPT)). Στον πίνακα CPT καταγράφεται για κάθε μεταβλητή X η κατανομή πιθανοτήτων $P(X|Par(X))$, όπου $Par(X)$ οι γονείς της μεταβλητής X. Αν τα δεδομένα περιέχουν n μεταβλητές $X_1,X_2,\ldots,X_n$, τότε η πιθανότητα εμφάνισης μιας παρατήρησης με τιμές $x_1,x_2,\ldots,x_n$ για τις αντίστοιχες μεταβλητές δίνεται από την Εξίσωση 9.16

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P\big(x_i | Par(X_i)\big) \tag{9.16}$$

Ένα Μπαϋεσιανό Δίκτυο μπορεί να χρησιμοποιηθεί ως κατηγοριοποιητής. Ένας από τους κόμβους αντιπροσωπεύει τη μεταβλητή της κλάσης. Για μια παρατήρηση υπολογίζονται οι πιθανότητες για κάθε δυνατή τιμή της κλάσης, και η παρατήρηση εκχωρείται στην πιο πιθανή κλάση. Έχουν προταθεί διάφοροι κατηγοριοποιητές Μπαϋεσιανών Δικτύων. Ένας από τους πιο γνωστούς είναι ο Tree Augmented Naïve Bayes (Friedman, Geiger & Goldszmidt, 1997), ο οποίος προβλέπει ότι μια μεταβλητή έχει οπωσδήποτε γονέα της τη μεταβλητή της κλάσης και πιθανώς να έχει γονέα της μια επιπλέον μεταβλητή.

Η δημιουργία ενός μοντέλου Μπαϋεσιανού Δικτύου περιλαμβάνει δύο εργασίες:

- την κατασκευή του γράφου,
- τον υπολογισμό του πίνακα πιθανοτήτων CPT.

Ο υπολογισμός του CPT είναι ευκολότερο καθήκον, ειδικά εάν δεν υπάρχουν χαμένες τιμές. Εάν δεν υπάρχουν κρυφά δεδομένα, ο υπολογισμός του CPT είναι απλός και γίνεται με τρόπο αντίστοιχο με τον υπολογισμό των πιθανοτήτων στον Αφελή Μπαϋεσιανό κατηγοριοποιητή. Η ύπαρξη χαμένων τιμών περιπλέκει τον υπολογισμό του CPT. Για την κατασκευή του γράφου υπάρχουν δύο εκδοχές. Κατά την πρώτη εκδοχή ο γράφος σχεδιάζεται από ανθρώπους, οι οποίοι είναι ειδικοί στο πρόβλημα το οποίο εξετάζεται. Η δεύτερη εκδοχή είναι να εξαχθεί ο γράφος από τα δεδομένα με αυτοματοποιημένο τρόπο. Η αυτόματη δημιουργία του γράφου είναι ένα δύσκολο καθήκον. Σε επιστημονικές εργασίες έχουν προταθεί διάφορες μέθοδοι για την εξαγωγή του γράφου. Ενδεικτικά αναφέρουμε τις Heckerman, Geiger and Chickering (1995), Cooper and Herskovits (1992) και Cheng, Greiner, Kelly, Bell and Liu (2002).

9.9.3 Πλεονεκτήματα και μειονεκτήματα των Μπαϋεσιανών Δικτύων

Τα Μπαϋεσιανά Δίκτυα Πίστης συγκεντρώνουν πολλά **πλεονεκτήματα**:

- Δημιουργούν ένα μοντέλο για την κατανομή πιθανοτήτων για ένα πρόβλημα. Υπό αυτήν την έννοια είναι ιδιαίτερα κατάλληλα για περιπτώσεις όπου υπάρχουν σύνθετες εξαρτήσεις μεταξύ της μεταβλητής της κλάσης και των μεταβλητών εισόδου ή και ακόμα μεταξύ των μεταβλητών εισόδου.
- Ο γράφος που δημιουργείται οπτικοποιεί τις σχέσεις μεταξύ της κλάσης και των μεταβλητών εισόδου. Για τον λόγο αυτό, τα Μπαϋεσιανά Δίκτυα Πίστης είναι εύκολα κατανοητά από τους ανθρώπους.
- Για τον σχεδιασμό του γράφου μπορεί να χρησιμοποιηθεί προηγούμενη γνώση ειδικών. Ακόμα και εάν ο γράφος εξαχθεί από τα δεδομένα με αυτοματοποιημένο τρόπο, υπάρχει δυνατότητα μεταβολής του από τους ειδικούς. Στην περίπτωση αυτή επιτρέπεται ο συγκερασμός της γνώσης ειδικών με το αποτέλεσμα του αλγορίθμου εξαγωγής του γράφου.
- Μπορούν να χειριστούν και αριθμητικές και ονομαστικές μεταβλητές.
- Μπορούν να επιτύχουν υψηλούς ρυθμούς ακρίβειας.
- Διαθέτουν στιβαρή θεωρητική θεμελίωση βασισμένη στη Στατιστική.

Μειονεκτήματα των Μπαϋεσιανών Δικτύων είναι τα ακόλουθα:

- Το σημαντικότερο μειονέκτημα τους είναι το γεγονός ότι δεν υπάρχει ένας καθιερωμένος και γενικά αποδεκτός τρόπος εξαγωγής του γράφου από τα δεδομένα.
- Για τον υπολογισμό των πιθανοτήτων ενός κλάδου του δικτύου απαιτείται ο υπολογισμός όλων των άλλων κλάδων επιφέροντας σημαντικό υπολογιστικό κόστος.
- Στους Αφελείς Μπαϋεσιανούς Κατηγοριοποιητές η υπόθεση ανεξαρτησίας των μεταβλητών εισόδου ισχύει σπάνια.

9.10 Μελέτη περίπτωσης – Εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων με χρήση μεθόδων κατηγοριοποίησης.

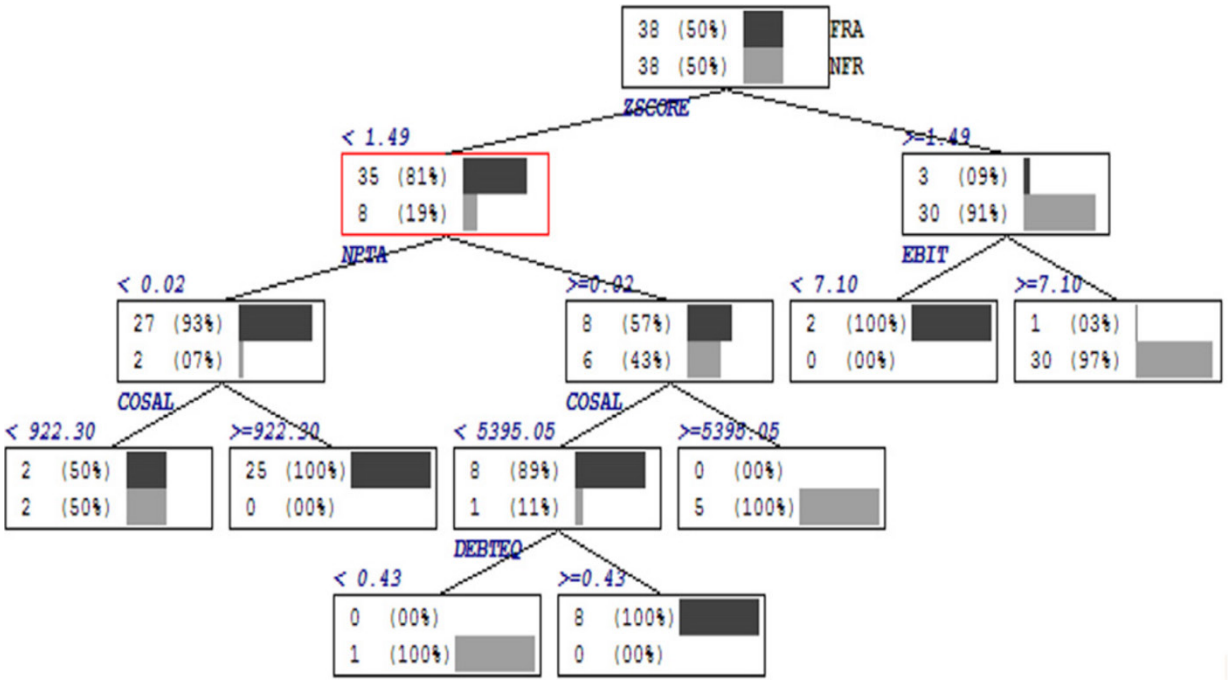
Η παραποίηση των χρηματοοικονομικών καταστάσεων από τη διοίκηση των επιχειρήσεων είναι ένα σημαντικό πρόβλημα της παγκόσμιας οικονομίας. Ο Wells (1997) εκτιμά ότι η απάτη κοστίζει στην αμερικανική οικονομία 400 δισεκατομμύρια δολάρια ετησίως, ενώ ο Koskivaara (2004) αποκαλεί το έτος 2002 «φριχτή χρονιά» ως προς την τήρηση των βιβλίων και ισχυρίζεται ότι η χειραγώγηση συνεχίζεται. Οι Spathis, Doumpos and Zopounidis (2002) επισημαίνουν ότι οι παραποιημένες χρηματοοικονομικές καταστάσεις αυξάνονται τα τελευταία χρόνια. Ο εντοπισμός των περιπτώσεων διοικητικής απάτης είναι ιδιαίτερα δύσκολος, καθώς τα έμπειρα διοικητικά στελέχη γνωρίζουν τα όρια των τυπικών ελεγκτικών διαδικασιών και ηθελημένα προσπαθούν να παραπλανήσουν τους ελεγκτές. Αυτοί οι περιορισμοί συνιστούν την εφαρμογή νέων, περίτεχνων αναλυτικών διαδικασιών. Οι Kirkos, Spathis and Manolopoulos (2007) διερευνούν τη δυνατότητα των μεθόδων εξόρυξης δεδομένων, και ειδικότερα των μεθόδων κατηγοριοποίησης, να εντοπίσουν περιπτώσεις παραποιη-

μένων χρηματοοικονομικών καταστάσεων.

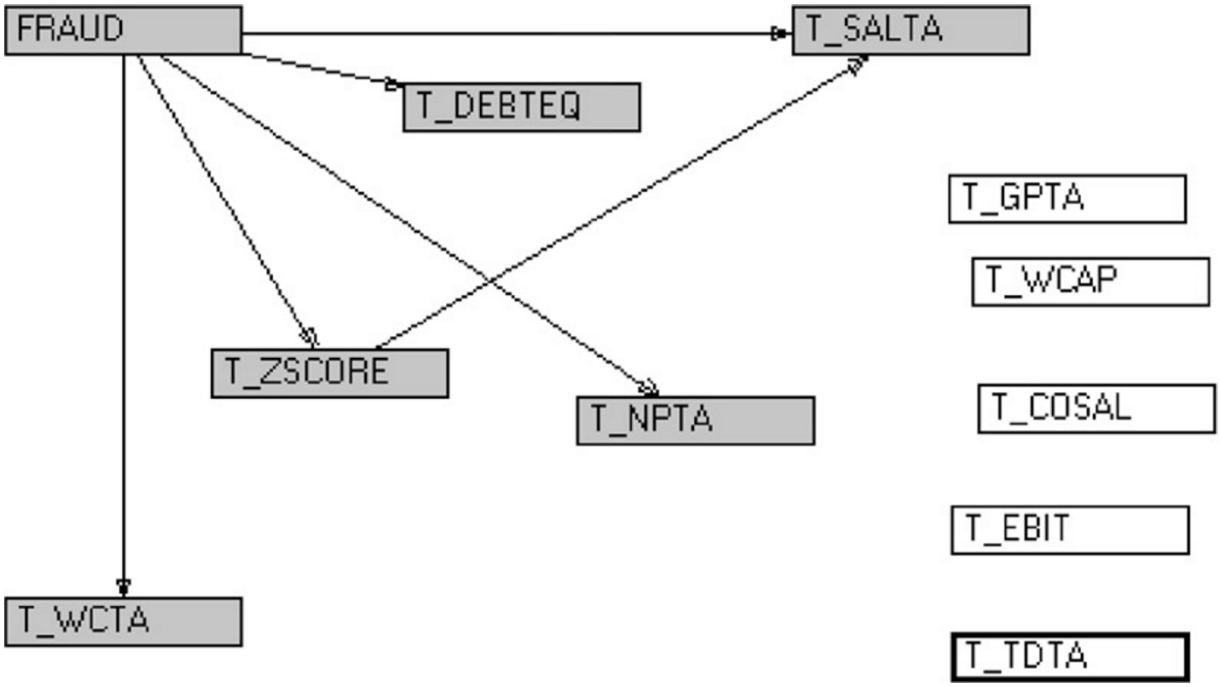
Τα δεδομένα αφορούν 76 περιπτώσεις ελληνικών, μη χρηματοπιστωτικών επιχειρήσεων. Οι μισές από αυτές τις επιχειρήσεις είχαν αποδεδειγμένα παραποιήσει τις χρηματοοικονομικές τους καταστάσεις. Το γεγονός αυτό προκύπτει από παρατηρήσεις των εξωτερικών ελεγκτών, από αποφάσεις δικαστηρίων, από αποφάσεις της Επιτροπής Κεφαλαιαγοράς του ΧΑΑ, καθώς και από αποφάσεις φορολογικών αρχών. Για τις υπόλοιπες 38 επιχειρήσεις δεν υπήρχε καμία απόδειξη ή ένδειξη παραποίησης, και για τον λόγο αυτό οι χρηματοοικονομικές τους καταστάσεις θεωρήθηκαν ειλικρινείς και σύννομες. Ως μεταβλητή κλάσης χρησιμοποιήθηκε μια δυαδική μεταβλητή που διαφοροποιούσε τις δύο κατηγορίες επιχειρήσεων. Σημειωτέον ότι τα δεδομένα αφορούν δημοσίως διαθέσιμα στοιχεία, τα οποία δημοσιεύτηκαν σε ισολογισμούς και σε αποτελέσματα χρήσης.

Η αρχική επιλογή μεταβλητών βασίστηκε στην προηγούμενη ερευνητική βιβλιογραφία. Μεγάλος αριθμός ερευνητικών εργασιών ασχολήθηκε με το θέμα της παραποίησης των χρηματοοικονομικών καταστάσεων. Στις εργασίες αυτές προτείνονται διάφοροι αριθμοδείκτες, οι οποίοι παρέχουν ενδείξεις παραποίησης και μπορούν να χρησιμοποιηθούν ως μεταβλητές εισόδου σε ένα μοντέλο πρόβλεψης. Ενδεικτικά αναφέρουμε τις εργασίες των Fanning and Cogger (1998), Loebbecke, Eining and Willingham (1989) και Persons (1995). Συνολικά επιλέχτηκαν 27 αρχικοί αριθμοδείκτες. Σε αυτές περιλαμβάνονταν και ο αριθμοδείκτης Z-Score του Altman. Ακολούθησε ανάλυση επιλογής σημαντικών χαρακτηριστικών με εφαρμογή της μεθόδου ANOVA. Σύμφωνα με τα αποτελέσματα, 10 αριθμοδείκτες παρουσίασαν χαμηλή τιμή *p (p<=0,05)*. Οι αριθμοδείκτες αυτοί επιλέχτηκαν, ώστε να αποτελέσουν τις μεταβλητές εισόδου στα μοντέλα που αναπτύχθηκαν.

Τρεις διαφορετικές μέθοδοι κατηγοριοποίησης εφαρμόστηκαν για την ανάπτυξη αντίστοιχων μοντέλων. Το πρώτο μοντέλο ήταν ένα Δένδρο Αποφάσεων τύπου ID3. Το δεύτερο μοντέλο ήταν ένα Νευρωνικό Δίκτυο τύπου Multilayer Perceptron με ένα κρυφό στρώμα, το οποίο περιείχε πέντε νευρώνες. Το τρίτο μοντέλο ήταν ένα Μπαϋεσιανό Δίκτυο. Το λογισμικό που χρησιμοποιήθηκε για το Μπαϋεσιανό Δίκτυο ήταν ικανό να εξάγει τον γράφο από τα δεδομένα. Το λογισμικό επέτρεπε στον χρήστη να τροποποιήσει τη δομή του γράφου ο οποίος κατασκευάστηκε αυτόματα, ωστόσο επιλέξαμε να χρησιμοποιήσουμε τον αυτοματοποιημένο γράφο χωρίς μεταβολές. Τα μοντέλα του Δένδρου Αποφάσεων και του Μπαϋεσιανού Δικτύου παρουσιάζονται στα σχήματα 9.11 και 9.12 αντίστοιχα.



Σχήμα 9.11 *Εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων με Δένδρο Αποφάσεων*



Σχήμα 9.12 *Εντοπισμός παραποιημένων χρηματοοικονομικών καταστάσεων με Μπαϋεσιανό Δίκτυο.*

Παρατηρούμε ότι το Δένδρο Αποφάσεων χρησιμοποιεί ως μεταβλητή διαχωρισμού πρώτου επιπέδου τη μεταβλητή Z-Score. Σύμφωνα με το κριτήριο μείωσης της εντροπίας, ο δείκτης Z-Score είναι αυτός που διαχωρίζει με τον καλύτερο τρόπο τις δύο κλάσεις. Σαράντα τρεις επιχειρήσεις είχαν τιμή Z-Score μικρότερη από 1,49. Οι τριάντα πέντε από αυτές τις επιχειρήσεις είχαν παραποιήσει τις χρηματοοικονομικές τους καταστάσεις. Αντιθέτως, από τις τριάντα τρεις επιχειρήσεις με δείκτη Z-Score μεγαλύτερο από ή ίσο με 1,49 μόνο οι τρεις είχαν παραποιήσει τις χρηματοοικονομικές τους καταστάσεις. Υπενθυμίζουμε ότι ο Altman είχε ορίσει την τιμή Z-Score=1,81 ως τιμή διαχωρισμού μεταξύ των υγειών και προβληματικών επιχειρήσεων για την αμερικανική βιομηχανία. Με βάση αυτό το γεγονός, μπορούμε να συμπεράνουμε ότι επιχειρήσεις σε οικονομική δυσπραγία τείνουν να χειραγωγήσουν τις χρηματοοικονομικές τους καταστάσεις. Μια άλλη ενδιαφέρουσα παρατήρηση είναι ότι και οι δύο αριθμοδείκτες που χρησιμοποιούνται ως μεταβλητές διαχωρισμού δεύτερου επιπέδου σχετίζονται με την κερδοφορία. Η μεταβλητή NPTA είναι τα καθαρά κέρδη προς σύνολο ενεργητικού (Net Profit to Total Assets) και η μεταβλητή EBIT είναι τα κέρδη προ φόρων και τόκων (Earnings Before Interest and Tax). Από τους δύο κλάδους του κόμβου Z-Score και τις μεταβλητές διαχωρισμού δευτέρου επιπέδου προκύπτει ότι επιχειρήσεις εμπλεκόμενες σε πρακτικές χειραγώγησης και με χαμηλό Z-Score παρουσιάζουν χαμηλή κερδοφορία. Αντιθέτως, η μη χειραγώγηση σχετίζεται κυρίως με επιχειρήσεις που έχουν υψηλό Z-Score και ικανοποιητική κερδοφορία.

Σύμφωνα με το μοντέλο του Μπαϋεσιανού Δικτύου υπάρχει σχέση εξάρτησης μεταξύ της μεταβλητής της κλάσης και πέντε αριθμοδεικτών. Είναι ενδιαφέρον ότι κάθε μια από αυτές τις μεταβλητές αναφέρεται σε μια διαφορετική διάσταση των οικονομικών στοιχείων της επιχείρησης. Ειδικότερα, η μεταβλητή Z-Score αναφέρεται στην οικονομική ευρωστία, η μεταβλητή DEBTEQ (Debt to Equity) στη μόχλευση, η μεταβλητή NPTA (Net Profit to Total Assets) στην κερδοφορία, η μεταβλητή SALTA (Sales to Total Assets) στο ύψος των πωλήσεων και η μεταβλητή WCTA (Working Capital to Total Assets) στη ρευστότητα. Φαίνεται ότι το Μπαϋεσιανό Δίκτυο οικοδομεί μια πιο γενικευμένη άποψη και συσχετίζει τη χειραγώγηση των χρηματοοικονομικών καταστάσεων με διάφορες εκφάνσεις της οικονομικής κατάστασης της επιχείρησης.

Για την εκτίμηση της ρεαλιστικής επίδοσης των τριών μεθόδων, δηλαδή της ικανότητας τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων, εφαρμόστηκε η τεχνική της επικύρωσης 10 τμημάτων (10 fold cross validation), η οποία παρουσιάζεται αναλυτικά στο Κεφάλαιο 10. Η συγκεκριμένη τεχνική θεωρείται ιδιαίτερα αξιόπιστη και κατάλληλη για σχετικά μικρά σύνολα παρατηρήσεων. Σύμφωνα με τα αποτελέσματα, το Μπαϋεσιανό Δίκτυο είχε γενική επίδοση 90,3% και επέτυχε να κατηγοριοποιήσει σωστά το 91,7% των περιπτώσεων απάτης και το 88,9% των περιπτώσεων μη απάτης. Οι αντίστοιχες επιδόσεις για το Νευρωνικό Δίκτυο ήταν 80%, 82,5% και 77,5%, ενώ για το Δένδρο Αποφάσεων ήταν 73,6%, 75% και 72,5%. Παρατηρούμε ότι το Μπαϋεσιανό Δίκτυο επέτυχε εξαιρετικά υψηλές επιδόσεις, ακολουθούμενο από το Νευρωνικό Δίκτυο και το Δένδρο Αποφάσεων. Παρατηρούμε επίσης ότι και τα τρία μοντέλα προβλέπουν σε μεγαλύτερο

ποσοστό τις περιπτώσεις απάτης από τις περιπτώσεις μη απάτης. Το γεγονός αυτό είναι σημαντικό. Η σημασία των διαφορετικών τύπων σφάλματος παρουσιάζεται αναλυτικά στο Κεφάλαιο 10.

Βιβλιογραφία / Αναφορές

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC Press.

Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning Bayesian Networks from Data: An Information-Theory Based Approach. *Artificial Intelligence, 137*(1-2), 43-90. doi: 10.1016/s0004-3702(02)00191-1

Cooper, G., & Herskovits, E. (1992). A Bayesian Method for the Induction Probabilistic Networks from Data. *Machine Learning, 9*(4), 309-347. doi: 10.1007/bf00994110

Fanning, K., & Cogger, K. (1998). Neural Network Detection of Management Fraud Using Published Financial Data. *International Journal of Intelligent Systems in Accounting, Finance and Management, 7*(1), 21-41. doi: 10.1002/(sici)1099-1174(199803)7:1<21::aid-isaf138>3.0.co;2-k

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In: U. Fayyad, G. Piatetsky-Shapiro & P. Smyth (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1-34). Menlo Park, CA: AAAI/MIT Press.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning, 29*(2-3), 131-163. doi: 10.1023/A:1007465528199

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Boston: Academic Press.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Heckerman, D. (1997). Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery, 1*(1), 79-119. doi: 10.1023/A:1009730122752

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning, 20*(3), 197-243. doi: 10.1007/bf00994016

Hwang, J., Lay, S., & Lippman, A. (1994). Nonparametric Multivariate Density Estimation: A Comparative Study. *IEEE Transactions on Signal Processing, 42*(10), 2795-2810. doi: 10.1109/78.324744

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics, 29*(2), 119-127. doi: 10.2307/2986296

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications, 32*(4), 995-1003. doi: 10.1016/j.eswa.2006.02.016

Koskivaara, E. (2004). Artificial Neural Networks in Analytical Review Procedures. *Managerial Auditing Journal, 19*(2), 191–223. doi: 10.1108/02686900410517821

Loebbecke, J., Eining, M., & Willingham, J. (1989). Auditor's Experience with Material Irregularities: Frequency, Nature and Detectability. *Auditing: A Journal of Practice and Theory, 9*, 1-28.

Loh, W. Y., & Shih, X. (1997). Split Selection Methods for Classification Trees. *Statistica Sinika, 7*, 815-840.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer + Business Media.

Olaru, C., & Wehenkel, L. (2003). A Complete Fuzzy Decision Tree Technique. *Fuzzy Sets and Systems, 138*(2), 221-254. doi: 10.1016/S0165-0114(03)00089-7

Parker, D. B. (1985). Learning-logic: Casting the Cortex of the Human Brain in Silicon. *Technical Report TR-47*. Boston, MA: Center for Computational Research in Economics and Management Science, MIT.

Persons, O. (1995). Using Financial Statements Data to Identify Factors Associated with Fraudulent Financial Reporting. *Journal of Applied Business Research, 11*(3), 38-46.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*(1), 81-106. doi: 10.1007/bf00116251

Quinlan, J. R. (1987). Simplifying Decision Trees. *International Journal of Man-Machine Studies, 27*(3), 221-234. doi: 10.1016/s0020-7373(87)80053-6

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufman.

Quinlan, J. R., & Rivest, R. L. (1989). Inferring Decision Trees Using the Minimum Description Length Principle. *Information and Computation, 80*(3), 227-248. doi: 10.1016/0890-5401(89)90010-2

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-propagating Errors. *Letters to Nature, 323*(6088), 533-536. doi: 10.1038/323533a0

Simonoff, J. S. (2003). *Analyzing Categorical Data*. New York, NY: Springer-Verlag.
Sonquist, J. A., Baker, E. L., & Morgan, J. N. (1971). *Searching for Structure*. An Arbor, MI: Institute for Social Research, University of Michigan.
Spathis, C., Doumpos, M., & Zopounidis, C. (2002). Detecting Falsified Financial Statements: A Comparative Study Using Multicriteria Analysis and Multivariate Statistical Techniques. *The European Accounting Review, 11*(3), 509-535. doi: 10.1080/0963818022000000966
Wells, J. T. (1997). *Occupational Fraud and Abuse*. Austin, TX: Obsidian Publishing.
Wong, B. K., Bodnovich, T. A., & Selvi, Y. (1997). Neural Networks Applications in Business: A Review and Analysis of the Literature (1988-1995). *Decision Support Systems, 19(4), 301-320.* doi: 10.1016/s0167-9236(96)00070-x

Κριτήρια Αξιολόγησης

Άσκηση Υπολογισμών 9.1

Χρησιμοποιήστε τα δεδομένα του Πίνακα 9.1 και υπολογίστε το Κέρδος Πληροφορίας σε περίπτωση που επιλεγεί το πεδίο «Εισόδημα» ως μεταβλητή διαχωρισμού για τη δημιουργία Δένδρου Αποφάσεων ID3.

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

Πίνακας 9.1 *Δεδομένα Άσκησης 1*

Λύση

Αρχικά πρέπει να υπολογιστεί η Εντροπία του συνόλου E(S), σύμφωνα με την Εξίσωση 9.1. Θεωρούμε ως θετική κλάση την έγκριση του δανείου. Στο σύνολο δεδομένων υπάρχουν εννέα θετικές και πέντε αρνητικές παρατηρήσεις. Η Εντροπία υπολογίζεται ως εξής:

$E(S)= - (9/14)*\log_2(9/14) – (5/14)*\log_2(5/14) = 0,94.$

Εάν επιλεγεί το Εισόδημα ως μεταβλητή διαχωρισμού, τότε το σύνολο δεδομένων θα διαχωριστεί σε τρία υποσύνολα, όπου στο πρώτο υποσύνολο S_1 θα περιλαμβάνονται οι υποψήφιοι με χαμηλό εισόδημα, στο δεύτερο υποσύνολο S_2 οι υποψήφιοι με υψηλό εισόδημα και στο τρίτο υποσύνολο S_3 οι υποψήφιοι με μεσαίο εισόδημα. Η συνολική Εντροπία διαχωρισμού θα υπολογιστεί σύμφωνα με την Εξίσωση 9.3. Αρχικά πρέπει να υπολογιστούν οι Εντροπίες των τριών υποσυνόλων.

Το υποσύνολο S_1 περιέχει τρεις θετικές και μια αρνητική παρατήρηση. Η Εντροπία του υπολογίζεται ως εξής:

$E(S_1) = -(3/4)*\log_2(3/4) – (1/4)*\log_2(1/4) = 0,811.$

Το υποσύνολο S_2 περιέχει δύο θετικές και δύο αρνητικές παρατηρήσεις. Η Εντροπία του υπολογίζεται ως εξής:

$E(S_2) = -(2/4)*\log_2(2/4) – (2/4)*\log_2(2/4) = 1.$

Το υποσύνολο S_3 περιέχει τέσσερεις θετικές και δύο αρνητικές παρατηρήσεις. Η Εντροπία του υπολογίζεται ως εξής:

$E(S_3) = -(4/6)*\log_2(4/6) – (4/6)*\log_2(4/6) = 0,918.$

Το υποσύνολο S_1 περιέχει τέσσερεις παρατηρήσεις, το υποσύνολο S_2 περιέχει τέσσερεις παρατηρήσεις, το υποσύνολο S_3 περιέχει έξι παρατηρήσεις, και το αρχικό σύνολο περιέχει δέκα τέσσερεις παρατηρήσεις. Σύμφωνα με την Εξίσωση 9.3, η Εντροπία διαχωρισμού θα υπολογιστεί ως εξής:

$E(S, Εισόδημα) = (4/14)*E(S_1) + (4/14)*E(S_2) + (6/14)*E(S_3) = 0,911.$

Το Κέρδος Πληροφορίας υπολογίζεται σύμφωνα με την Εξίσωση 9.4
IG(S, Εισόδημα) = E(S) - E(S, Εισόδημα) = 0,94 – 0,911 = 0,029.

Άσκηση Υπολογισμών 9.2

Χρησιμοποιήστε τα δεδομένα του παρακάτω πίνακα. Εφαρμόστε το θεώρημα του Bayes για να προβλέψετε τις πιθανότητες έγκρισης και απόρριψης της αίτησης ενός υποψηφίου με μέσο εισόδημα και μικρή ηλικία.

ΕΙΣΟΔΗΜΑ	ΗΛΙΚΙΑ	ΕΓΚΡΙΣΗ
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΓΑΛΗ	No
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΧΑΜΗΛΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No
ΧΑΜΗΛΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΙΚΡΗ	Yes
ΜΕΣΟ	ΜΕΣΑΙΑ	Yes
ΥΨΗΛΟ	ΜΕΣΑΙΑ	Yes
ΜΕΣΟ	ΜΕΓΑΛΗ	No

Πίνακας 9.2 *Δεδομένα Άσκησης 2*

Λύση

Οι πιθανότητες έγκρισης και απόρριψης του δανείου θα υπολογιστούν σύμφωνα με την Εξίσωση 9.13. Ο πίνακας περιέχει 14 περιπτώσεις και στις τρεις από αυτές οι υποψήφιοι έχουν μέσο εισόδημα και μικρή ηλικία. Η πιθανότητα P(X) υπολογίζεται ως εξής:

P(X) = 3/14 = 0,21

Για τις οκτώ περιπτώσεις του συνόλου το δάνειο εγκρίνεται, ενώ για τις τέσσερεις απορρίπτεται. Οι αντίστοιχες πιθανότητες υπολογίζονται ως εξής:

P(Yes) = 8/14 = 0,57

P(No) = 6/14 = 0,429

Από τις οκτώ περιπτώσεις όπου το δάνειο εγκρίνεται, στις δύο το εισόδημα είναι μέσο και η ηλικία μικρή. Από τις έξι περιπτώσεις όπου το δάνειο δεν εγκρίνεται, στη μια περίπτωση το εισόδημα είναι μέσο και η ηλικία μικρή. Οι αντίστοιχες πιθανότητες υπολογίζονται ως εξής:

P(X|Yes) = 2/8 = 0,25

P(X|No) = 1/6 = 0,167

Σύμφωνα με την Εξίσωση 9.13, οι πιθανότητες έγκρισης και απόρριψης του δανείου για μέσο εισόδημα και μικρή ηλικία είναι:

P(Yes|X)= (0,25*0,57)/0,21 = 0,667

P(No|X)= (0,167*0,429)/0,21 = 0,333

Η πιθανότητα έγκρισης του δανείου είναι διπλάσια από την πιθανότητα απόρριψης του.

Άσκηση Εφαρμογής 9.3

Χρησιμοποιήστε το αρχείο «analcatdata_bankruptcy.arff» (θα το βρείτε στην ιστοσελίδα δεδομένων

του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή StatLib). Το σύνολο δεδομένων προέρχεται από το βιβλίο του Simonoff (2003) και σχετίζεται με τη χρεοκοπία επιχειρήσεων. Υπάρχουν 50 γραμμές, κάθε μια από τις οποίες αναφέρεται σε μια επιχείρηση. Οι μισές επιχειρήσεις έχουν χρεοκοπήσει. Στο σύνολο δεδομένων υπάρχουν 7 πεδία (στήλες). Το πρώτο πεδίο περιέχει τα ονόματα των επιχειρήσεων, και ακολουθούν 5 πεδία με αριθμοδείκτες. Το τελευταίο πεδίο είναι το πεδίο της κλάσης και περιέχει μια ένδειξη («1» ή «0») για το εάν η επιχείρηση χρεοκόπησε ή εξακολουθεί τη λειτουργία της αντίστοιχα.

Αναπτύξτε μοντέλα πρόβλεψης χρεοκοπίας με χρήση των μεθόδων α)Δένδρου Αποφάσεων C4.5, β)Νευρωνικού Δικτύου Multilayer Perceptron, γ)Μπαϋεσιανού Δικτύου. Στην επιλογή «Test Option» επιλέξτε «Cross-validation». Με την επιλογή αυτή τα μοντέλα δοκιμάζονται χρησιμοποιώντας άγνωστες παρατηρήσεις. Μελετήστε τα αποτελέσματα των μοντέλων και επιλέξτε το μοντέλο που επιτυγχάνει τις καλύτερες επιδόσεις κατηγοριοποίησης. Μελετήστε το Δένδρο Αποφάσεων. Παρατηρήστε τα βάρη των συνδέσεων του Νευρωνικού Δικτύου. Προβάλετε το Δένδρο Αποφάσεων με γραφικό τρόπο. Επαναλάβετε το πείραμα αφού προηγουμένως έχετε επιλέξει την επιλογή «Use training set» στο πεδίο «Test Options». Με την επιλογή αυτή τα μοντέλα δοκιμάζονται χρησιμοποιώντας τις παρατηρήσεις με τις οποίες εκπαιδεύτηκαν. Συγκρίνετε τις επιδόσεις των μοντέλων με τις προηγούμενες επιδόσεις που είχαν υπολογιστεί με την τεχνική «Cross-validation». Τι παρατηρείτε;

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «analcatdata_bankruptcy.arff» πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» εμφανίζονται διάφορες πληροφορίες για τα δεδομένα. Παρατηρήστε τα πεδία (Attributes). Ως πεδίο κλάσης ορίζεται αυτόματα το τελευταίο πεδίο, δηλαδή το πεδίο «Bankrupt». Κάνοντας κλικ σε ένα αριθμητικό πεδίο εμφανίζονται η ελάχιστη και μέγιστη τιμή, η μέση τιμή και η τυπική απόκλιση. Επίσης, εμφανίζεται η κατανομή των τιμών. Μπορείτε να κάνετε αρχική διερευνητική ανάλυση παρατηρώντας την κατανομή των τιμών και των κλάσεων.

Το πεδίο με τα ονόματα των επιχειρήσεων δεν προσφέρει κάτι χρήσιμο στην ανάλυση μας. Το επιλέγετε και το απομακρύνετε πιέζοντας το κουμπί «Remove».

Βήμα 2. Μεταβείτε στο tab «Classify».

Επιλέξτε μέθοδο κατηγοριοποίησης πιέζοντας το κουμπί «Choose» στο πεδίο «Classifier». Επιλέξτε πρώτα τη μέθοδο weka/classifiers/trees/J48 για το Δένδρο Αποφάσεων C4.5 και πατήστε το κουμπί «Start». Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Μπορείτε να δείτε το Δένδρο Αποφάσεων και τις επιδόσεις του μοντέλου. Το μοντέλο κατηγοριοποιεί σωστά το 78% του συνόλου των παρατηρήσεων, το 68% της κλάσης «0» και το 88% της κλάσης «1». Κάνοντας δεξί κλικ στο μοντέλο στο πεδίο «Results list» και επιλέγοντας «Visualize tree» παρουσιάζεται το δένδρο με γραφικό τρόπο.

Βήμα 3. Επιλέξτε τη μέθοδο weka/classifiers/functions/MultilayerPerceptron και πατήστε το κουμπί «Start». Εμφανίζονται τα βάρη των συνδέσεων. Το μοντέλο κατηγοριοποιεί σωστά το 90% των συνολικών περιπτώσεων, το 88% της κλάσης «0» και το 92% της κλάσης «1».

Βήμα 4. Επιλέξτε τη μέθοδο weka/classifiers/bayes/BayesNet και πατήστε το κουμπί «Start». Το μοντέλο κατηγοριοποιεί σωστά το 88% των συνολικών περιπτώσεων, και το 88% των κλάσεων «0» και «1».

Καλύτερες επιδόσεις επιτυγχάνει το Νευρωνικό Δίκτυο.

Βήμα 5. Επαναλάβετε τα τρία προηγούμενα βήματα έχοντας ενεργοποιήσει την επιλογή «Use training set». Τα αποτελέσματα που λαμβάνετε έχουν υπολογιστεί χρησιμοποιώντας τις παρατηρήσεις με τις οποίες εκπαιδεύτηκαν τα μοντέλα. Παρατηρήστε τις σημαντικές αυξήσεις των επιδόσεων. Για παράδειγμα, το Δένδρο Αποφάσεων αυξάνει το ποσοστό ακρίβειας στο 96% (από 78%). Ωστόσο, οι επιδόσεις αυτές είναι πλασματικές. Η πραγματική αξία των μοντέλων βρίσκεται στην ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων.

Άσκηση Εφαρμογής 9.4

Χρησιμοποιήστε το αρχείο «credit-a» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή UCI repository). Το σύνολο δεδομένων προσφέρθηκε από τον καθηγητή Ross Quinlan και χρησιμοποιήθηκε στο Quinlan (1987). Πρόκειται για δεδομένα αιτήσεων για πιστωτικές κάρτες. Υπάρχουν συνολικά 16 πεδία, όπου το τελευταίο είναι το πεδίο της

κλάσης. Στο πεδίο κλάσης χρησιμοποιούνται τα σύμβολα «+» και «-» για τις «καλές» και τις «κακές» αιτήσεις αντίστοιχα. Επίσης το αρχείο περιέχει 690 παρατηρήσεις, εκ των οποίων οι 307 ανήκουν στην κλάση «+» και 383 στην κλάση «-». Τα ονόματα των πεδίων έχουν αλλαχθεί και οι τιμές των δεδομένων έχουν κωδικοποιηθεί για λόγους απορρήτου.

Εφαρμόστε επιλογή χαρακτηριστικών με τη μέθοδο CFS Subset Evaluator. Στη συνέχεια αναπτύξτε μοντέλο νευρωνικού δικτύου τύπου Multilayer Perceprton και επικυρώστε το με τη μέθοδο «Cross-validation». Πειραματιστείτε με τις παραμέτρους της μεθόδου και προσπαθήστε να αυξήσετε τις επιδόσεις.

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «credit-a.arff» πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» εκτελέστε την επιλογή χαρακτηριστικών. Στο πεδίο "Filter" πιέστε το κουμπί "Choose" και επιλέξτε weka/filters/supervised/attribute/AttributeSelection. Αυτομάτως επιλέγεται η μέθοδος CfsSubsetEval. Εάν επιθυμείτε, μπορείτε να επιλέξτε μια άλλη μέθοδο κάνοντας κλικ στα περιεχόμενα του πεδίου «Filter». Αφού ορίσετε τη μέθοδο που επιθυμείτε (τα αποτελέσματα που ακολουθούν ισχύουν για τη CFS), κάνετε κλικ στο κουμπί «Apply». Θα διαπιστώσετε ότι στο πεδίο «Attributes» μειώνεται το πλήθος των στηλών. Ειδικότερα, παραμένουν οι στήλες A4, A6, A8, A9, A11, A14, A15 και η στήλη της κλάσης, ενώ οι υπόλοιπες στήλες απομακρύνονται.

Βήμα 2. Μεταβείτε στο tab «Classify» και επιλέξτε κατηγοριοποιητή κάνοντας κλικ στο κουμπί «Choose» του πεδίου «Classifier». Από τις διαθέσιμες μεθόδους επιλέξτε τη μέθοδο weka/classifiers/functions/MultilayerPerceptron.

Εκπαιδεύστε το μοντέλο και επικυρώστε το με τη μέθοδο «CrossValidation». Για να εκτελέσετε αυτήν την εργασία βεβαιωθείτε ότι είναι επιλεγμένη η μέθοδος «Cross-validation» στο πεδίο «Test-options» και στη συνέχεια κάντε κλικ στο κουμπί «Start».

Θα πρέπει να περιμένετε μερικά δευτερόλεπτα. Η εκπαίδευση των Νευρωνικών δικτύων είναι αργή. Παρατηρήστε το μικρό πουλί στο κάτω δεξιά μέρος της οθόνης. Όσο το πουλί κινείται, το λογισμικό εκτελεί υπολογισμούς.

Όταν ολοκληρωθεί η εκπαίδευση και η επικύρωση θα εμφανιστούν τα αποτελέσματα στο πεδίο «Classifier Output». Το μοντέλο κατηγοριοποιεί σωστά το 83,4783% των συνολικών παρατηρήσεων, το 82,1% της κλάσης «+» και το 84,6% της κλάσης «-».

Βήμα 3. Κάντε κλικ στα περιεχόμενα του πεδίου «Classifier» και στο όνομα «MultilayerPerceptron». Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων του νευρωνικού δικτύου. Μεταβάλετε τις τιμές ορισμένων παραμέτρων και επαναλάβετε την εκπαίδευση και επικύρωση του μοντέλου. Οι σημαντικότερες παράμετροι είναι οι «hiddenLayers», «learningRate», «momentum», και «trainingTime». Οδηγίες για τις παραμέτρους μπορείτε να λάβετε κάνοντας κλικ στο κουμπί «More».

Η παράμετρος «hiddenLayers» ορίζει το πλήθος των κρυφών στρωμάτων και των κρυφών νευρώνων. Η προκαθορισμένη τιμή είναι η «a». Η κωδική αυτή τιμή σημαίνει ότι το μοντέλο έχει ένα κρυφό στρώμα με πλήθος νευρώνων ίσο με το άθροισμα των μεταβλητών εισόδου και του πλήθους των τιμών κλάσης δια δύο ((attributes + classes) /2). Για περισσότερες πληροφορίες συμβουλευτείτε τον οδηγό του WEKA στο τελευταίο κεφάλαιο, ή κάντε κλικ στο κουμπί «More».

Πειραματιστείτε αρκετές φορές με διάφορες τιμές παραμέτρων προσπαθώντας να αυξήσετε τις επιδόσεις του Νευρωνικού Δικτύου.

Ορίζοντας Learning rate = 0.2, momentum=0.1, Training Time=700 και Hidden Layers=7 Το Νευρωνικό Δίκτυο επιτυγχάνει ακρίβεια 84.6377% και κατηγοριοποιεί σωστά το 83.7% των περιπτώσεων της κλάσης «+» και το 85.4% των περιπτώσεων της κλάσης «-».

10 Εναλλακτικές Μέθοδοι και ειδικά θέματα Κατηγοριοποίησης

Σύνοψη

Στο δέκατο Κεφάλαιο ολοκληρώνεται η κάλυψη της θεματικής ενότητας της Κατηγοριοποίησης. Παρουσιάζονται τρεις πρόσθετες μέθοδοι κατηγοριοποίησης, οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ), οι k-Πλησιέστεροι Γείτονες και η Λογιστική Παλινδρόμηση. Βασική ιδέα των ΜΔΥ είναι η κατασκευή ενός υπερεπιπέδου, το οποίο διαχωρίζει τις κλάσεις. Οι δύο κλάσεις θεωρούνται γραμμικά διαχωρίσιμες. Για τον καθορισμό του βέλτιστου υπερεπιπέδου εισάγεται η έννοια του περιθωρίου. Το βέλτιστο υπερεπίπεδο διαχωρισμού είναι αυτό που εξασφαλίζει το μέγιστο περιθώριο. Οι πραγματικές περιπτώσεις γραμμικού διαχωρισμού των κλάσεων είναι μάλλον σπάνιες. Για τον λόγο αυτό, τα σημεία προβάλλονται σε έναν χώρο περισσότερων διαστάσεων με τη βοήθεια μιας διανυσματικής συνάρτησης. Στον χώρο αυτόν τα σημεία είναι γραμμικώς διαχωρίσιμα. Η συνάρτηση πυρήνα ορίζει το εσωτερικό γινόμενο της διανυσματικής συνάρτησης, το οποίο απαιτείται για τον υπολογισμό της συνάρτησης απόφασης. Στη μέθοδο των k-Πλησιέστερων Γειτόνων η μάθηση βασίζεται στην αναλογία. Κάθε παρατήρηση θεωρείται ως ένα σημείο μέσα σε έναν πολυδιάστατο χώρο. Για την πρόβλεψη της κλάσης μιας νέας παρατήρησης, εντοπίζονται τα k πλησιέστερα σημεία και η νέα παρατήρηση εκχωρείται στην κλάση που πλειοψηφεί μεταξύ των k γειτονικών σημείων. Η Λογιστική (ή Λογαριθμική) Παλινδρόμηση είναι η κλασική μέθοδος την οποία χρησιμοποιούν οι οικονομολόγοι για την αντιμετώπιση προβλημάτων κατηγοριοποίησης. Σε ένα πρόβλημα δυαδικής κλάσης, ο λογάριθμος του λόγου των πιθανοτήτων να ανήκει η παρατήρηση στις δύο τιμές της κλάσης εκφράζεται ως γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών. Άλλες εκδοχές της Παλινδρόμησης χρησιμοποιούνται για την πρόβλεψη αριθμητικών τιμών. Στο παρόν Κεφάλαιο παρουσιάζονται η Απλή Γραμμική Παλινδρόμηση, η Πολλαπλή Γραμμική Παλινδρόμηση και η Πολυωνυμική Παλινδρόμηση.
Τα τελευταία χρόνια ιδιαίτερη άνθηση γνωρίζουν οι λεγόμενοι σύνθετοι κατηγοριοποιητές. Οι σύνθετοι κατηγοριοποιητές μπορούν να χωριστούν σε δύο βασικές κατηγορίες, στους συνδυασμούς κατηγοριοποιητών και στους υβριδικούς κατηγοριοποιητές. Στους συνδυασμούς κατηγοριοποιητών δημιουργείται ένας αριθμός ατομικών κατηγοριοποιητών, οι οποίοι προβλέπουν την κλάση και η τελική απόφαση υπολογίζεται με συνάθροιση των ατομικών αποφάσεων. Στους υβριδικούς κατηγοριοποιητές εφαρμόζονται ετερογενείς τεχνικές, κάθε μια από τις οποίες επιλύει ένα διαφορετικό πρόβλημα. Η τελική απόφαση κατηγοριοποίησης λαμβάνεται από έναν μόνο κατηγοριοποιητή. Η ακρίβεια ενός μοντέλου πρέπει να εκτιμάται έναντι παρατηρήσεων, οι οποίες δεν ανήκουν στο σύνολο εκπαίδευσης. Η διαδικασία αυτή ονομάζεται επικύρωση του μοντέλου και έχουν προταθεί σχετικές τεχνικές. Στο παρόν Κεφάλαιο παρουσιάζονται η μέθοδος holdout, η διασταυρούμενη επικύρωση 10 τμημάτων, η μέθοδος «άφησε ένα έξω» και η μέθοδος bootstrap. Δύο σημαντικά θέματα, τα οποία συναντώνται συχνά σε ρεαλιστικά προβλήματα κατηγοριοποίησης, είναι το πρόβλημα της ανισοκατανομής των κλάσεων και το πρόβλημα του διαφορετικού κόστους σφάλματος. Τα δύο αυτά θέματα αναλύονται με σύντομο, αλλά ουσιαστικό τρόπο. Για την εκτίμηση και παρουσίαση της ικανότητας των μοντέλων να προβλέπουν συγκεκριμένη τιμή κλάσης έχουν προταθεί ειδικές τεχνικές. Στο παρόν κεφάλαιο παρουσιάζονται ο πίνακας σύγχυσης (confusion matrix) και η καμπύλες ROC (Receiver Operating Characteristics). Τέλος, παρατίθεται μια μελέτη περίπτωσης, όπου εφαρμόζονται τεχνικές κατηγοριοποίησης για την πρόβλεψη του τύπου του εξωτερικού ελεγκτή. Εφαρμόζονται Δένδρα Αποφάσεων, Νευρωνικά Δίκτυα και k-Πλησιέστεροι Γείτονες, καθώς και συνδυασμοί κατηγοριοποιητών τύπου bagging, οι οποίοι βελτιώνουν περαιτέρω τις επιδόσεις των ατομικών τεχνικών.

Προηγούμενη γνώση

Στο παρόν Κεφάλαιο παρουσιάζονται μέθοδοι και αναπτύσσονται ειδικά θέματα κατηγοριοποίησης. Για την κατανόηση αυτών των θεμάτων απαιτείται η προηγούμενη ανάγνωση του ένατου κεφαλαίου και ειδικά των υποκεφαλαίων από 9.1 έως και 9.6, τα οποία εισάγουν τον αναγνώστη σε βασικές έννοιες κατηγοριοποίησης. Επίσης, χρήσιμη είναι η προηγούμενη ανάγνωση του Κεφαλαίου 6, το οποίο αποτελεί εισαγωγή στην Εξόρυξη Δεδομένων και του Κεφαλαίου 7, το οποίο αναφέρεται στην προεπεξεργασία των δεδομένων. Πρόσθετη πληροφόρηση για τα παραπάνω θέματα μπορεί να αναζητήσει ο ενδιαφερόμενος αναγνώστης σε ένα από τα πολλά συγγράμματα Εξόρυξης Δεδομένων. Ενδεικτικά αναφέρουμε τα βιβλία των Han, Kamber and Pei (2011) και των Maimon and Rokach (2010). Για μεθόδους που χρησιμοποιούν συναρτήσεις πυρήνα και ειδικότερα και τις Μηχανές Διανυσμάτων Υποστήριξης, ενδιαφέρον παρουσιάζει η ιστοθέση της kernel-machines.org.

10.1 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) (Support Vector Machines (SVM)) προτάθηκαν από τον Vapnik (1995) και γρήγορα γνώρισαν μεγάλη διάδοση λόγω της στιβαρής θεωρητικής θεμελίωσης τους και των υψηλών επιδόσεων τους. Οι ΜΔΥ αποτέλεσαν αντικείμενο ενδιαφέροντος πολλών ερευνητών και εφαρμόστηκαν για την ανάπτυξη μοντέλων σε πλήθος προβλημάτων κατηγοριοποίησης.

Βασική ιδέα των ΜΔΥ είναι η κατασκευή ενός **υπερεπιπέδου** (hyperplane), το οποίο διαχωρίζει τις κλάσεις και λειτουργεί ως συνάρτηση απόφασης. Οι νέες παρατηρήσεις κατηγοριοποιούνται ανάλογα με την πλευρά του υπερεπιπέδου στην οποία βρίσκονται. Ας θεωρήσουμε μια απλή περίπτωση όπου η κλάση είναι δυαδική και οι παρατηρήσεις είναι γραμμικά διαχωρίσιμες. Το **κυρτό περίβλημα** (convex hull) ενός συνόλου σημείων είναι το μικρότερο κυρτό πολύγωνο, το οποίο περικλείει όλα τα σημεία του συνόλου. Οι δύο κλάσεις είναι **γραμμικά διαχωρίσιμες**, όταν τα κυρτά περιβλήματα τους δεν επικαλύπτονται. Παράδειγμα παρατηρήσεων δυαδικής κλάσης, οι οποίες είναι γραμμικά διαχωρίσιμες, απεικονίζεται στο Σχήμα 10.1.Α. Οι παρατηρήσεις συμβολίζονται ως μικροί κύκλοι, ενώ το διαφορετικό χρώμα συμβολίζει τις διαφορετικές κλάσεις. Η μια τιμή κλάσης μπορεί να οριστεί ως θετική και να συμβολιστεί με την τιμή +1, ενώ η άλλη τιμή να οριστεί ως αρνητική και να συμβολιστεί με την τιμή -1.

Το γενικό υπερεπίπεδο διαχωρισμού ορίζεται από την Εξίσωση 10.1

$$w^T x + b = 0 \tag{10.1}$$

όπου *w* είναι ένα διάνυσμα βαρών, το οποίο είναι κάθετο στο επίπεδο και ορίζει τον προσανατολισμό του και *b* είναι το κατώφλι. Η μεταβολή της τιμής του *b* έχει σαν αποτέλεσμα την παράλληλη μετατόπιση του επιπέδου. Για μια παρατήρηση x_1 θετικής κλάσης ισχύει ότι

$$w^T x_1 + b > 0 \tag{10.2}$$

ενώ για μια παρατήρηση x_2 αρνητικής κλάσης ισχύει ότι

$$w^T x_2 + b < 0 \tag{10.3}$$

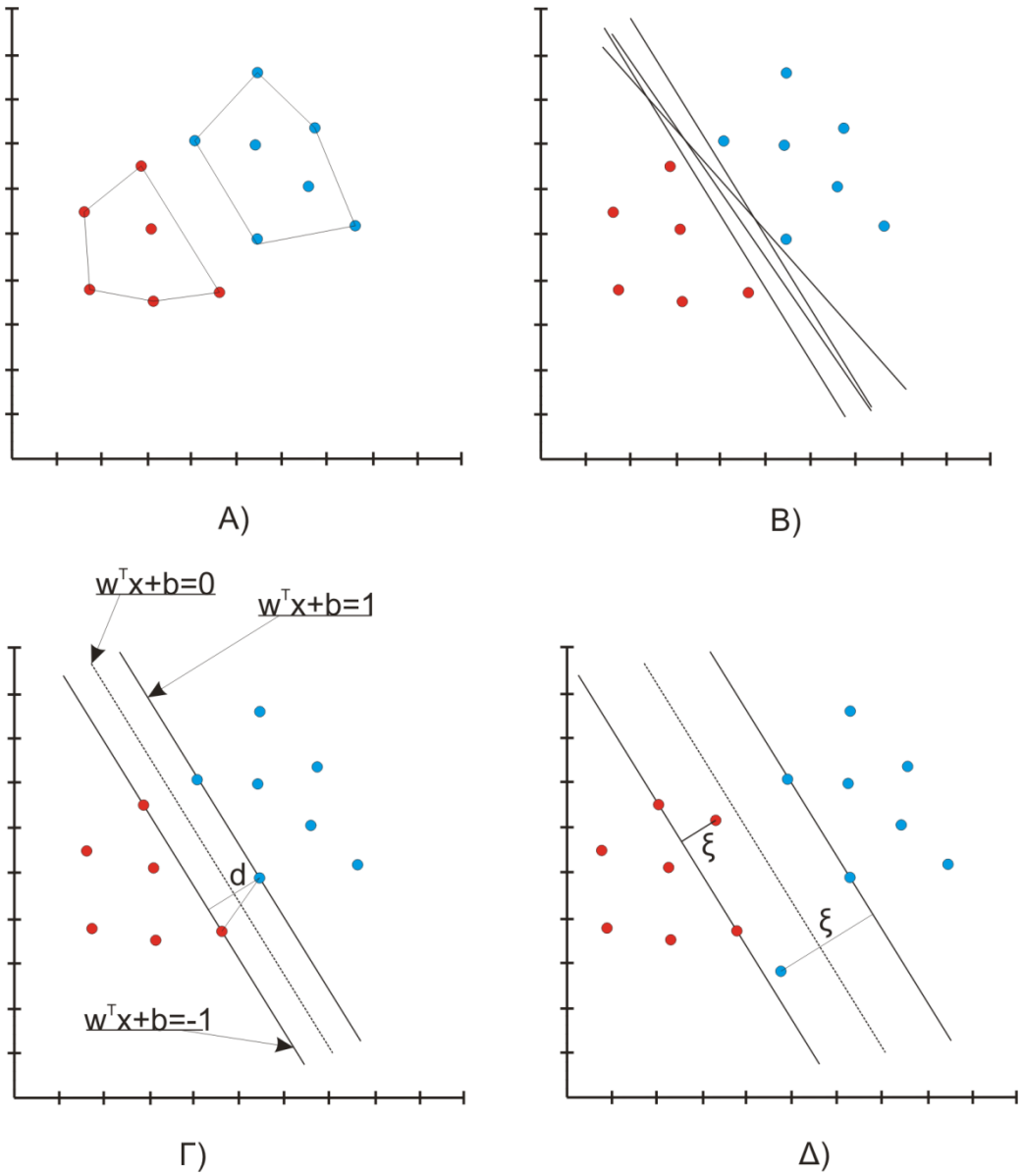
Πλέον το πρόβλημα της κατηγοριοποίησης ανάγεται σε πρόβλημα καθορισμού του υπερεπιπέδου διαχωρισμού. Όπως φαίνεται στο σχήμα 10.1.Β υπάρχουν πολλά υπερεπίπεδα, τα οποία θα μπορούσαν να χρησιμοποιηθούν, και το ερώτημα είναι ποιο από αυτά είναι το καλύτερο. Για τον υπολογισμό του βέλτιστου επιπέδου εισάγεται η έννοια του περιθωρίου (margin). Ως **περιθώριο** ορίζεται η μικρότερη απόσταση ενός σημείου από το υπερεπίπεδο διαχωρισμού. Η κλίμακα του περιθωρίου επηρεάζεται από το διάνυσμα βαρών *w*. Θεωρούμε τα σημεία x_i, τα οποία είναι πλησιέστερα στο υπερεπίπεδο. Μπορούμε να ρυθμίσουμε τις τιμές των *w* και *b* έτσι ώστε η απόσταση των σημείων αυτών από το υπερεπίπεδο να είναι ίση με 1 (Εξίσωση 10.4)

$$|(w^T x_i) + b| = 1 \tag{10.4}$$

Θεωρούμε δύο σημεία x_1 και x_2 τα οποία είναι πλησιέστερα στο υπερεπίπεδο, δηλαδή η απόσταση τους από αυτό είναι ίση με 1 και τα οποία βρίσκονται εκατέρωθεν του υπερεπιπέδου, δηλαδή η τιμή κλάσης του ενός είναι +1 και του άλλου -1. Από τα σημεία αυτά μπορούμε να ορίσουμε το περιθώριο ως την απόσταση τους *d*, μετρημένη κάθετα στο υπερεπίπεδο, όπως φαίνεται και στο σχήμα 10.1.Γ. Το περιθώριο υπολογίζεται σύμφωνα με τη Σχέση 10.5

$$\left(\frac{w}{\|w\|}(x_1 - x_2)\right) = \frac{2}{\|w\|}$$

(10.5)



Σχήμα 10.1 *Μηχανές Διανυσμάτων Υποστήριξης*

Το βέλτιστο υπερεπίπεδο διαχωρισμού των κλάσεων είναι αυτό που εξασφαλίζει το **μέγιστο περιθώριο**. Τα σημεία, τα οποία βρίσκονται στο όριο του περιθωρίου, ονομάζονται **διανύσματα υποστήριξης**. Προφανώς η κάθετη απόσταση από το υπερεπίπεδο των σημείων x_1 και x_2 είναι ίση με το μισό του περιθωρίου, δηλαδή $1/\|w\|$. Το πρόβλημα μετατρέπεται σε ένα πρόβλημα βελτιστοποίησης. Η ποσότητα $1/\|w\|$ πρέπει να μεγιστοποιηθεί για κάθε σημείο, με τον περιορισμό ότι η απόσταση του πλησιέστερου σημείου θα είναι ίση με 1. Για n σημεία x_i το παραπάνω πρόβλημα διατυπώνεται ως εξής:

$$Maximize\ \frac{1}{\|w\|}$$

(10.6)

με τον περιορισμό ότι

$$\min_{i=1,2,..,n} |w^T x_i + b| = 1$$

(10.7)

Με δεδομένο ότι η κλάση y_i μιας παρατήρησης x_i μπορεί να πάρει τιμές +1 ή -1, καθώς και ότι το $w^T x_i + b$ θα έχει τιμή >=1 για παρατηρήσεις θετικής κλάσης και <=-1 για παρατηρήσεις αρνητική κλάσης, προκύπτει ότι το γινόμενο του $(w^T x_i + b)$ με την τιμή της κλάσης θα δίνει αποτέλεσμα μεγαλύτερο ή ίσο του 1

$$y_i * (w^T x_i + b) \geq 1$$

(10.8)

Το πρόβλημα επαναδιατυπώνεται ως:

$$Minimize \frac{1}{2} \|w^2\|$$

(10.9)

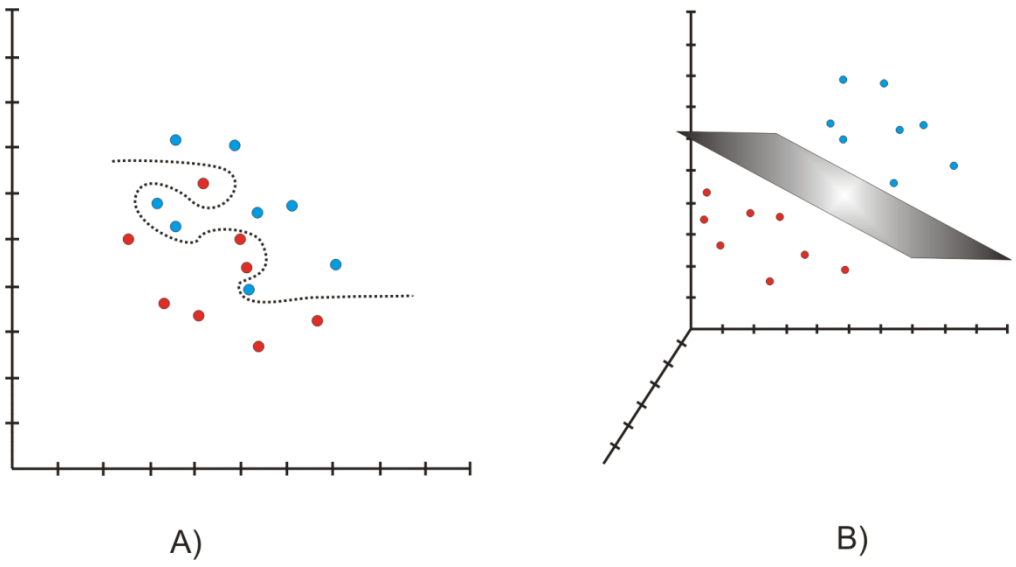
με τον περιορισμό της Σχέσης 10.8.

Το πρόβλημα μπορεί να λυθεί με τη χρήση του τετραγωνικού προγραμματισμού. Διαθέσιμα λογισμικά επιλύουν προβλήματα τετραγωνικού προγραμματισμού. Σε προβλήματα του πραγματικού κόσμου μπορεί να μην είναι όλες οι παρατηρήσεις γραμμικά διαχωρίσιμες. Για να ξεπεράσει το πρόβλημα του απόλυτου γραμμικού διαχωρισμού, ο Vapnik εισήγαγε τις μεταβλητές χαλαρότητας ξ_i. Με τη συμμετοχή των μεταβλητών χαλαρότητας, η Σχέση 10.8 τροποποιείται ως ακολούθως:

$$y_i * (w^T x_i + b) \geq 1 - \xi_i$$

(10.10)

όπου ξ_i >=0.



Σχήμα 10.2 *Γραμμικός διαχωρισμός κλάσεων σε χώρο περισσότερων διαστάσεων*

Αν για ένα σημείο x_i η μεταβλητή ξ_i είναι μεγαλύτερη από 1, τότε το σημείο κατηγοριοποιείται εσφαλμένα, όπως φαίνεται και στο σχήμα 10.1.Δ. Το άθροισμα των ξ_i μπορεί να θεωρηθεί το πλήθος των σφαλμάτων κατηγοριοποίησης. Η 9.10 μπορεί να τροποποιηθεί ώστε να περιλαμβάνει τις μεταβλητές χαλαρότητας.

$$Minimize \frac{1}{2}\|w^2\| + C\sum_{i=1}^{n}\xi_i$$

(10.11)

Η σταθερά *C* είναι μια παράμετρος, που ορίζει το ισοζύγιο μεταξύ πολυπλοκότητας και εμπειρικού σφάλματος. Οι περιπτώσεις δυνατότητας γραμμικού διαχωρισμού των κλάσεων είναι μάλλον σπάνιες σε πραγματικά προβλήματα. Εάν όμως τα σημεία x_i προβληθούν με μία μη γραμμική διανυσματική συνάρτηση $\varphi(x_i)$ σε έναν χώρο περισσότερων διαστάσεων, τότε είναι πιθανό οι απεικονίσεις τους στον νέο χώρο να είναι γραμμικώς διαχωρίσιμες. Στο Σχήμα 10.2.Α απεικονίζονται τα σημεία στον αρχικό δισδιάστατο χώρο. Τα σημεία δεν είναι γραμμικώς διαχωρίσιμα. Στο Σχήμα 10.2.Β τα σημεία προβάλλονται σε έναν τρισδιάστατο χώρο, και εκεί είναι γραμμικώς διαχωρίσιμα. Εφόσον στον χώρο αυτόν ισχύει ο γραμμικός διαχωρισμός, μπορεί να εφαρμοστεί η μέθοδος των διανυσμάτων υποστήριξης που παρουσιάστηκε προηγουμένως. Η συνάρτηση απόφασης επαναδιατυπώνεται ως εξής:

$$f(x) = w^T\varphi(x) + b$$

(10.12)

Ο προσδιορισμός της συνάρτησης φ μπορεί να είναι εξαιρετικά δύσκολος και ο χώρος προβολής μπορεί να έχει πάρα πολλές διαστάσεις. Όμως για τον υπολογισμό της συνάρτησης απόφασης *f*, απαιτείται μόνο ο ορισμός του εσωτερικού γινόμενου $\varphi(x_i)*\varphi(x_j)$. Ορίζουμε μια συνάρτηση $K(x_i, x_j)$, οποία υπολογίζει το εσωτερικό γινόμενο των απεικονίσεων $\varphi(x_i)$ και $\varphi(x_j)$ (Σχέση 10.13). Η συνάρτηση *K* καλείται **συνάρτηση πυρήνα** (kernel function)

$$K(x_i, x_j) = \phi(x_i) * \phi(x_j)$$

(10.13)

Διάφορες συναρτήσεις μπορούν να χρησιμοποιηθούν ως συναρτήσεις πυρήνα. Σε αυτές περιλαμβάνονται η Συνάρτηση Ακτινωτής Βάσης (Radial Base Function – RBF), η Σιγμοειδής, η πολυωνυμική και η αντίστροφη πολυτετραγωνική συνάρτηση. Ο πυρήνας καθορίζει τη μορφή του υπερεπιπέδου διαχωρισμού και συνεπώς επηρεάζει την απόδοση του κατηγοριοποιητή. Η επιλογή της καλύτερης συνάρτησης πυρήνα είναι θέμα το οποίο διερευνάται (Steinwart, 2003).

Έχουν προταθεί κατάλληλες παραλλαγές των Μηχανών Διανυσμάτων Υποστήριξης, που τις καθιστούν ικανές να υπολογίζουν αριθμητικές τιμές και όχι τιμές κλάσης. Η μέθοδος ονομάζεται **Παλινδρόμηση Διανυσμάτων Υποστήριξης** (Support Vector Regression (SVR)). Η κεντρική ιδέα των SVR (Smola & Schoelkopf, 2004) είναι να οριστεί μια συνάρτηση $f(x_i)$, της οποίας το αποτέλεσμα να μην αποκλίνει περισσότερο από μια ποσότητα ε από τις πραγματικές τιμές y_i.

Οι ΜΔΥ αρχικά σχεδιάστηκαν για την επίλυση διχότομων προβλημάτων, προβλημάτων δηλαδή με δύο δυνατές τιμές κλάσης, Ωστόσο, έχουν προταθεί παραλλαγές των ΜΔΥ που τις καθιστούν ικανές να αναπτύσσουν μοντέλα κατηγοριοποίησης για προβλήματα με πολλαπλές τιμές κλάσης. Μια προσέγγιση ονομάζεται one-against-the-rest (Vapnik, 1995). Σύμφωνα με την προσέγγιση αυτή, για ένα πρόβλημα με *k* δυνατές τιμές κλάσης κατασκευάζονται *k* δυαδικοί κατηγοριοποιητές, οι οποίοι προβλέπουν τιμή +1 για τη μια τιμή κλάσης και τιμή -1 για όλες τις υπόλοιπες. Οι άγνωστες παρατηρήσεις εκχωρούνται στην κλάση με τη μεγαλύτερη τιμή απόφασης. Σύμφωνα με μια άλλη προσέγγιση, η οποία ονομάζεται one-against-one (Kreßel, 1999), αναπτύσσονται δυαδικοί κατηγοριοποιητές για κάθε δυνατό ζευγάρι τιμών κλάσης. Για προβλήματα με *k* δυνατές τιμές κλάσης αναπτύσσονται συνολικά *k(k-1)/2* κατηγοριοποιητές. Σχήματα ψηφοφορίας χρησιμοποιούνται για την τελική κατηγοριοποίηση.

Οι Μηχανές Διανυσμάτων Υποστήριξης είναι πολύ δημοφιλείς, χάρη στα ιδιαίτερα χαρακτηριστικά τους και τα πολλά πλεονεκτήματα τους. Τα κύρια **πλεονεκτήματα** τους είναι τα ακόλουθα:

- Η χρήση της συνάρτησης πυρήνα τις καθιστά πολύ αποτελεσματικές σε περιπτώσεις όπου υπάρχουν μη γραμμικές σχέσεις στα δεδομένα.
- Επιτυγχάνουν υψηλές επιδόσεις κατηγοριοποίησης, κυρίως στην περίπτωση δυαδικών κλάσεων.
- Διαθέτουν στιβαρή θεωρητική θεμελίωση.
- Είναι ανθεκτικές στην υπερπροσαρμογή και διαθέτουν πολύ καλή δυνατότητα γενίκευσης με κατάλληλη ρύθμιση της παραμέτρου C.
- Δεν παγιδεύονται σε τοπικά ελάχιστα.
- Είναι αποτελεσματικές σε περιπτώσεις συνόλων δεδομένων με πολλές στήλες και σχετικά λίγες γραμμές.

Τα βασικότερα **μειονεκτήματα** των Μηχανών διανυσμάτων Υποστήριξης είναι τα ακόλουθα:

- Δεν υπάρχει κάποια μεθοδολογία για την επιλογή της συνάρτησης πυρήνα καθώς και των παραμέτρων του πυρήνα.
- Δεν παρέχουν ερμηνεύσιμα μοντέλα. Η συμβολή της εκάστοτε μεταβλητής εισόδου στο τελικό αποτέλεσμα κατηγοριοποίησης είναι αδιαφανής.
- Έχουν σχετικά μεγάλους χρόνους εκπαίδευσης, αν και σημαντικά χαμηλότερους από αυτούς των Νευρωνικών Δικτύων.
- Έχουν μεγάλες απαιτήσεις σε μνήμη υπολογιστή.
- Σε περίπτωση κλάσεων με πολλαπλές τιμές το πρόβλημα διατυπώνεται σαν συνδυασμός προβλημάτων δυαδικών κλάσεων.

10.2 k-Πλησιέστεροι Γείτονες

Οι **Κατηγοριοποιητές Βασισμένοι σε Παραδείγματα** (Instance Based Classifiers (IBC)) είναι μια οικογένεια κατηγοριοποιητών, όπου η μάθηση βασίζεται στην αναλογία. Οι κατηγοριοποιητές IBC δεν παράγουν κάποιο μοντέλο γενίκευσης. Μέθοδοι κατηγοριοποίησης όπως τα Νευρωνικά Δίκτυα, τα Δένδρα Αποφάσεων ή τα Μπαϋεσιανά Δίκτυα Πίστης ολοκληρώνουν την εκπαίδευση με τη δημιουργία κάποιου μοντέλου. Ακολούθως, το μοντέλο χρησιμοποιείται για την κατηγοριοποίηση νέων παρατηρήσεων. Σε αντίθεση με αυτές τις μεθόδους, στους κατηγοριοποιητές IBC δεν υπάρχει κάποιο στάδιο εκπαίδευσης και δεν παράγεται κάποιο μοντέλο, μέχρι να χρειαστεί να κατηγοριοποιηθεί μια νέα παρατήρηση. Για τον λόγο αυτό, οι κατηγοριοποιητές IBC καλούνται και «οκνηροί» (lazy classifiers). Όταν χρειαστεί να κατηγοριοποιήσουν μια νέα παρατήρηση, τη συγκρίνουν με γνωστές παρατηρήσεις του συνόλου εκπαίδευσης. Αυτό απαιτεί την αποθήκευση όλων ή τουλάχιστον ενός μέρους των παρατηρήσεων εκπαίδευσης. Αντιθέτως, σε άλλες τεχνικές, όπως τα SVM, μπορούν να απορριφθούν όλες οι παρατηρήσεις εκπαίδευσης που δεν είναι διανύσματα υποστήριξης

Η μέθοδος των **k-Πλησιέστερων Γειτόνων** (k-Nearest Neighbors - kNN) είναι ένας αλγόριθμος της οικογένειας των Κατηγοριοποιητών Βασισμένων σε Παραδείγματα. Για λόγους απλότητας θεωρούμε αρχικά ένα πρόβλημα κατηγοριοποίησης, όπου οι παρατηρήσεις αποτελούνται από δύο αριθμητικά πεδία και το γνώρισμα της κλάσης. Κάθε παρατήρηση μπορεί να θεωρηθεί ως ένα σημείο στον χώρο των δύο διαστάσεων. Μια παρατήρηση X απέχει από μια άλλη παρατήρηση Y, απόσταση d(X,Y) μέσα στον δισδιάστατο χώρο. Η απόσταση d(X,Y) μπορεί να υπολογιστεί ως η Ευκλείδεια απόσταση σύμφωνα με την Εξίσωση 10.14:

$$d(X,Y) = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2)}$$

(10.14)

όπου x_1, y_1 οι τιμές των X και Y για την πρώτη διάσταση και x_2, y_2 οι τιμές των X και Y για τη δεύτερη διάσταση. Σύμφωνα με τον αλγόριθμο k-NN, ο χρήστης προκαθορίζει την τιμή της σταθερής παραμέτρου k. Ο αλγόριθμος αναζητά μέσα στον δισδιάστατο χώρο τα k σημεία-παρατηρήσεις που βρίσκονται πλησιέστερα στη νέα παρατήρηση. Ο κατηγοριοποιητής εκχωρεί τη νέα παρατήρηση στην κλάση που πλειοψηφεί μεταξύ των k πλησιέστερων γειτόνων. Εάν οριστεί ότι $k=1$, τότε η νέα παρατήρηση εκχωρείται στην κλάση της πιο όμοιας παρατήρησης εκπαίδευσης

Τα παραπάνω παρουσιάζονται διαγραμματικά στο Σχήμα 10.3. Στο παράδειγμα υπάρχουν δύο δυνατές τιμές κλάσης, οι οποίες συμβολίζονται με το χρώμα των σημείων. Το κίτρινο σημείο συμβολίζει τη νέα πα-

ρατήρηση που θα κατηγοριοποιηθεί. Στο παράδειγμα η τιμή του *k* έχει οριστεί να είναι 5. Εντοπίζονται τα 5 πλησιέστερα σημεία. Παρατηρούμε ότι τρία από αυτά είναι κόκκινα και δύο είναι μπλε. Η νέα παρατήρηση εκχωρείται στην «κόκκινη» κλάση.

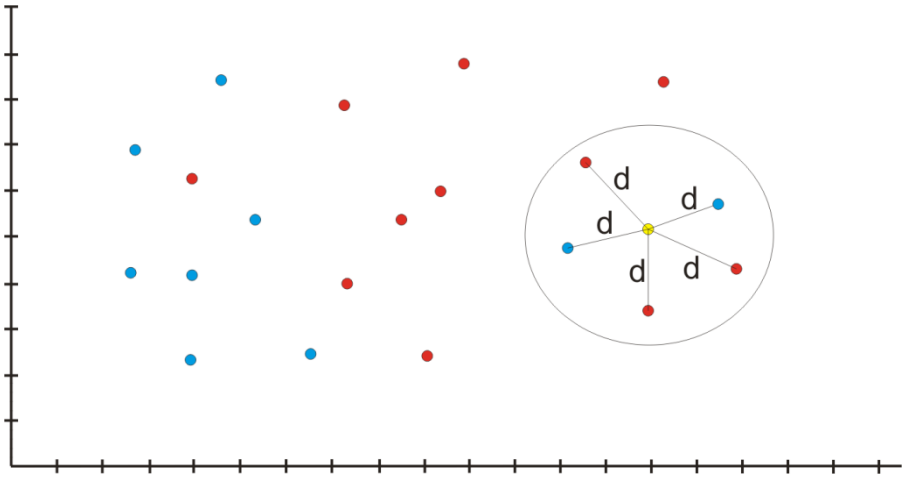
Κατ' αντιστοιχία, ο ίδιος αλγόριθμος ισχύει για παρατηρήσεις με *n* αριθμητικές διαστάσεις. Οι παρατηρήσεις θεωρούνται σημεία στον n-διάστατο χώρο και η Ευκλείδεια απόσταση υπολογίζεται σύμφωνα με την Εξίσωση 10.15.

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

(10.15)

Μια βελτίωση στον παραπάνω αλγόριθμο είναι να μην λαμβάνεται η απόφαση κατηγοριοποίησης με ισότιμη ψηφοφορία μεταξύ των επιλεγμένων γειτόνων, αλλά να συνεισφέρουν περισσότερο τα σημεία τα οποία είναι πλησιέστερα στη νέα παρατήρηση. Ένας απλός τρόπος για να επιτευχθεί αυτό είναι να εκχωρηθούν συντελεστές βαρύτητας ψήφου στα επιλεγμένα σημεία. Οι συντελεστές θα μπορούσαν να είναι ίσοι με *1/d*, όπου *d* η απόσταση του εκάστοτε σημείου από τη νέα παρατήρηση.

Μία αδυναμία στον υπολογισμό της ομοιότητας με βάση την Ευκλείδεια απόσταση είναι το γεγονός ότι οι μεταβλητές με μεγάλο εύρος τιμών επηρεάζουν περισσότερο το αποτέλεσμα από τις μεταβλητές με μικρό εύρος τιμών. Εάν πχ οι παρατηρήσεις έχουν δύο γνωρίσματα Α και Β και το Α παίρνει τιμές από 1 έως 1000, ενώ το Β παίρνει τιμές από 1 έως 10, τότε το γνώρισμα Α επηρεάζει δυσανάλογα την απόσταση σε σχέση με το γνώρισμα Β. Το πρόβλημα αυτό αντιμετωπίζεται με κανονικοποίηση των αριθμητικών τιμών. Αυτό μπορεί να επιτευχθεί διαιρώντας τις τιμές των γνωρισμάτων με την περιοχή τιμών των γνωρισμάτων.



Σχήμα 10.3 *Κατηγοριοποιητής k-NN με k=5*

Ένα άλλο συγγενές πρόβλημα είναι το γεγονός ότι ο υπολογισμός της ομοιότητας με βάση την Ευκλείδεια απόσταση υποθέτει την ισότιμη συμμετοχή όλων των γνωρισμάτων, κάτι που γενικώς δεν ισχύει. Το πρόβλημα αυτό αντιμετωπίζεται με τον καθορισμό «βαρών» για την κάθε διάσταση. Ο καθορισμός των βαρών επιτρέπει την αναδιατύπωση του υπολογισμού της απόστασης σύμφωνα με την Εξίσωση 10.16:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} w_i * (x_i - y_i)^2}$$

(10.16)

όπου w_i είναι το βάρος που αντιστοιχεί στην *i*-οστή διάσταση. Ο καθορισμός των βαρών αποτελεί ένα ενεργό πεδίο έρευνας που έχει αποδώσει διάφορες μεθόδους υπολογισμού των βαρών.

Ένας άλλος περιορισμός του υπολογισμού της ομοιότητας με βάση την Ευκλείδεια απόσταση ή κάποια παραλλαγή της είναι το γεγονός ότι αυτές οι προσεγγίσεις προϋποθέτουν γνωρίσματα αριθμητικών τιμών. Για να αντιμετωπιστεί αυτό το πρόβλημα, έχουν προταθεί συναρτήσεις που υπολογίζουν την απόσταση παρατηρήσεων που αποτελούνται από ονομαστικές τιμές. Στην απλούστερη εκδοχή τους οι συναρτήσεις αυτές επιστρέφουν την τιμή 0 εάν οι τιμές του ίδιου ονομαστικού γνωρίσματος δύο διαφορετικών παρατηρήσεων είναι ίδιες, αλλιώς επιστρέφουν την τιμή 1. Επίσης, έχουν προταθεί αλγόριθμοι που υπολογίζουν την απόσταση αντικειμένων που αποτελούνται και από αριθμητικές και από ονομαστικές τιμές. (Wilson & Martinez, 1997).

Η μέθοδος k-NN εκτός από κατηγοριοποίηση μπορεί να χρησιμοποιηθεί και για παλινδρόμηση, δηλαδή για πρόβλεψη αριθμητικών τιμών. Για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής μιας νέας παρατήρησης, ο αλγόριθμος εντοπίζει τις *k* πλησιέστερες παρατηρήσεις και επιστρέφει ως πρόβλεψη τη μέση τιμή των εξαρτημένων μεταβλητών των επιλεγμένων παρατηρήσεων.

Οι κατηγοριοποιητές k-NN διαθέτουν αξιόλογα **πλεονεκτήματα**:

- Είναι αποτελεσματικοί όταν υπάρχουν σύνθετες εξαρτήσεις μεταξύ των μεταβλητών.
- Διαθέτουν απλό αλγόριθμο.
- Σε πολλές περιπτώσεις επέτυχαν υψηλές επιδόσεις κατηγοριοποίησης.

Ορισμένα από τα βασικά **μειονεκτήματα** τους είναι τα ακόλουθα:

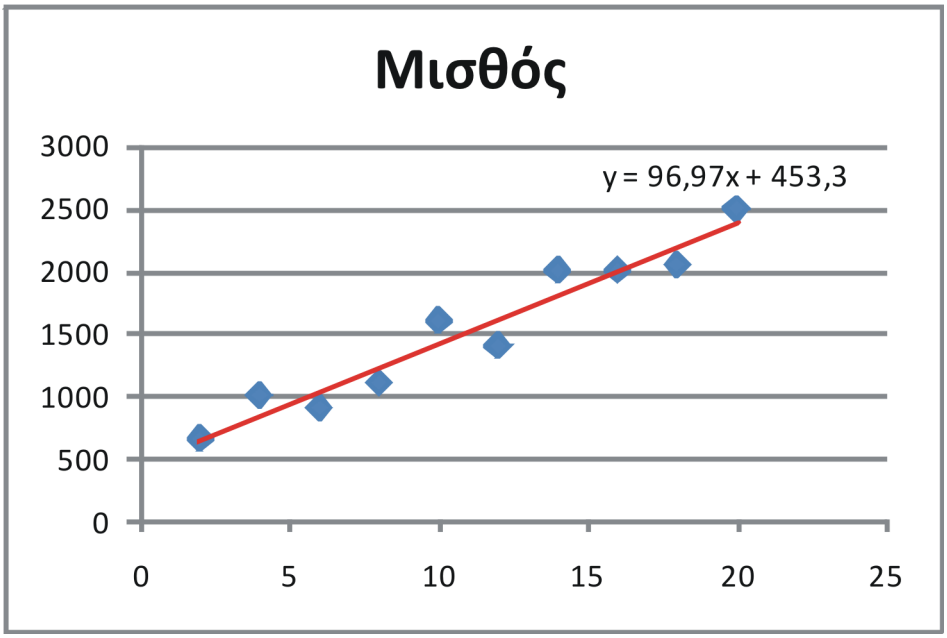
- Το γεγονός ότι γίνονται πολλές συγκρίσεις μεταξύ παρατηρήσεων απαιτεί πολύ αποτελεσματικές τεχνικές καταλογοποίησης (indexing).
- Η κατηγοριοποίηση νέων παρατηρήσεων διαρκεί πολύ περισσότερο χρόνο, ειδικά στις περιπτώσεις όπου ο αριθμός των εν δυνάμει «γειτόνων» είναι μεγάλος.
- Τα αποτελέσματα τους μπορούν να επηρεαστούν σε σημαντικό βαθμό από το πλήθος των γειτόνων k.
- Είναι ευαίσθητοι σε τοπικά χαρακτηριστικά των δεδομένων.
- Είναι ευαίσθητοι στην ύπαρξη μη σημαντικών μεταβλητών εισόδου.

10.3 Παλινδρόμηση

Ο όρος **Ανάλυση Παλινδρόμησης** (Regression Analysis) αναφέρεται σε μια οικογένεια στατιστικών τεχνικών, που στοχεύουν στη διερεύνηση σχέσεων μεταξύ μεταβλητών. Πιο συγκεκριμένα, η παλινδρόμηση μοντελοποιεί τη σχέση επίδρασης μιας ή περισσότερων μεταβλητών σε μια άλλη μεταβλητή. Η μεταβλητή της οποίας η τιμή υπολογίζεται καλείται **εξαρτημένη μεταβλητ**ή (dependent variable) ή **μεταβλητή απόκρισης** (respond variable). Οι μεταβλητές οι οποίες χρησιμοποιούνται για τον υπολογισμό της εξαρτημένης μεταβλητής ονομάζονται **ανεξάρτητες** (independent) ή **επεξηγηματικές** (explanatory). Με την Παλινδρόμηση ο χρήστης κατανοεί τον τρόπο με τον οποίο επηρεάζουν οι μεταβολές των ανεξάρτητων μεταβλητών την εξαρτημένη μεταβλητή. Επίσης, η Παλινδρόμηση μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων της τιμής της εξαρτημένης μεταβλητής. Τεχνικές Ανάλυσης Παλινδρόμησης έχουν χρησιμοποιηθεί κατά κόρον σε οικονομικές μελέτες πάσης φύσεως και αποτελούν ένα από τα βασικά και παραδοσιακά εργαλεία διεξαγωγής οικονομικών αναλύσεων.

10.3.1 Απλή Γραμμική Παλινδρόμηση

Η απλούστερη εκδοχή παλινδρόμησης είναι όταν η μεταβλητή απόκρισης εξαρτάται από μια μόνο επεξηγηματική μεταβλητή. Ως παράδειγμα, ας υποθέσουμε ότι ο μισθός εξαρτάται μόνο από την εκπαίδευση του εργαζομένου. Για τις ανάγκες του παραδείγματος μετρούμε τον βαθμό εκπαίδευσης με βάση τον χρόνο εκπαίδευσης. Αφού συγκεντρωθούν στοιχεία για ένα σύνολο εργαζομένων, τα στοιχεία αυτά καταγράφονται σε ένα διάγραμμα διασποράς. Κάθε σημείο αντιστοιχεί σε έναν εργαζόμενο, ενώ οι άξονες αντιστοιχούν στον μισθό και τον χρόνο εκπαίδευσης. Τα στοιχεία αυτά παρουσιάζονται στο Σχήμα 10.4. Είναι προφανές ότι ο μισθός αυξάνεται με τα χρόνια εκπαίδευσης. Αυτό που δεν είναι προφανές είναι η ακριβής σχέση ανάμεσα σε αυτές τις δύο μεταβλητές.



Σχήμα 10.4 *Γραμμική Παλινδρόμηση*

Ο αναλυτής διατυπώνει μια υπόθεση σχετικά με τη σχέση ανάμεσα στις μεταβλητές, τις οποίες μελετά. Υποθέτουμε ότι η σχέση μεταξύ του μισθού και του χρόνου εκπαίδευσης είναι γραμμική. Στην περίπτωση αυτή, η σχέση ανάμεσα στις δύο μεταβλητές μπορεί να αναπαρασταθεί με μια ευθεία γραμμή. Μια γραμμική σχέση ανάμεσα στη μεταβλητή Y και X αποδίδεται με την Εξίσωση 10.17.

$$Y = a + b * X$$

(10.17)

Η παράμετρος α είναι η τιμή του *Y* όταν το *X* ισούται με 0. Στο παράδειγμα μας είναι η αμοιβή του εργαζόμενου με μηδενικό χρόνο εκπαίδευσης. Η παράμετρος *β* καθορίζει την κλίση της ευθείας. Οι δύο παράμετροι, *α* και *β*, είναι άγνωστες και πρέπει να υπολογιστούν με βάση την πληροφορία, η οποία βρίσκεται στο σύνολο δεδομένων. Για τον υπολογισμό τους εφαρμόζεται η μέθοδος των ελαχίστων τετραγώνων. Εάν έχουμε *n* παρατηρήσεις (x_i, y_i) τότε οι συντελεστές α και β υπολογίζονται σύμφωνα με τις ακόλουθες εξισώσεις.

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(10.18)

$$a = \bar{y} - \beta\bar{x}$$

(10.19)

όπου $\bar{x}$ η μέση τιμή των x_i και $\bar{y}$ η μέση τιμή των y_i. Για τα δεδομένα τα οποία παρουσιάζονται στο Σχήμα 10.4, η σχέση ανάμεσα στον μισθό και τα χρόνια εκπαίδευσης αποδίδεται από την εξίσωση μισθός=453,3+96,97*χρόνια_εκπαίδευσης. Το Τετραγωνικό Σφάλμα μας δίνει μια εκτίμηση του βαθμού προσέγγισης των πραγματικών τιμών από τη συνάρτηση. Αν y_i είναι οι πραγματικές τιμές και $\hat{y}_i$ είναι οι υπολογισμένες τιμές, τότε το Τετραγωνικό Σφάλμα δίνεται από τη Σχέση 10.29.

$$TΣ = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(10.20)

Ο **συντελεστής προσδιορισμού** r2 (coefficient of determination) είναι ένα μέτρο του βαθμού της συνολικής μεταβλητότητας της εξαρτώμενης μεταβλητής, που εξηγείται από την παλινδρόμηση. Ο συντελεστής προσδιορισμού ορίζεται από την Εξίσωση 10.21.

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(10.21)

Ο συντελεστής προσδιορισμού παίρνει τιμές μεταξύ 0 και 1. Μεγάλη τιμή του σημαίνει ότι η παλινδρόμηση εξηγεί τη μεταβλητότητα της εξαρτημένης μεταβλητής. Αυτό είναι σημαντικό, εάν ο αναλυτής επιθυμεί να χρησιμοποιήσει το μοντέλο για τη διατύπωση προβλέψεων σχετικά με τις τιμές της *Y*.

Σημειώνεται ότι στην Απλή Γραμμική Παλινδρόμηση γίνονται οι παρακάτω παραδοχές:

- **Γραμμικότητα**. Υποθέτουμε ότι οι μέσες τιμές της Y, για τα διάφορα επίπεδα της X, είναι γραμμικές συναρτήσεις της X.
- **Ομοσκεδαστικότητα-Σταθερότητα Διασποράς**. Οι κατανομές της Y έχουν ίδια διασπορά για όλα τα επίπεδα της X.
- **Ανεξαρτησία**. Οι τιμές της Y, που αντιστοιχούν στα διάφορα επίπεδα της X, είναι μεταξύ τους ανεξάρτητες.
- **Κανονικότητα**. Η κατανομή της Y για όλα τα επίπεδα της X είναι κανονική.

10.3.2 Πολλαπλή Γραμμική Παλινδρόμηση

Σε πολλά προβλήματα η εξαρτημένη μεταβλητή εξαρτάται όχι από μια, αλλά από περισσότερες μεταβλητές. Η Πολλαπλή Παλινδρόμηση επιτρέπει την προσθήκη πρόσθετων παραγόντων και ποσοτικοποιεί την επίδραση τους στην εξαρτημένη μεταβλητή. Η σχέση ανάμεσα στην εξαρτημένη μεταβλητή Y και στις *n* ανεξάρτητες μεταβλητές X_i δίνεται από την Εξίσωση 10.22

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

(10.22)

Κατ' αντιστοιχία με την απλή γραμμική παλινδρόμηση, οι παράμετροι $a, b_1, b_2, \ldots b_n$ πρέπει να υπολογιστούν. Στην απλή γραμμική παλινδρόμηση, η Εξίσωση 10.17 περιγράφει μια ευθεία γραμμή. Στην πολλαπλή γραμμική παλινδρόμηση με δύο ανεξάρτητες μεταβλητές, η Εξίσωση 10.22 περιγράφει ένα επίπεδο. Η μέθοδος των ελαχίστων τετραγώνων εφαρμόζεται για τον υπολογισμό των παραμέτρων a, b_1, b_2, δηλαδή του επιπέδου. Το επίπεδο ορίζεται με τέτοιο τρόπο, ώστε το άθροισμα των τετραγωνικών λαθών ανάμεσα στις προβλεπόμενες και τις πραγματικές τιμές του *Y* να ελαχιστοποιείται. Το α είναι το σημείο τομής του επιπέδου με τον άξονα του *Y*. Το b_1 και το b_2 είναι οι κλίσεις του επιπέδου ως προς τους άξονες των X_1 και X_2 αντίστοιχα. Στην πολλαπλή παλινδρόμηση μπορούμε να έχουμε *n* ανεξάρτητες μεταβλητές X_i και στην περίπτωση αυτή κατασκευάζεται ένα υπερεπίπεδο στον αντίστοιχο χώρο. Σημειώνεται ότι για τη σωστή εκτίμηση των τιμών των παραμέτρων απαιτείται μεγάλος αριθμός παρατηρήσεων.

Ένα μοντέλο πολλαπλής παλινδρόμησης μπορεί να χρησιμοποιηθεί για την πρόβλεψη των τιμών της εξαρτημένης μεταβλητής. Σε αυτήν την περίπτωση η τιμή του συντελεστή προσδιορισμού r^2 είναι σημαντική. Επίσης, το μοντέλο μπορεί να χρησιμοποιηθεί για την εκτίμηση της σημαντικότητας των ανεξάρτητων μεταβλητών. Οι συντελεστές b_i αποτελούν μια εκτίμηση της επίδρασης της εκάστοτε μεταβλητής X_i στην *Y*. Με τη χρήση στατιστικών τεχνικών, η αναλυτική περιγραφή των οποίων βρίσκεται έξω από τα όρια του παρόντος συγγράμματος, ο αναλυτής μπορεί να ελέγξει και να αποδεχτεί ή να απορρίψει τη μηδενική υπόθεση, ότι η πραγματική τιμή ενός συντελεστή είναι μηδενική, και να εκτιμήσει το κατά πόσο ο συντελεστής είναι στατιστικά σημαντικός.

Ένα ενδεχόμενο πρόβλημα στην πολλαπλή παλινδρόμηση είναι η παράλειψη σημαντικών μεταβλητών, η μη συμμετοχή δηλαδή στο μοντέλο της Εξίσωσης 10.22 ανεξάρτητων μεταβλητών, οι οποίες επηρεάζουν ουσιαστικά την εξαρτημένη μεταβλητή Y. Η παράλειψη σημαντικών μεταβλητών έχει ουσιαστικές επιπτώσεις στην παλινδρόμηση. Ο συντελεστής προσδιορισμού r^2 μειώνεται. Επίσης, προκαλούνται μεταβολές στην τιμή

του σταθερού συντελεστή *α*. Αν η μεταβλητή η οποία παραλήφθηκε έχει θετική επίπτωση στην ανεξάρτητη μεταβλητή, τότε η τιμή του *α* αυξάνεται, ενώ αν έχει αρνητική επίπτωση, τότε η τιμή του *α* μειώνεται. Σε περίπτωση όπου η μεταβλητή η οποία παραλήφθηκε συσχετίζεται με κάποια άλλη μεταβλητή, τότε ο συντελεστής αυτής της μεταβλητής τροποποιείται. Ο αναλυτής πρέπει να προσπαθεί να συμπεριλάβει όλες τις σημαντικές μεταβλητές, αν και αυτό δεν είναι πάντα δυνατόν, καθώς μερικές σημαντικές μεταβλητές μπορεί να μην είναι παρατηρήσιμες.

Ένα άλλο γνωστό πρόβλημα στην πολλαπλή παλινδρόμηση είναι το πρόβλημα της πολυσυγγραμμικότητας (multicollinearity). Πολυσυγγραμμικότητα υπάρχει όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες μεταξύ τους, και οι τιμές της μιας μπορούν να υπολογιστούν από την άλλη. Η πολυσυγγραμμικότητα δεν έχει επιπτώσεις στην ικανότητα του μοντέλου να προβλέπει τις τιμές της εξαρτημένης μεταβλητής, έχει όμως επιπτώσεις στους συντελεστές των ανεξάρτητων μεταβλητών. Εάν ο χρήστης χρησιμοποιεί το μοντέλο για την εκτίμηση της σημαντικότητας των ανεξάρτητων μεταβλητών και υπάρχει πρόβλημα πολυσυγγραμμικότητας, τότε τα αποτελέσματα δεν είναι ασφαλή. Οι τιμές των συντελεστών μπορεί να αλλάξουν πολύ, αν προστεθεί ή αφαιρεθεί μια νέα μεταβλητή ή εάν συμβούν μικρές μεταβολές στα δεδομένα. Ένας απλός τρόπος αντιμετώπισης του προβλήματος είναι η απομάκρυνση μεταβλητών από το μοντέλο.

10.3.3 Πολυωνυμική Παλινδρόμηση

Υπάρχουν προβλήματα όπου η σχέση ανάμεσα στην εξαρτημένη και την ανεξάρτητη μεταβλητή δεν είναι γραμμική και δεν μπορεί να αποδοθεί από μια συνάρτηση της μορφής Y=α+β*X. Η **Μη Γραμμική Παλινδρόμηση** είναι μια παλινδρόμηση όπου ανάμεσα στην εξαρτημένη και στην ανεξάρτητη μεταβλητή υπάρχει μη γραμμική σχέση.

Η **Πολυωνυμική Παλινδρόμηση** είναι μια περίπτωση Μη Γραμμικής Παλινδρόμησης, όπου η σχέση ανάμεσα στα Y και X περιγράφεται με τη χρήση πολυωνύμου:

$$Y = a + b_1 X + b_2 X^2 + \cdots + b_k X^k$$

(10.23)

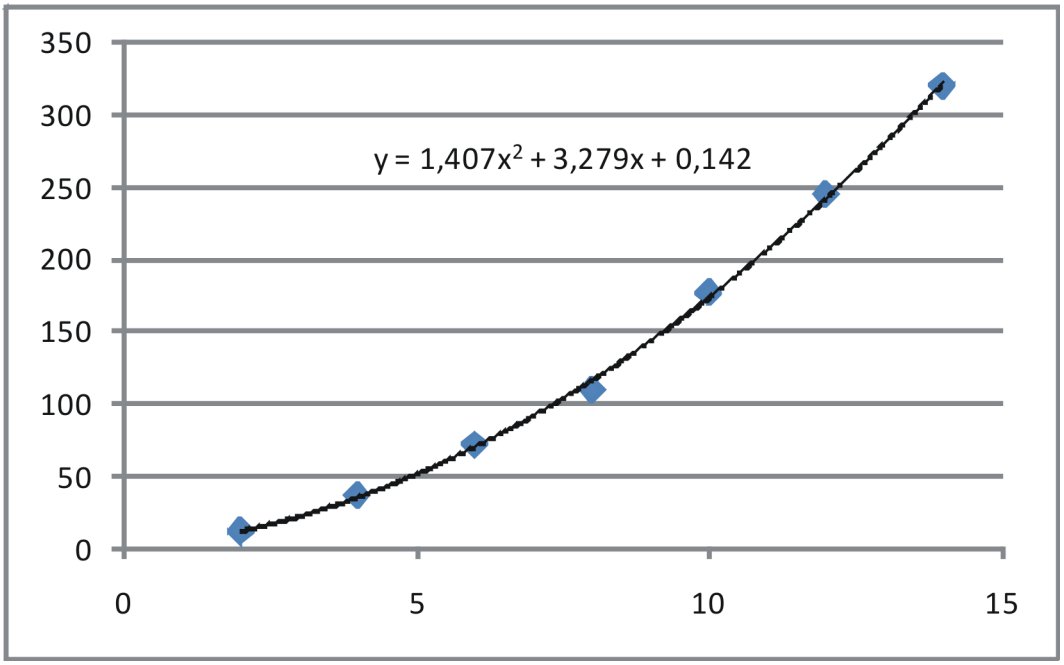
Μια πολυωνυμική συνάρτηση μπορεί να προσεγγίσει την πραγματική παλινδρόμηση, εάν από το διάγραμμα διασποράς ή από τις θεωρητικές γνώσεις για το πρόβλημα προκύπτει ότι η συνάρτηση είναι καμπυλόγραμμη. Το πολυώνυμο της Εξίσωσης 10.23 είναι βαθμού k. Στην πράξη πολυώνυμα βαθμού μεγαλύτερου από 2 αποφεύγονται, εκτός εάν δικαιολογούνται θεωρητικά. Παράδειγμα πολυωνυμικής παλινδρόμησης βαθμού 2 παρουσιάζεται στο Σχήμα 10.5

Για τον υπολογισμό των συντελεστών a, b_1, b_2 μπορεί να εφαρμοστεί η μέθοδος των ελαχίστων τετραγώνων. Μια συνάρτηση παλινδρόμησης με πολυώνυμο βαθμού 2 μπορεί να γραφεί ως

$$Y = a + b_1 X_1 + b_2 X_2$$

(10.24)

όπου X_1=X και X_2=X^2. Με τον τρόπο αυτό, η πολυωνυμική παλινδρόμηση μετατρέπεται σε πολλαπλή γραμμική παλινδρόμηση, και ο υπολογισμός των συντελεστών γίνεται με τη μέθοδο των ελαχίστων τετραγώνων.



Σχήμα 10.5 *Πολυωνυμική Παλινδρόμηση*

10.3.4 Λογιστική (ή Λογαριθμική) Παλινδρόμηση

Όλα τα είδη παλινδρόμησης, τα οποία περιγράψαμε μέχρι αυτό το σημείο, μοντελοποιούν τη σχέση ανάμεσα σε ένα σύνολο ανεξάρτητων μεταβλητών και σε μια εξαρτημένη μεταβλητή, η οποία παίρνει αριθμητικές τιμές. Η παλινδρόμηση όμως μπορεί να χρησιμοποιηθεί και για την πρόβλεψη τιμών ονομαστικών πεδίων, μπορεί δηλαδή να εφαρμοστεί σε προβλήματα κατηγοριοποίησης. Ένας πολύ συνηθισμένος τύπος παλινδρόμησης, που χρησιμοποιείται για κατηγοριοποίηση, είναι **Λογιστική Παλινδρόμηση** (Logistic Regression). Στην ελληνική γλώσσα θα τη συναντήσουμε και με το όνομα **Λογαριθμική Παλινδρόμηση**.

Θεωρούμε την περίπτωση δυαδικής κλάσης. Μπορούμε να χρησιμοποιήσουμε την τιμή 1 για τη μια τιμή της κλάσης (πχ χρεοκοπία) και την τιμή 0 για την άλλη τιμή της κλάσης (μη χρεοκοπία). Εάν p είναι η πιθανότητα να πάρει η εξαρτημένη μεταβλητή Y την τιμή 1, τότε η πιθανότητα να πάρει το Y την τιμή 0 είναι $1-p$.

$$P(Y = 1) = p; \qquad P(Y = 0) = 1 - p$$

(10.25)

Μπορούμε να συνδέσουμε την πιθανότητα p με μια γραμμική έκφραση $\alpha+\Sigma bx$ με τη βοήθεια μιας συνάρτησης, η οποία επιστρέφει τιμές στο διάστημα [0..1]. Η συνάρτηση Logit έχει αυτήν την ιδιότητα. Η Λογιστική Παλινδρόμηση περιγράφεται από τη Σχέση 10.26

$$Logit(Y = 1) = ln\left(\frac{P(Y = 1)}{P(Y = 0)}\right) = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

(10.26)

Η πιθανότητα να παίρνει το Y την τιμή 1 υπολογίζεται από την Εξίσωση 10.27.

$$P(Y = 1) = \frac{1}{1 + e^{-(a+b_1X_1+\cdots+b_nX_n)}}$$

(10.27)

Για την εκτίμηση της σημαντικότητας των ανεξάρτητων μεταβλητών, ο αναλυτής μπορεί να χρησιμοποιή-

σει ειδικούς ελέγχους, όπως η στατιστική Wald ή το Likelihood-ratio test. Η Λογιστική Παλινδρόμηση μπορεί να χρησιμοποιηθεί και για την πρόβλεψη κλάσεων με περισσότερες από δύο τιμές.

Η Λογιστική Παλινδρόμηση είναι η παραδοσιακή μέθοδος που χρησιμοποιούν οι οικονομολόγοι για να αντιμετωπίσουν προβλήματα κατηγοριοποίησης. Ο κυριότερος λόγος γι' αυτό είναι το γεγονός ότι κατά κανόνα οι οικονομολόγοι δεν είναι εξοικειωμένοι με μεθόδους κατηγοριοποίησης, οι οποίες προέρχονται από τη Μηχανική Μάθηση

Η Λογιστική Παλινδρόμηση συγκεντρώνει αρκετά **πλεονεκτήματα:**

- Είναι μια μέθοδος αρκετά απλή, δοκιμασμένη και ευρύτατα χρησιμοποιούμενη.
- Ο υπολογισμός των συντελεστών $b_1 \ldots b_n$ είναι ένα μέτρο της σημαντικότητας των ανεξάρτητων μεταβλητών. Υπό τη έννοια αυτή, η Λογιστική Παλινδρόμηση παρέχει μοντέλα ερμηνεύσιμα.
- Η Λογιστική Παλινδρόμηση επιτυγχάνει ικανοποιητικές επιδόσεις κατηγοριοποίησης.

Μειονεκτήματα της Λογιστικής Παλινδρόμησης είναι τα εξής:

- Το βασικό μειονέκτημα της Λογιστικής Παλινδρόμησης είναι η διατύπωση αυθαίρετων υποθέσεων, όπως η ύπαρξη γραμμικής σχέσης με τον λογάριθμο του κλάσματος των πιθανοτήτων.
- Σύμφωνα με τα πολλά ερευνητικά αποτελέσματα, άλλες μέθοδοι, όπως τα Νευρωνικά Δίκτυα ή οι Μηχανές διανυσμάτων Υποστήριξης, επιτυγχάνουν τουλάχιστον εφάμιλλες ή και καλύτερες επιδόσεις κατηγοριοποίησης.

10.4 Σύνθετοι Κατηγοριοποιητές

Η κατηγοριοποίηση είναι ένα από τα βασικότερα αντικείμενα της Μηχανικής Μάθησης και της Εξόρυξης Δεδομένων. Σήμερα υπάρχουν διαθέσιμες αρκετές μέθοδοι κατηγοριοποίησης, ορισμένες από τις οποίες παρουσιάστηκαν στα πλαίσια του παρόντος και του προηγούμενου κεφαλαίου. Εκτός όμως από αυτήν την ποικιλία «ατομικών» και ριζικά διαφορετικών μεθόδων, υπάρχουν και οι λεγόμενες σύνθετες τεχνικές. Στους σύνθετους κατηγοριοποιητές γίνεται συνδυασμός μοντέλων ή μεθόδων. Αποτελέσματα ερευνητικών εργασιών παρέχουν ισχυρές ενδείξεις ότι οι σύνθετοι κατηγοριοποιητές μπορούν να επιτύχουν υψηλότερες επιδόσεις από τις ατομικές τεχνικές. Τα τελευταία χρόνια μάλιστα, οι σύνθετοι κατηγοριοποιητές γνωρίζουν μεγάλη άνθηση. Ενδεικτικά αναφέρουμε ότι ο Kirkos (2015), σε μια εργασία επισκόπησης βιβλιογραφίας σχετικά με την πρόβλεψη χρεοκοπίας με χρήση ευφυών τεχνικών, διαπιστώνει ότι στις είκοσι από τις συνολικά σαράντα δύο εργασίες εφαρμόστηκαν σύνθετοι κατηγοριοποιητές.

Ο συνδυασμός ατομικών τεχνικών είναι μια απαιτητική εργασία. Οι δυνατότητες συνδυασμού των βασικών μεθόδων είναι πάρα πολλές, και στη σχετική βιβλιογραφία προτείνονται συνεχώς νέες τεχνικές. Παρά τις πολλές διαφορές τους, οι σύνθετοι κατηγοριοποιητές μπορούν να χωριστούν σε δύο βασικές κατηγορίες:

- στους συνδυασμούς κατηγοριοποιητών,
- στους υβριδικούς κατηγοριοποιητές.

Στους **Συνδυασμούς Κατηγοριοποιητών** δημιουργείται ένας σχετικά μεγάλος αριθμός επιμέρους κατηγοριοποιητών. Όλοι αυτοί οι κατηγοριοποιητές εκτελούν την ίδια εργασία, δηλαδή δίνουν απαντήσεις στο ίδιο πρόβλημα κατηγοριοποίησης. Οι αποφάσεις των επιμέρους μοντέλων συναθροίζονται και προκύπτει η τελική απόφαση. Ένα παράδειγμα, το οποίο αναφέρεται συχνά για την καλύτερη κατανόηση των συνδυασμών κατηγοριοποιητών, είναι αυτό του ασθενή, ο οποίος επισκέπτεται πολλούς γιατρούς, συγκρίνει τις γνωματεύσεις τους και αποδέχεται την πλειοψηφούσα γνωμάτευση.

Προφανώς δεν έχει νόημα να αναπαραχθεί πολλές φορές ο ίδιος κατηγοριοποιητής. Κατά συνέπεια οι επιμέρους κατηγοριοποιητές πρέπει να διαφέρουν μεταξύ τους ουσιαστικά. Η επιθυμητή διαφοροποίηση μπορεί να επιτευχθεί με πολλούς τρόπους:

- **Εφαρμογή διαφορετικών ατομικών μεθόδων** και ανάπτυξη αντίστοιχων μοντέλων. Μπορούν να αναπτυχθούν μοντέλα Νευρωνικών Δικτύων, Δένδρων Αποφάσεων, Μηχανών Διανυσμάτων Υποστήριξης και άλλων μεθόδων, και να συνδυαστούν οι προβλέψεις τους. Κατά κανόνα όλα τα μοντέλα εκπαιδεύονται χρησιμοποιώντας τα ίδια δεδομένα.
- **Χρήση διαφορετικών δεδομένων εκπαίδευσης**. Εφαρμόζεται μόνο μια μέθοδος. Από το αρχικό

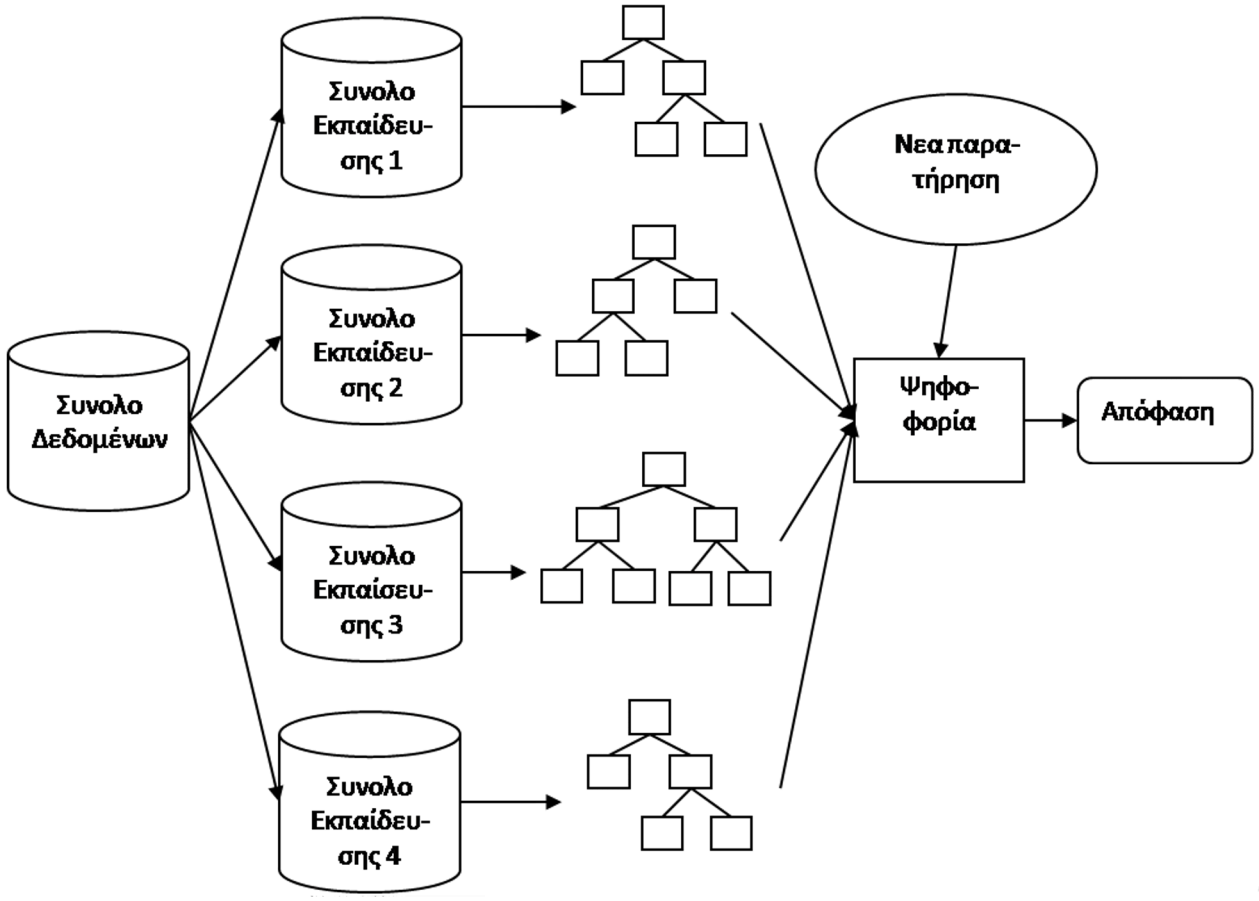
σύνολο δεδομένων δημιουργούνται πολλά σύνολα εκπαίδευσης. Η διαφοροποίηση των μοντέλων πηγάζει από τη χρήση διαφορετικών δεδομένων για την εκπαίδευση.

- **Χρήση διαφορετικών χώρων χαρακτηριστικών** (feature space). Επιλέγονται διαφορετικά σύνολα μεταβλητών εισόδου και εκπαιδεύονται αντίστοιχα μοντέλα. Τα διαφορετικά σύνολα μεταβλητών εισόδου μπορούν να προκύψουν από την εφαρμογή διαφορετικών μεθόδων επιλογής χαρακτηριστικών (feature selection).
- **Διαφορετική ρύθμιση παραμέτρων της ίδιας βασικής μεθόδου**. Εφαρμόζεται μια βασική μέθοδος και χρησιμοποιείται το ίδιο σύνολο δεδομένων για την εκπαίδευση των μοντέλων. Η διαφοροποίηση των μοντέλων προέρχεται από τη διαφορετική ρύθμιση των παραμέτρων. Για παράδειγμα, αν χρησιμοποιηθεί η μέθοδος των Νευρωνικών Δικτύων μπορούν να εκπαιδευτούν μοντέλα με διαφορετικές αρχιτεκτονικές, ρυθμούς εκπαίδευσης, εποχές κλπ.

Ένα βασικό ζήτημα στον σχεδιασμό συνδυασμού κατηγοριοποιητών είναι ο καθορισμός του τρόπου με τον οποίο ένας κατηγοριοποιητής επηρεάζει τους άλλους κατηγοριοποιητές. Οι κατηγοριοποιητές μπορούν να συνδυαστούν με σειριακό ή με παράλληλο τρόπο. Στον σειριακό τρόπο διεξάγεται μια επαναληπτική διαδικασία. Σε κάθε επανάληψη χρησιμοποιείται γνώση, η οποία αποκτήθηκε στο προηγούμενο στάδιο, και επηρεάζει την εκπαίδευση στο τρέχων στάδιο. Η αποκτηθείσα γνώση μπορεί να εκφραστεί σαν χειρισμός και τροποποίηση των δεδομένων εκπαίδευσης. Μια διαφορετική προσέγγιση είναι να χρησιμοποιηθεί ο κατηγοριοποιητής που δημιουργήθηκε σε ένα στάδιο για τη δημιουργία του νέου κατηγοριοποιητή στο επόμενο στάδιο.

Το πιο γνωστό **παράδειγμα σειριακού συνδυασμού** με ταυτόχρονο χειρισμό των δεδομένων εκπαίδευσης είναι η τεχνική **boosting.** Στον αλγόριθμο AdaBoost (Freund and Schapire, 1996), από ένα αρχικό σύνολο δεδομένων δημιουργείται ένα νέο σύνολο εκπαίδευσης, ίδιου μεγέθους με το αρχικό. Για τη δημιουργία του νέου συνόλου εκπαίδευσης εφαρμόζεται δειγματοληψία με επανατοποθέτηση. Αυτό σημαίνει ότι όταν επιλέγεται μια παρατήρηση και τοποθετείται στο νέο σύνολο εκπαίδευσης δεν απομακρύνεται από το αρχικό σύνολο δεδομένων. Κατά συνέπεια, στην επόμενη επιλογή είναι πιθανό να επιλεγεί πάλι η ίδια παρατήρηση. Με τον τρόπο αυτό, μια παρατήρηση μπορεί να συμμετέχει στο νέο σύνολο εκπαίδευσης πολλές φορές. Από το νέο σύνολο δεδομένων κατασκευάζεται ένας κατηγοριοποιητής. Σε κάθε παρατήρηση εκχωρείται ένας συντελεστής βαρύτητας. Εάν το μοντέλο κατηγοριοποιεί εσφαλμένα μια παρατήρηση τότε ο συντελεστής βαρύτητας αυξάνεται, διαφορετικά μειώνεται. Οι συντελεστές βαρύτητας χρησιμοποιούνται στο επόμενο στάδιο εκπαίδευσης, έτσι ώστε ο επόμενος κατηγοριοποιητής να δώσει περισσότερη προσοχή σε αυτές τις παρατηρήσεις. Με τον τρόπο αυτό, δημιουργείται μια σειρά διαδοχικών κατηγοριοποιητών. Κάθε κατηγοριοποιητής σχετίζεται με έναν συντελεστή βαρύτητας, ο οποίος είναι συνάρτηση της ακρίβειας του. Η τελική απόφαση λαμβάνεται από τις αποφάσεις των επιμέρους μοντέλων με ψηφοφορία και με χρήση των βαρών. Ο αλγόριθμος δίνει υψηλά ποσοστά ακρίβειας, υπάρχει όμως κίνδυνος υπερπροσαρμογής στις παρατηρήσεις που κατηγοριοποιούνται λανθασμένα.

Σύμφωνα με την προσέγγιση του **παράλληλου συνδυασμού κατηγοριοποιητών**, δημιουργούνται πολλαπλά σύνολα εκπαίδευσης από το αρχικό σύνολο δεδομένων. Για κάθε σύνολο εκπαίδευσης δημιουργείται ένα διαφορετικό μοντέλο. Η τελική απόφαση κατηγοριοποίησης μιας νέας παρατήρησης λαμβάνεται με κάποια μορφή συνάθροισης των προβλέψεων των επιμέρους μοντέλων. Ένας πολύ δημοφιλής αλγόριθμος παράλληλου συνδυασμού είναι ο **Bagging (**Breiman, 1996). Στον αλγόριθμο Bagging, τα σύνολα εκπαίδευσης δημιουργούνται με δειγματοληψία με επανατοποθέτηση. Για την κατηγοριοποίηση μιας νέας παρατήρησης γίνεται ψηφοφορία μεταξύ των μοντέλων, και η παρατήρηση εκχωρείται στην κλάση που συγκέντρωσε τις περισσότερες ψήφους. Σύμφωνα με τον Breiman, το Bagging λειτουργεί πολύ αποτελεσματικά με «ασταθείς» μεθόδους, όπου μικρές αλλαγές στο σύνολο εκπαίδευσης οδηγούν σε σημαντικά διαφορετικά μοντέλα. Παράδειγμα τέτοιας μεθόδου είναι τα Δένδρα Αποφάσεων. Η τεχνική Bagging παρουσιάζεται διαγραμματικά στο Σχήμα 10.6.



Σχήμα 10.6 *Bagging*

Ένα πολύ σημαντικό ζήτημα στην κατασκευή συνδυασμού κατηγοριοποιητών είναι ο καθορισμός του τρόπου με τον οποίο συνδυάζονται οι αποφάσεις των επιμέρους μοντέλων. Υπάρχουν δύο διαθέσιμες προσεγγίσεις, οι **απλές συνδυαστικές μέθοδοι** και οι **μετα-συνδυαστικές μέθοδοι**. Στις απλές συνδυαστικές μεθόδους, οι επιμέρους αποφάσεις συνδυάζονται σύμφωνα με κάποια συνάρτηση. Ο απλούστερος τρόπος είναι η απλή ψηφοφορία και η εκχώρηση της παρατήρησης στην πλειοψηφούσα κλάση. Ωστόσο, έχουν προταθεί και πολλά άλλα συνδυαστικά σχήματα, όπως η Άθροιση Κατανομής (Distribution Summation), η Καταμέτρηση Borda (Borda Count) και η Θεωρία Dempster-Shafer (Dempster Shafer Theory (DST)). Η Άθροιση Κατανομής συναθροίζει τις πιθανότητες ένταξης σε κάθε κλάση, τις οποίες υπολογίζει το εκάστοτε μοντέλο. Σύμφωνα με την καταμέτρηση Borda, οι επιμέρους κατηγοριοποιητές ταξινομούν τις υποψήφιες τιμές κλάσης σε σειρά προτεραιότητας. Κάθε θέση στην κλίμακα ταξινόμησης συσχετίζεται με έναν αριθμό «πόντων». Οι πόντοι της κάθε κλάσης συναθροίζονται, και η παρατήρηση εκχωρείται στην κλάση που συγκεντρώνει τους περισσότερους πόντους. Στην τεχνική DST επικρατεί η κλάση για την οποία μεγιστοποιείται η τιμή μιας συνάρτησης, η οποία χρησιμοποιεί τις πιθανότητες που υπολογίζουν οι βασικοί κατηγοριοποιητές. Οι μετα-συνδυαστικές μέθοδοι χρησιμοποιούν τους βασικούς κατηγοριοποιητές και τις προβλέψεις τους για περαιτέρω μάθηση. Στη μέθοδο Stack Generalization (Wolpert, 1992) τα αποτελέσματα των βασικών κατηγοριοποιητών χρησιμοποιούνται ως είσοδος από τους κατηγοριοποιητές του επόμενου επιπέδου.

Οι **Υβριδικοί Κατηγοριοποιητές** είναι η δεύτερη μεγάλη κατηγορία των σύνθετων κατηγοριοποιητών. Στα υβριδικά συστήματα, όπως και στους συνδυασμούς κατηγοριοποιητών, χρησιμοποιούνται διάφορες τεχνικές. Ωστόσο υπάρχουν σημαντικές διαφορές σε σχέση με τις μεθόδους συνδυασμού κατηγοριοποιητών. Η πρώτη διαφορά είναι ότι εφαρμόζονται ετερογενείς τεχνικές, κάθε μια από τις οποίες επιλύει ένα διαφορετικό πρόβλημα. Η δεύτερη διαφορά είναι ότι η τελική απόφαση κατηγοριοποίησης λαμβάνεται από έναν μόνο κατηγοριοποιητή, ενώ στους συνδυασμούς κατηγοριοποιητών γίνεται συνδυασμός των αποφάσεων πολλών κατηγοριοποιητών.

Οι Lin, Hu and Tsai (2012) ορίζουν τους ακόλουθους τρεις τύπους υβριδικών κατηγοριοποιητών:

- διαδοχικές τεχνικές,

- συνδυασμός κατηγοριοποίησης και ανάλυσης συστάδων,
- ολοκλήρωση δύο τεχνικών με συμπληρωματικό τρόπο.

Στις διαδοχικές υβριδικές τεχνικές πραγματοποιείται μια επεξεργασία σε ένα πρώτο στάδιο και στη συνέχεια, τα αποτελέσματα αυτής της επεξεργασίας χρησιμοποιούνται ως είσοδος στο επόμενο στάδιο. Για παράδειγμα, στο πρώτο στάδιο μπορεί να υπολογίζονται κάποιες τιμές, οι οποίες θα χρησιμοποιηθούν στο επόμενο στάδιο ως πρόσθετα δεδομένα εισόδου. Μια διασταλτική ερμηνεία του όρου των διαδοχικών υβριδικών τεχνικών θα μπορούσε να αποδεχτεί τη μείωση της διαστασιμότητας ή του πλήθους των παρατηρήσεων εκπαίδευσης με την εφαρμογή μεθόδων soft computing, και την ακόλουθη ανάπτυξη ενός κατηγοριοποιητή, ως περίπτωση υβριδικών μοντέλων.

Ο δεύτερος τύπος υβριδικών κατηγοριοποιητών είναι ο συνδυασμός τεχνικών κατηγοριοποίησης με τεχνικές ανάλυσης συστάδων. Οι τεχνικές ανάλυσης συστάδων μπορούν να εφαρμοστούν για τον εντοπισμό και την απομάκρυνση εξαιρέσεων και παρατηρήσεων με ακραίες τιμές. Επίσης, η ανάλυση συστάδων μπορεί να εντοπίσει ομάδες ομοειδών παρατηρήσεων, οι οποίες θα θεωρηθούν κλάσεις και θα χρησιμοποιηθούν για περαιτέρω κατηγοριοποίηση.

Πιθανώς, η πιο γνήσια εκδοχή υβριδικών συστημάτων είναι αυτά τα οποία ολοκληρώνουν δύο ετερογενείς μεθόδους σε μια ενιαία διαδικασία εκπαίδευσης. Μια διαδεδομένη τεχνική, για την ανάπτυξη υβριδικών κατηγοριοποιητών αυτού του τύπου, είναι ο συνδυασμός Εξελικτικών Αλγορίθμων με μεθόδους κατηγοριοποίησης. Εξελικτικοί αλγόριθμοι που εφαρμόζονται συχνά για τη δημιουργία υβριδικών κατηγοριοποιητών είναι οι Γενετικοί Αλγόριθμοι (Genetic Algorithms) και η Βελτιστοποίηση Σμήνους Σημείων (Particle Swarm Optimization). Οι Γενετικοί Αλγόριθμοι παρουσιάζονται στο Κεφάλαιο 3. Η Βελτιστοποίηση Σμήνους Σημείων είναι μια τεχνική, που προσομοιάζει την κίνηση σμήνους πτηνών ή κοπαδιού ψαριών, με στόχο την εύρεση της βέλτιστης θέσης μέσα στο σμήνος.

Οι εξελικτικοί αλγόριθμοι μπορούν να εφαρμοστούν με διάφορους τρόπους. Η απλούστερη τεχνική είναι να χρησιμοποιηθούν για τη ρύθμιση των παραμέτρων άλλων μεθόδων. Για παράδειγμα, μπορεί να δημιουργηθεί ένας πληθυσμός Νευρωνικών Δικτύων με διαφορετικές αρχιτεκτονικές, ρυθμούς μάθησης, εποχές εκπαίδευσης κλπ. και με τη χρήση των εξελικτικών αλγορίθμων να επιλεγεί το δίκτυο με την καλύτερη ρύθμιση παραμέτρων. Άλλη εκδοχή είναι να ενσωματωθούν οι εξελικτικοί αλγόριθμοι στη διαδικασία εκπαίδευσης του κατηγοριοποιητή. Μια τρίτη εκδοχή είναι να γίνει επιλογή σημαντικών χαρακτηριστικών και ταυτόχρονη ρύθμιση παραμέτρων. Με τον τρόπο αυτό, επιτυγχάνεται κατάλληλη ρύθμιση των παραμέτρων για τις συγκεκριμένες μεταβλητές εισόδου. Τέλος, σε αυτήν την κατηγορία των υβριδικών κατηγοριοποιητών ανήκουν και τα Νεύρω-Ασαφή συστήματα, τα οποία συνδυάζουν την Ασαφή Λογική με τα Νευρωνικά Δίκτυα.

Στην πρόσφατη έρευνα έχουν προταθεί σχήματα, τα οποία επιτρέπουν τη συναρμογή συνδυασμών κατηγοριοποιητών και υβριδικών τεχνικών. Εξελικτικοί αλγόριθμοι έχουν εφαρμοστεί για τη βελτιστοποίηση συνδυασμών κατηγοριοποιητών. Στους συνδυασμούς κατηγοριοποιητών μια σημαντική ιδιότητα, η οποία επηρεάζει και την επίδοση, είναι η ουσιαστική διαφοροποίηση των επιμέρους μοντέλων. Οι Γενετικοί Αλγόριθμοι μπορούν να χρησιμοποιηθούν για να επιλέξουν από μια δεξαμενή διαθέσιμων βασικών κατηγοριοποιητών εκείνους τους κατηγοριοποιητές που παρουσιάζουν αυξημένη διαφοροποίηση.

10.5 Επικύρωση Κατηγοριοποιητών

Σύμφωνα με ότι έχει αναφερθεί μέχρι τώρα, οι μέθοδοι κατηγοριοποίησης επεξεργάζονται ένα σύνολο παρατηρήσεων και εκπαιδεύουν ένα μοντέλο. Το μοντέλο συνίσταται σε έναν μηχανισμό λήψης απόφασης, για το εάν οι παρατηρήσεις του δείγματος ανήκουν σε μια κλάση. Η ακρίβεια του μοντέλου μέχρι στιγμής έχει οριστεί σε σχέση με την ικανότητα του να κατατάσσει σωστά τις παρατηρήσεις του συνόλου εκπαίδευσης. **Το ερώτημα που προκύπτει είναι τι θα συμβεί εάν το μοντέλο αντιμετωπίσει «άγνωστες» παρατηρήσεις**, παρατηρήσεις δηλαδή που δεν ανήκουν στο σύνολο εκπαίδευσης. Ουσιαστικά, το ερώτημα που τίθεται είναι εάν το μοντέλο ενσωματώνει γενικευμένους κανόνες ευρύτερης ισχύος, οι οποίοι καθορίζουν τον προσδιορισμό της κλάσης μιας παρατήρησης στον πραγματικό κόσμο, ή εάν το μοντέλο ενσωματώνει εξειδικευμένους κανόνες, που καθορίζουν τον προσδιορισμό της κλάσης των παρατηρήσεων του συγκεκριμένου συνόλου. Προφανώς η πραγματική αξία ενός μοντέλου βρίσκεται στην ικανότητα του να προβλέπει την κλάση άγνωστων παρατηρήσεων του πραγματικού κόσμου.

Οι αυξημένες επιδόσεις ενός μοντέλου έναντι του συνόλου εκπαίδευσης δεν συνεπάγονται και αυξημένη ικανότητα κατηγοριοποίησης άγνωστων παρατηρήσεων. Στο Κεφάλαιο 9 έγινε αναφορά στο πρόβλημα της υπερπροσαρμογής των μοντέλων (data overfitting). Η υπερπροσαρμογή παρουσιάζεται όταν ένα μοντέλο εί-

ναι υπερβολικά περίπλοκο. Το μοντέλο αυτό είναι ικανό να αφομοιώσει τις ιδιαιτερότητες των δεδομένων εκπαίδευσης, αντί να καταγράψει σχέσεις γενικότερης ισχύος. Το αποτέλεσμα της υπερπροσαρμογής είναι ιδιαίτερα ψηλές επιδόσεις έναντι του συνόλου εκπαίδευσης, αλλά δυσανάλογα χαμηλές επιδόσεις έναντι άγνωστων παρατηρήσεων.

Για τους παραπάνω λόγους, η ακρίβεια ενός μοντέλου πρέπει να εκτιμάται έναντι άγνωστων παρατηρήσεων. Ο καθορισμός της ακρίβειας ενός μοντέλου είναι ιδιαίτερα σημαντικός, γιατί μας επιτρέπει να αποφανθούμε εάν το μοντέλο μπορεί να χρησιμοποιηθεί για τη λήψη αποφάσεων στον πραγματικό κόσμο. Επίσης, μας επιτρέπει να συγκρίνουμε διαφορετικά μοντέλα, ώστε να επιλέξουμε το καλύτερο. Για την εκτίμηση τις ικανότητας ενός μοντέλου να προβλέπει άγνωστες παρατηρήσεις έχουν προταθεί διάφορες μέθοδοι, όπως η διάσπαση του δείγματος σε δείγμα εκπαίδευσης και δείγμα επικύρωσης (holdout method), η επικύρωση 10 τμημάτων (10-fold cross validation), η μέθοδος «άφησε ένα έξω» (leave one out) και τέλος, η μέθοδος bootstrapping.

Μέθοδος Holdout. Κατά τη μέθοδο holdout, το σύνολο δεδομένων διασπάται σε δύο υποσύνολα, κάθε ένα από τα οποία περιέχει διαφορετικές παρατηρήσεις. Το ένα υποσύνολο χρησιμοποιείται για την εκπαίδευση του μοντέλου και ονομάζεται **σύνολο εκπαίδευσης** (training set). Αφού ολοκληρωθεί η εκπαίδευση, το μοντέλο αποπειράται να προβλέψει την κλάση των παρατηρήσεων του δεύτερου υποσυνόλου, και ακολούθως συγκρίνονται οι προβλέψεις του μοντέλου με την πραγματική κλάση των παρατηρήσεων. Το δεύτερο υποσύνολο είναι γνωστό ως **σύνολο επικύρωσης** (validation set ή holdout set). Μια ενδεδειγμένη πρακτική είναι να χρησιμοποιούνται τα δύο τρίτα του αρχικού συνόλου ως σύνολο εκπαίδευσης και το ένα τρίτο ως σύνολο επικύρωσης. Η επίδοση του μοντέλου είναι το ποσοστό των ορθών προβλέψεων. Μια παραλλαγή της μεθόδου holdout είναι η μέθοδος της τυχαίας υποδειγματοληψίας (random subsampling). Σύμφωνα με τη μέθοδο αυτή, γίνεται επανάληψη της μεθόδου holdout πολλές φορές. Σε κάθε επανάληψη δημιουργούνται νέα σύνολα εκπαίδευσης και επικύρωσης, εφαρμόζοντας τυχαία δειγματοληψία.

Διασταυρούμενη Επικύρωση 10 τμημάτων. Στη μέθοδο επικύρωσης 10 τμημάτων (10 fold cross validation) το σύνολο δεδομένων διαιρείται σε 10 υποσύνολα. Κάθε υποσύνολο περιέχει διαφορετικές παρατηρήσεις. Η επιλογή των υποσυνόλων είναι τυχαία. Ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο επικύρωσης και τα υπόλοιπα εννέα συνενώνονται και δημιουργούν το σύνολο εκπαίδευσης. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης και δοκιμάζεται έναντι του συνόλου επικύρωσης. Η διαδικασία επαναλαμβάνεται δέκα φορές, κάθε φορά χρησιμοποιώντας ένα διαφορετικό σύνολο ως σύνολο επικύρωσης και τα υπόλοιπα εννέα ως σύνολο εκπαίδευσης. Στο τέλος υπολογίζεται η μέση επίδοση του μοντέλου. Η μέθοδος μπορεί να διαφοροποιηθεί ως προς το πλήθος των τμημάτων. Γενικότερα ονομάζεται μέθοδος επικύρωσης *k* τμημάτων, όπου *k* συμβολίζει τον αριθμό των δημιουργημένων υποσυνόλων και των επαναλήψεων. Μια άλλη εκδοχή της μεθόδου είναι η **στρωματοποιημένη** επικύρωση 10 τμημάτων (stratified 10 fold cross validation). Σύμφωνα με αυτήν την εκδοχή, κάθε υποσύνολο περιέχει περίπου ίσο αριθμό παρατηρήσεων για την κάθε κλάση.

Μέθοδος «άφησε ένα έξω». Η μέθοδος «άφησε ένα έξω» (leave one out) ουσιαστικά αποτελεί παραλλαγή της μεθόδου επικύρωσης *k* τμημάτων. Στην παραλλαγή αυτή το *k=n*, όπου n είναι το πλήθος των παρατηρήσεων, οι οποίες απαρτίζουν το σύνολο δεδομένων. Για κάθε μια παρατήρηση, το μοντέλο εκπαιδεύεται χρησιμοποιώντας τις υπόλοιπες *n-1* παρατηρήσεις και επικυρώνεται έναντι της επιλεγμένης παρατήρησης. Η διαδικασία επαναλαμβάνεται *n* φορές. Στο τέλος υπολογίζεται το ποσοστό ορθών παρατηρήσεων.

Μέθοδος bootstrap. Στη μέθοδο bootstrap δημιουργούνται πάλι πολλαπλά σύνολα επικύρωσης με δειγματοληψία. Η διαφορά έγκειται στο γεγονός ότι η δειγματοληψία γίνεται με επανατοποθέτηση. Κάθε παρατήρηση που επιλέγεται να συμμετάσχει στο δείγμα επικύρωσης δεν αφαιρείται από το αρχικό δείγμα. Κατά τον τρόπο αυτό, μια παρατήρηση μπορεί να επιλεγεί περισσότερες από μία φορές για να συμμετάσχει στο ίδιο σύνολο επικύρωσης.

Ο Kohavi (1995) πραγματοποίησε μια συγκριτική μελέτη σχετικά με τις μεθόδους επικύρωσης κατηγοριοποιητών. Σύμφωνα με τα αποτελέσματα του, η πλέον κατάλληλη μέθοδος είναι η στρωματοποιημένη επικύρωση 10 τμημάτων.

10.6 Ανισοκατανομή κλάσεων και κόστος σφάλματος

Το ποσοστό των επιτυχών προβλέψεων ενός κατηγοριοποιητή είναι ένα ισχυρό μέτρο της ικανότητας πρόβλεψης, σε ορισμένες περιπτώσεις όμως δεν είναι επαρκές. Σε πολλά προβλήματα του πραγματικού κόσμου **οι παρατηρήσεις δεν είναι ισομερώς κατανεμημένες στις διάφορες κλάσεις**. Το φαινόμενο είναι πολύ συνηθισμένο σε ιατρικά δεδομένα, όπου η πιθανότητα εμφάνισης μιας ασθένειας είναι μικρή. Σε μια βάση δεδομένων με αποτελέσματα εξετάσεων, ένα μικρό ποσοστό των παρατηρήσεων θα αφορά περιπτώσεις ασθένειας.

Σε οικονομικά δεδομένα, μια από τις χαρακτηριστικότερες περιπτώσεις ανισοκατανομής των κλάσεων είναι η πρόβλεψη χρεοκοπίας. Στις ΗΠΑ το ποσοστό αποτυχίας των επιχειρήσεων είναι περίπου 2%. Αυτό σημαίνει ότι ένας κατηγοριοποιητής, ο οποίος προβλέπει πάντα μη χρεοκοπία, θα είχε ακρίβεια 98%. Η ακρίβεια του μοντέλου είναι εξαιρετικά υψηλή, στην πραγματικότητα όμως το μοντέλο είναι απολύτως αποτυχημένο, αφού αδυνατεί να προβλέψει τις περιπτώσεις χρεοκοπίας, που είναι και το ζητούμενο. Σε σύνολα δεδομένων όπου οι παρατηρήσεις δεν είναι ισομερώς κατανεμημένες στις κλάσεις, το γενικό ποσοστό επιτυχών προβλέψεων δεν επαρκεί για την εκτίμηση των δυνατοτήτων του κατηγοριοποιητή. Επίσης, η ανισοκατανομή των κλάσεων έχει επιπτώσεις στην εκπαίδευση των κατηγοριοποιητών. Μελέτες έχουν δείξει ότι τα παραγόμενα μοντέλα τείνουν να προβλέπουν καλύτερα την πλειοψηφούσα κλάση και σημαντικά χειρότερα τη μειοψηφούσα κλάση (Weiss & Provost, 2003).

Η φυσική κατανομή των παρατηρήσεων σε κλάσεις συχνά δεν είναι η καλύτερη κατανομή για την εκπαίδευση ενός κατηγοριοποιητή. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί διάφορες τεχνικές επαναδειγματοληψίας (resampling). Η τυχαία υποδειγματοληψία (random undersampling) είναι μια τεχνική, η οποία απομακρύνει με τυχαίο τρόπο παρατηρήσεις της πλειοψηφούσας κλάσης, μέχρι να επιτευχθεί ίσο πλήθος παρατηρήσεων για την κάθε κλάση. Μειονέκτημα αυτής της τεχνικής είναι ότι διαγράφονται παρατηρήσεις, οι οποίες πιθανώς περιείχαν ουσιαστική πληροφορία. Η τυχαία υπερδειγματοληψία (random oversampling) αναπαράγει με τυχαίο τρόπο τις παρατηρήσεις της μειοψηφούσας κλάσης, μέχρι να ισοσταθμιστεί το πλήθος των κλάσεων. Μειονέκτημα της τεχνικής είναι ότι μπορεί να οδηγήσει σε υπερπροσαρμογή. Για την αντιμετώπιση αυτών των προβλημάτων έχουν προταθεί τεχνικές, οι οποίες επιλεκτικά αναπαράγουν παρατηρήσεις της μειοψηφούσας κλάσης ή/και διαγράφουν παρατηρήσεις της πλειοψηφούσας κλάσης (Kubat & Martin, 1997; Japkowicz, 2000; Chawla, Boywer, Hall & Kegelmeyer, 2002).

Μια άλλη περίπτωση όπου το γενικό ποσοστό επιτυχών προβλέψεων δεν επαρκεί, είναι όταν οι **αποτυχίες πρόβλεψης διαφορετικών κλάσεων δεν έχουν το ίδιο κόστος**. Ας επανέρθουμε στο παράδειγμα πρόβλεψης χρεοκοπίας. Οι περιπτώσεις εσφαλμένων προβλέψεων είναι οι εξής δύο: α) Μια επιχείρηση που θα χρεοκοπήσει κατηγοριοποιείται ως βιώσιμη. β) Μια επιχείρηση που δεν θα χρεοκοπήσει κατηγοριοποιείται ως χρεοκοπημένη. Η πρώτη περίπτωση είναι ένα σφάλμα Τύπου Ι. Η δεύτερη περίπτωση είναι ένα σφάλμα Τύπου ΙΙ. Ποιο είναι το κόστος των σφαλμάτων για έναν τραπεζικό οργανισμό; Στην περίπτωση σφάλματος Τύπου ΙΙ, η τράπεζα δεν θα ενέκρινε ένα επιτυχημένο δάνειο και θα έχανε το σχετικό κέρδος. Στην περίπτωση σφάλματος Τύπου Ι, η τράπεζα θα ενέκρινε ένα αποτυχημένο δάνειο και θα έχανε το αντίστοιχο κεφάλαιο. Είναι προφανές ότι τα δύο σφάλματα έχουν διαφορετικό κόστος. Στον πραγματικό κόσμο είναι δύσκολο να βρεθούν προβλήματα όπου το κόστος διαφορετικού τύπου σφαλμάτων είναι το ίδιο (Witten & Frank, 2000).

Οι μέθοδοι κατηγοριοποίησης είναι σχεδιασμένες έτσι ώστε να ελαχιστοποιούν τον ρυθμό σφάλματος, δηλαδή τον συνολικό αριθμό εσφαλμένων προβλέψεων ή την πιθανότητα εσφαλμένων προβλέψεων. Η προσέγγιση αυτή υποθέτει ότι το κόστος διαφορετικών τύπων σφάλματος είναι το ίδιο. Σε ρεαλιστικά προβλήματα όμως αυτό που συχνά έχει σημασία είναι η μείωση του κόστους εσφαλμένων κατηγοριοποιήσεων, και όχι η αύξηση του ρυθμού ακρίβειας. Για μια τράπεζα σημασία έχει η λήψη επικερδών αποφάσεων και όχι η διατύπωση πολλών επιτυχών προβλέψεων.

Η εκπαίδευση μοντέλων με τρόπο τέτοιο ώστε να μειώνεται το συνολικό κόστος εσφαλμένων προβλέψεων ονομάζεται **ευαίσθητη ως προς το κόστος εκπαίδευση** (cost sensitive learning). Για εκπαίδευση ευαίσθητη ως προς το κόστος, απαιτείται καταρχάς ο καθορισμός του συνολικού σφάλματος. Έχουν προταθεί διάφοροι ορισμοί για το συνολικό κόστος σφάλματος. Για περιπτώσεις δυαδικής κλάσης, οι Chen, Huang and Lin (2009) το ορίζουν σύμφωνα με τη Σχέση 10.28

$$MC = (k_1 * T1Err + k_2 * T2Err)$$

(10.28)

όπου *T1Err* είναι το πλήθος σφαλμάτων Τύπου Ι, *T2Err* το πλήθος σφαλμάτων Τύπου ΙΙ, και k_1, k_2 το κόστος σφάλματος Τύπου Ι και ΙΙ αντίστοιχα. Οι Chen, Ribeiro, Vieira, Duarte and Neves (2011) ορίζουν το συνολικό κόστος σφάλματος σύμφωνα με τη Σχέση 10.29.

$$MC = \frac{k_1}{k_1 + k_2} * T1Err + \frac{k_2}{k_1 + k_2} * T2Err$$

(10.29)

Πρόσθετοι ορισμοί έχουν προταθεί από άλλους ερευνητές. Ένα ανοικτό ζήτημα σε κάθε πρόβλημα εκπαίδευσης ευαίσθητης ως προς το κόστος, είναι ο καθορισμός της αναλογίας του κόστους διαφορετικών σφαλμάτων k_1/k_2. Για τον καθορισμό της αναλογίας απαιτείται καλή γνώσης του πεδίου εφαρμογής και η γνώμη ειδικών. Σε πολλά προβλήματα δεν υπάρχει μια αναγνωρισμένη και γενικώς αποδεκτή αναλογία. Για παράδειγμα, σε εργασίες πρόβλεψης χρεοκοπίας έχουν χρησιμοποιηθεί αναλογίες, οι οποίες κυμαίνονται από *1/1* έως και *1/100*.

Στην ευαίσθητη ως προς το κόστος εκπαίδευση, τα μοντέλα εκπαιδεύονται με τέτοιο τρόπο ώστε να μειωθεί το κόστος των εσφαλμένων κατηγοριοποιήσεων. Μια τέτοιου τύπου εκπαίδευση μπορεί να επιτευχθεί με παρεμβάσεις είτε στο επίπεδο των δεδομένων είτε στο επίπεδο του αλγορίθμου. Σύμφωνα με την πρώτη προσέγγιση πραγματοποιείται επαναδειγματοληψία, ώστε να μεταβληθεί κατάλληλα η αναλογία ανάμεσα στις παρατηρήσεις που ανήκουν στην «ακριβή» κλάση και στις παρατηρήσεις που ανήκουν στη «φθηνή» κλάση. Ο Elkan (2001) ασχολείται με την περίπτωση της δυαδικής κλάσης, και αναφέρεται σε τρόπους αλλαγής της αναλογίας θετικών και αρνητικών παρατηρήσεων, έτσι ώστε να επιτευχθεί εκπαίδευση με μείωση του κόστους σε μοντέλα που εκπαιδεύονται με μεθόδους, οι οποίες δεν είναι ευαίσθητες ως προς το κόστος. Σύμφωνα με τη δεύτερη προσέγγιση, η πληροφορία για το κόστος ενσωματώνεται και αξιοποιείται στον αλγόριθμο εκπαίδευσης. Παράδειγμα τέτοιας τεχνικής υπάρχει στην εργασία των Chen et al. (2011). Στην εργασία αυτή οι ερευνητές συνδυάζουν τους Γενετικούς Αλγορίθμους με τη μέθοδο κατηγοριοποίησης Learning Vector Quantization, και ενσωματώνουν το κόστος στη συνάρτηση καταλληλότητας (fitness function).

10.7 Επιδόσεις ανά κλάση

Σε περιπτώσεις όπου οι κλάσεις δεν είναι ισομερώς κατανεμημένες ή όπου οι εσφαλμένες κατηγοριοποιήσεις διαφορετικών κλάσεων έχουν διαφορετικό κόστος, είναι σημαντική η εκτίμηση της ικανότητας πρόβλεψης του κατηγοριοποιητή για την κάθε κλάση.

Για να εκτιμήσουμε τις ανά κλάση επιδόσεις ενός κατηγοριοποιητή, εισάγουμε την αναγκαία ορολογία. Για την περίπτωση δυαδικής κλάσης ισχύουν οι ακόλουθοι όροι:

Θετικές παρατηρήσεις (positive) ονομάζονται οι παρατηρήσεις, οι οποίες ανήκουν σε μια τιμή της κλάσης (πχ χρεοκοπία)

Αρνητικές παρατηρήσεις (negative) ονομάζονται οι παρατηρήσεις, οι οποίες ανήκουν στην άλλη τιμή της κλάσης (πχ μη χρεοκοπία)

Αληθινές Θετικές Προβλέψεις (true positive – tp) είναι το πλήθος των επιτυχών προβλέψεων για θετικές παρατηρήσεις (πχ η επιχείρηση είναι χρεοκοπημένη και ο κατηγοριοποιητής προβλέπει σωστά την κλάση).

Αληθινές Αρνητικές Προβλέψεις (true negative – tn) είναι το πλήθος των επιτυχημένων προβλέψεων για αρνητικές παρατηρήσεις (πχ η επιχείρηση δεν είναι χρεοκοπημένη και ο κατηγοριοποιητής προβλέπει σωστά την κλάση).

Ψευδείς Θετικές Προβλέψεις (false positive – fp) είναι το πλήθος των αποτυχημένων προβλέψεων για αρνητικές παρατηρήσεις (η επιχείρηση δεν είναι χρεοκοπημένη, ο κατηγοριοποιητής όμως την προβλέπει ως χρεοκοπημένη).

Ψευδείς Αρνητικές Προβλέψεις (false negative – fn) είναι το πλήθος των αποτυχημένων προβλέψεων για θετικές παρατηρήσεις (η επιχείρηση είναι χρεοκοπημένη, ο κατηγοριοποιητής όμως την προβλέπει ως μη χρεοκοπημένη).

Ένας τρόπος παρουσίασης των επιδόσεων ανά κλάση ενός κατηγοριοποιητή είναι με τη χρήση του **πίνακα σύγχυσης** (confusion matrix). Ο Πίνακας Σύγχυσης είναι ένας δισδιάστατος πίνακας, όπου οι στήλες αντιστοιχούν στις προβλέψεις και οι γραμμές στις πραγματικές τιμές κλάσης. Στα κελιά του πίνακα αναγράφονται οι αληθινές θετικές, οι αληθινές αρνητικές, οι ψευδείς θετικές και οι ψευδείς αρνητικές προβλέψεις. Στο Σχήμα 10.7 απεικονίζεται ένας Πίνακας Σύγχυσης.

	Πρόβλεψη Αρνητικής Κλάσης	Πρόβλεψη Θετικής Κλάσης
Πραγματική Αρνητική Κλάση	tn	fp
Πραγματική Θετική Κλάση	fn	tp

tn = true negative
tp = true positive
fn = false negative
fp = false positive

Σχήμα 10.7 *Πίνακας Σύγχυσης*

Ορισμένα πρόσθετα μέτρα για τις επιδόσεις ενός κατηγοριοποιητή είναι τα ακόλουθα:

$$sensitivity = \frac{tp}{pos} \tag{10.30}$$

$$specificity = \frac{tn}{negat} \tag{10.31}$$

$$precision = \frac{tp}{tp + fp} \tag{10.32}$$

$$accuracy = sensitivity * \frac{pos}{pos + negat} + specificity * \frac{neg}{pos + negat} = \frac{tp + tn}{pos + negat} \tag{10.33}$$

όπου pos είναι το πλήθος των θετικών παρατηρήσεων και negat είναι το πλήθος των αρνητικών παρατηρήσεων. Σύμφωνα με τα παραπάνω, η ακρίβεια (accuracy) ορίζεται ως το ποσοστό των ορθών θετικών προβλέψεων επί το ποσοστό των θετικών παρατηρήσεων συν το ποσοστό των ορθών αρνητικών προβλέψεων επί το ποσοστό των αρνητικών παρατηρήσεων ή ισοδύναμα ως το πλήθος των ορθών προβλέψεων προς το πλήθος των παρατηρήσεων.

10.8 Καμπύλες ROC

Ένα ισχυρό μέτρο για την εκτίμηση της ανά κλάση ακρίβειας του κατηγοριοποιητή είναι οι λεγόμενες καμπύλες ROC (Receiver Operating Characteristics). Οι καμπύλες ROC σχεδιάζονται σε έναν δυσδιάστατο επίπεδο χώρο. Ο οριζόντιος άξονας εκφράζει το μέγεθος 1-specificity, το οποίο ονομάζεται και False Positive Rate.

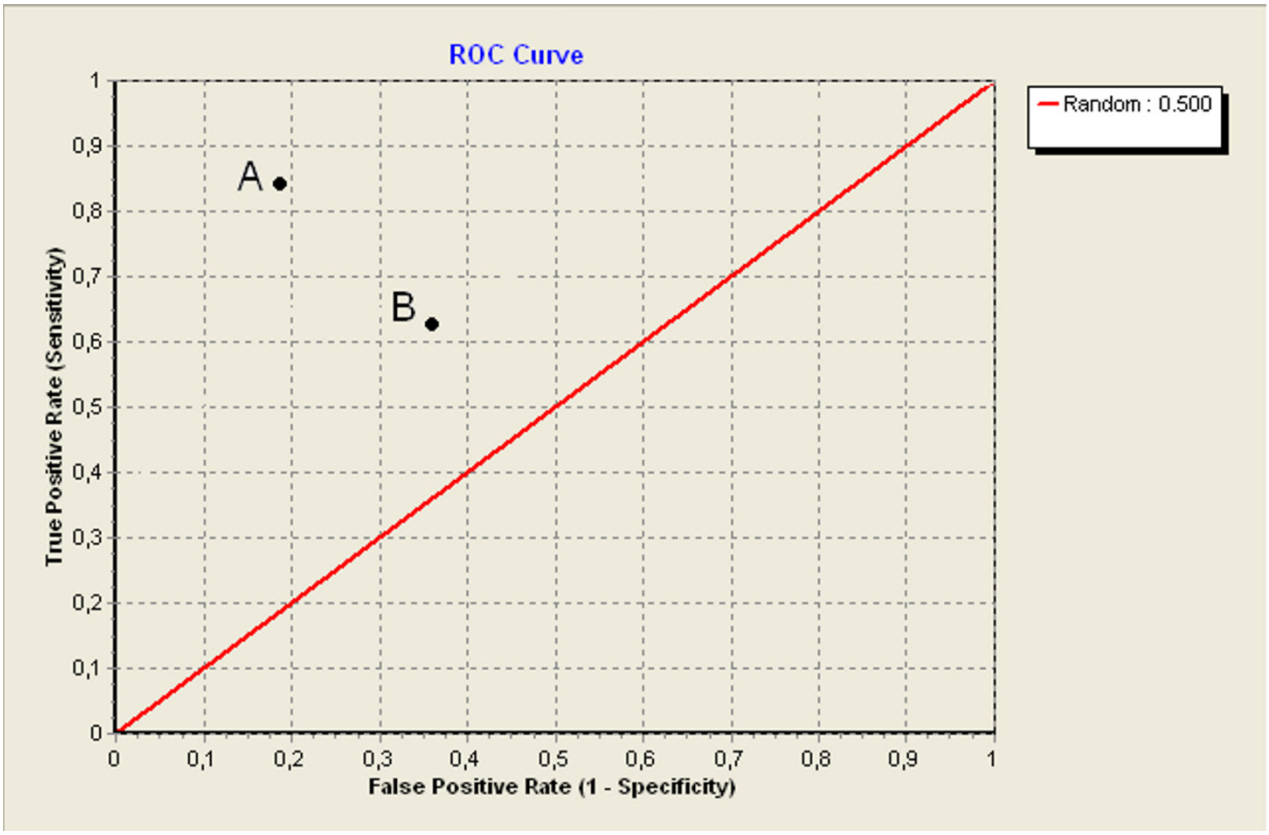
$$False\_Positive\_Rate = 1 - Specificity = \frac{fp}{negat} \tag{10.34}$$

Ο κατακόρυφος άξονας εκφράζει το μέγεθος sensitivity, το οποίο ονομάζεται και True Positive Rate

$$True\_Positive\_Rate = Sensitivity = \frac{tp}{pos}$$

(10.35)

Ουσιαστικά, ο οριζόντιος άξονας εκφράζει το ποσοστό των αρνητικών παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν λάθος, και ο κατακόρυφος άξονας εκφράζει το ποσοστό των θετικών παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν σωστά. Το Σχήμα 10.8 απεικονίζει τον δυσδιάστατο χώρο καμπύλων ROC. Κάθε σημείο του χώρου αυτού εκφράζει ένα ισοζύγιο ανάμεσα στο ποσοστό ορθών θετικών προβλέψεων και εσφαλμένων θετικών προβλέψεων. Το σημείο 0,0 είναι ένας κατηγοριοποιητής, που δεν προβλέπει ποτέ θετική παρατήρηση. Το σημείο 1,1 είναι ένας κατηγοριοποιητής, που προβλέπει πάντα θετική παρατήρηση. Η διαγώνια γραμμή, από το σημείο 0,0 στο σημείο 1,1 είναι ένας κατηγοριοποιητής που προβλέπει τυχαία την κλάση. Οι κατηγοριοποιητές που βρίσκονται κάτω από τη διαγώνια γραμμή είναι χειρότεροι από την τυχαία πρόβλεψη. Οι κατηγοριοποιητές που βρίσκονται πάνω από τη διαγώνια γραμμή είναι καλύτεροι από την τυχαία πρόβλεψη. Το σημείο 0,1 είναι ο άριστος κατηγοριοποιητής, οι οποίος προβλέπει σωστά όλες τις θετικές και αρνητικές παρατηρήσεις. Γενικώς, όσο πιο μετατοπισμένο είναι προς τα επάνω και προς τα αριστερά ένα σημείο, τόσο καλύτερη θεωρείται η επίδοση. Στο Σχήμα 10.8 το σημείο A είναι καλύτερο από το σημείο B.

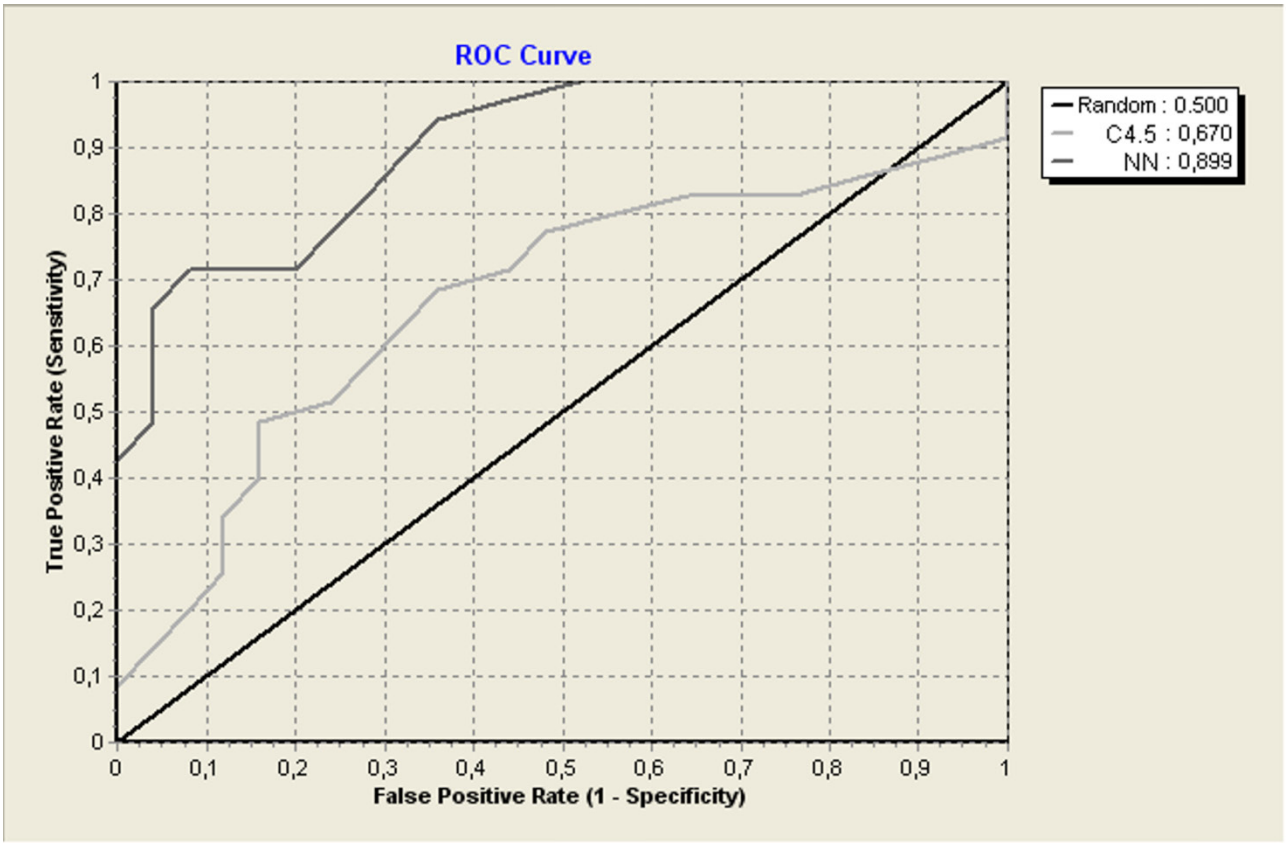


Σχήμα 10.8 *Σημεία στον χώρο καμπύλων ROC*

Η επίδοση των κατηγοριοποιητών στον χώρο ROC συμβολίζεται με μία καμπύλη. Για να συγκρίνουμε κατηγοριοποιητές χρειαζόμαστε ένα μέτρο σύγκρισης. Τέτοιο μέτρο σύγκρισης είναι η Περιοχή Κάτω από την Καμπύλη ROC (Area Under ROC Curve (AUC)). Η AUC εκφράζει το ποσοστό του χώρου που βρίσκεται κάτω από την καμπύλη, και παίρνει τιμές από 0 έως 1. Η διαγώνια γραμμή τυχαίας πρόβλεψης έχει *AUC = 0,5*. Συνεπώς, κάθε κατηγοριοποιητής καλύτερος της τυχαίας πρόβλεψης έχει *AUC>0,5*. Όσο μεγαλύτερη περιοχή AUC έχει ένας κατηγοριοποιητής τόσο καλύτερος είναι. Στο Σχήμα 10.9, παρουσιάζονται οι καμπύλες ROC ενός Δένδρου Αποφάσεων και ενός Νευρωνικού Δικτύου, τα οποία προβλέπουν περιπτώσεις, όπου οι εξωτερικοί ελεγκτές εκδίδουν δυσμενή σχόλια. Όπως φαίνεται στο σχήμα, η τιμή AUC του Νευρωνικού Δικτύου είναι μεγαλύτερη από την αντίστοιχη του Δένδρου Αποφάσεων, γεγονός που σημαίνει ότι το μοντέλο του Νευρωνικού Δικτύου προβλέπει πιο αποτελεσματικά τις περιπτώσεις έκδοσης δυσμενών σχολίων.

10.9 Μελέτη Περίπτωσης – Πρόβλεψη τύπου εξωτερικού ελεγκτή με χρήση μεθόδων κατηγοριοποίησης

Αναφερθήκαμε και στο προηγούμενο κεφάλαιο στη σημασία του εξωτερικού ελέγχου, ιδιαίτερα στη σημερινή εποχή. Ο πρωτεύων σκοπός του εξωτερικού ελέγχου είναι να διασφαλίσει την αντικειμενική παρουσίαση της οικονομικής κατάστασης της επιχείρησης, και έτσι να μειώσει την ασυμμετρία στη ροή πληροφορίας ανάμεσα στα διοικητικά στελέχη, τους μετόχους και τους πιστωτές. Ο έλεγχος μπορεί να έχει μια σειρά από ευεργετικά αποτελέσματα. Ο διαχωρισμός της ιδιοκτησίας από τη διοίκηση στις μοντέρνες επιχειρήσεις δημιουργεί κίνητρα στους μάνατζερς να λειτουργήσουν προς ίδιον όφελος και σε βάρος των μετόχων και πιστωτών (Kane & Velury, 2004). Ο έλεγχος αυξάνει την αξιοπιστία των χρηματοοικονομικών καταστάσεων και μειώνει το ρίσκο σφαλερής πληροφόρησης. Συνακόλουθα μειώνει τον επενδυτικό κίνδυνο. Με αυτόν τον τρόπο ο αξιόπιστος έλεγχος μπορεί επίσης να αυξήσει τις τιμές των μετοχών και να μειώσει το κόστος του χρήματος. Ο έλεγχος μπορεί να βελτιώσει την αποδοτικότητα των επιχειρηματικών διαδικασιών και να βοηθήσει στη συμμόρφωση με τις κανονιστικές διατάξεις (Knechel, Niemi & Sundgren, 2008). Αυτά τα ευεργετικά αποτελέσματα αυξάνονται με τη διεξαγωγή ελέγχου υψηλής ποιότητας (Broye & Weill, 2008).



Σχήμα 10.9 *Σύγκριση κατηγοριοποιητών με καμπύλες ROC*

Παρά τα σημαντικά του αποτελέσματα, ο έλεγχος πάσχει από μια εσωτερική αντίθεση. Η αντίθεση προέρχεται από το γεγονός ότι ο ελεγκτής πρέπει να παραμείνει ανεξάρτητος και να προστατεύσει τα συμφέροντα των επενδυτών και των πιστωτών, όμως η πρόσληψη του και η αμοιβή του αποφασίζεται από τη διοίκηση της ελεγχόμενης επιχείρησης. Αυτή η αντίφαση μπορεί να θέσει σε κίνδυνο την αντικειμενικότητα του ελεγκτή. Το ερώτημα της ποιότητας του ελέγχου παραμένει ανοικτό. Ωστόσο, είναι γεγονός ότι όταν επιλέγεται έλεγχος υψηλής ποιότητας το φαινόμενο της ασυμμετρίας μειώνεται (Broye & Weill, 2008).

Είναι γενικά αποδεκτό ότι οι ελεγκτικές εταιρείες χωρίζονται σε δύο κατηγορίες Τη μια κατηγορία απαρτίζουν οι τέσσερεις μεγάλες ελεγκτικές εταιρείες (4 Μεγάλοι Ελεγκτές - 4ΜΕ). Αυτές οι εταιρείες είναι η KPMG, η PriceWaterhouseCoopers, η Ernst & Young και η Deloitte & Touche. Όλοι οι υπόλοιποι ελεγκτές εντάσσονται στη δεύτερη κατηγορία (Όχι 4 Μεγάλοι Ελεγκτές – Ο4ΜΕ). Οι 4ΜΕ θεωρούνται ότι διεξάγουν πιο ποιοτικό έλεγχο (DeAngelo, 1981; Palmrose, 1988). Χάρη στο μέγεθος τους, οι 4ΜΕ είναι σε θέση να αντιστέκονται περισσότερο στις πιέσεις των πελατών τους. Επίσης, επενδύουν περισσότερο σε τεχνολογία, εκπαίδευση και υποδομές, και έχουν περισσότερα κίνητρα να διατηρήσουν την επαγγελματική φήμη τους.

Μελέτες έχουν δείξει ότι ψευδής δήλωση αυξημένων εσόδων είναι σπανιότερη σε εταιρείες που ελέγχονται από τους 4ΜΕ (Francis, Maydew & Sparks, 1999). Σε προηγούμενες μελέτες το μέγεθος του ελεγκτή έχει χρησιμοποιηθεί ως μέτρο της ποιότητας του ελέγχου. (Teoh & Wong, 1993).

Η πρόσληψη του εξωτερικού ελεγκτή είναι μια σύνθετη διαδικασία. Οι μέτοχοι επιδιώκουν να προσλάβουν έναν ελεγκτή υψηλής ποιότητας για να περιορίσουν ενδεχόμενο κίνδυνο χειραγώγησης των οικονομικών στοιχείων, και να επιβεβαιώσουν την αξιοπιστία των οικονομικών καταστάσεων. Επίσης, τα διοικητικά στελέχη, τα οποία θέλουν να σηματοδοτήσουν την αξιοπιστία τους και την ευθυγράμμιση τους με τα συμφέροντα των μετόχων, επιθυμούν επίσης να προσλάβουν ελεγκτές υψηλής ποιότητας. Ωστόσο, οι ελεγκτές υψηλής ποιότητας επενδύουν περισσότερο σε τεχνολογία και εκπαίδευση και επομένως έχουν υψηλότερη αμοιβή. Ένα άλλο ζήτημα είναι ότι, σε περίπτωση αποτυχίας της επιχείρησης και συνακόλουθα αποτυχίας του ελέγχου, η πρόσληψη του ελεγκτή πιθανόν να πρέπει να αιτιολογηθεί. Η υιοθέτηση μιας αποτελεσματικής διαδικασίας πρόσληψης του εξωτερικού ελεγκτή αυξάνει την πιθανότητα να προσλάβει η επιχείρηση τον κατάλληλο ελεγκτή στην κατάλληλη τιμή.

Η ερευνητική βιβλιογραφία αποκαλύπτει ότι οι ερευνητές απορρίπτουν τη μηδενική υπόθεση, ότι οι επιχειρήσεις είναι κατανεμημένες τυχαία μεταξύ των 4ΜΕ και των Ο4ΜΕ. Σε ερευνητικές εργασίες έχει μελετηθεί το θέμα της πρόσληψης εξωτερικού ελεγκτή. Ωστόσο, οι μέθοδοι που χρησιμοποιήθηκαν ήταν κλασσικές στατιστικές, όπως η Λογιστική Παλινδρόμηση. Οι Kirkos, Spathis and Manolopoulos (2010) εφαρμόζουν μεθόδους Εξόρυξης Δεδομένων για την ανάπτυξη μοντέλων, τα οποία προβλέπουν την κατηγορία του εξωτερικού ελεγκτή. Η μελέτη αυτή αυξάνει την κατανόηση σχετικά με την επιλογή κατηγορίας εξωτερικών ελεγκτών. Οι ελεγκτικές εταιρείες μπορούν να χρησιμοποιήσουν τα αποτελέσματα και να ανακαλύπτουν τα χαρακτηριστικά των εταιρειών στις οποίες μπορούν να στοχεύσουν.

Τα δεδομένα της έρευνας προέρχονται από τη βάση οικονομικών δεδομένων FAME (Financial Analysis Made Easy), η οποία περιλαμβάνει στοιχεία για Βρετανικές και Ιρλανδικές επιχειρήσεις. Επιλέχθηκαν οι επιχειρήσεις, οι οποίες ήταν εισηγμένες στο χρηματιστήριο και δραστηριοποιούνταν στους τομείς της βιομηχανίας, των κατασκευών, της πληροφορικής και της εξόρυξης μεταλλευμάτων, και οι οποίες άλλαξαν εξωτερικό ελεγκτή τα έτη 2003-2005. Οι επιλεγμένες επιχειρήσεις συνταιριάστηκαν με ίσο αριθμό επιχειρήσεων, οι οποίες δεν άλλαξαν τον εξωτερικό τους ελεγκτή.

Η αρχική επιλογή ανεξάρτητων μεταβλητών στηρίχθηκε στην προηγούμενη έρευνα. Περιλήφθηκαν μεταβλητές που αφορούσαν το μέγεθος της ελεγχόμενης εταιρείας (Krishnan, Krishnan & Stephens, 1996), το πλήθος των θυγατρικών εταιρειών, το ύψος του δανεισμού (Knechel et al., 2008; Broye & Weil, 2008), την αποθήκη και τους εισπρακτέους λογαριασμούς (Icerman & Hillison, 1991), την κερδοφορία (Citron & Manalis, 2001), την έκδοση δυσμενών σχολίων (Chow & Rice, 1982; Citron & Taffler, 1992; Krishnan et al., 1996), τις τάσεις αύξησης του μεγέθους της εταιρείας (Velury, Reish & O'Reilly, 2003), τις αμοιβές των εξωτερικών ελεγκτών, τη χρηματιστηριακή αξία της επιχείρησης (Kane & Velury, 2004), καθώς και μερικοί ακόμα γνωστοί αριθμοδείκτες, όπως το Quick Ratio και το Z-Score του Altman.

Συνολικά επιλέχθηκαν τριάντα πέντε μεταβλητές. Οι μεταβλητές αυτές υποβλήθηκαν σε έλεγχο σημαντικότητας με εφαρμογή της μεθόδου ANOVA. Δεκαοκτώ μεταβλητές παρουσίαζαν μικρή τιμή p και επιλέχθηκαν να συμμετέχουν στο τελικό άνυσμα εισόδου. Η στατιστική ανάλυση των μεταβλητών αποκάλυψε και ορισμένες πρώτες ενδείξεις συσχέτισης τιμών των μεταβλητών με τον τύπο του εξωτερικού ελεγκτή. Το μέγεθος της ελεγχόμενης επιχείρησης είναι σημαντικό, και οι μεγάλες επιχειρήσεις τείνουν τα προσλάβουν μεγάλους ελεγκτές. Σημαντικό είναι επίσης το ύψος του χρέους. Επιχειρήσεις με μεγαλύτερο ποσοστό χρέους τείνουν να προσλάβουν μεγάλους ελεγκτές. Αξιόλογη διαφοροποίηση παρουσιάζεται και σε μεταβλητές που αναφέρονται στη ρευστότητα. Είναι αξιοσημείωτο ότι ο αριθμοδείκτης Z-Score παρουσίασε υψηλή τιμή *p*, παρέχοντας ισχυρές ενδείξεις ότι η οικονομική ευρωστία ή δυσπραγία δεν επηρεάζει την επιλογή τύπου εξωτερικού ελεγκτή. Επίσης, όλες οι μεταβλητές, οι οποίες αναφέρονταν σε τάσεις οικονομικών μεγεθών, απορρίφθηκαν ως μη σημαντικές.

Τρεις μέθοδοι εξόρυξης δεδομένων, τα Δένδρα Αποφάσεων τύπου C4.5, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron και οι k-Πλησιέστεροι Γείτονες εφαρμόστηκαν για την πρόβλεψη του τύπου του εξωτερικού ελεγκτή. Οι τρεις αυτές μέθοδοι συγκρίθηκαν με τη Λογιστική Παλινδρόμηση, η οποία ήταν και η μοναδική μέθοδος που είχε εφαρμοστεί σε προηγούμενες εργασίες. Το Δένδρο Απόφασης εκπαιδεύτηκε με επίπεδο εμπιστοσύνης 0,25 και περιείχε εικοσιπέντε κόμβους και δεκατρία φύλλα. Ως μεταβλητή διαχωρισμού πρώτου επιπέδου επιλέχθηκε το Σύνολο Χρέους (Total Debt). Σύμφωνα με το κριτήριο του Λόγου Κέρδους (Gain Ratio), η μεταβλητή αυτή διαχωρίζει βέλτιστα τις δυο κατηγορίες. Η μεγάλη πλειοψηφία των επιχειρήσεων με υψηλό επίπεδο χρέους (95 από 98 παρατηρήσεις) επιλέγει ελεγκτή 4ΜΕ. Συμπεραίνουμε ότι οι επιχειρήσεις με υψηλό επίπεδο χρέους επιδιώκουν ποιοτικότερο έλεγχο. Οι αμοιβή των ελεγκτών και οι εισπρακτέοι λογα-

ριασμοί επιλέχθηκαν επίσης ως μεταβλητές διαχωρισμού υψηλού επιπέδου.

Διάφορες εναλλακτικές αρχιτεκτονικές δοκιμάστηκαν για το Νευρωνικό Δίκτυο, και τελικά επιλέχθηκε μια αρχιτεκτονική με ένα κρυφό στρώμα, το οποίο περιείχε ένδεκα κρυφούς νευρώνες. Επειδή το Νευρωνικό Δίκτυο δεν παρέχει κάποια κατανοητή ερμηνεία σχετικά με τη σημαντικότητα των μεταβλητών εισόδου, εφαρμόστηκε ένας έμμεσος έλεγχος, ο οποίος αποτελούνταν από μια επαναληπτική διαδικασία. Σε κάθε στάδιο της επανάληψης αφαιρούνταν μια από τις μεταβλητές εισόδου, και ελέγχονταν η ακρίβεια του μοντέλου. Αν η αφαίρεση μιας μεταβλητής προκαλούσε σημαντική πτώση της ακρίβειας, τότε η μεταβλητή θεωρούνταν σημαντική. Σύμφωνα με τα αποτελέσματα, η σημαντικότερη μεταβλητή ήταν το Σύνολο Χρέους. Επισημαίνεται ότι η ίδια μεταβλητή είχε επιλεγεί ως διαχωριστής πρώτου επιπέδου από το Δένδρο Απόφασης.

Για τη μέθοδο των k-Πλησιέστερων Γειτόνων, το πλήθος των γειτονικών σημείων k ορίστηκε να είναι ίσο με πέντε. Δυστυχώς, για τη συγκεκριμένη μέθοδο δεν υπήρχε η δυνατότητα ελέγχου της σημαντικότητας των μεταβλητών εισόδου. Η τελευταία μέθοδος που εφαρμόστηκε ήταν η Λογιστική Παλινδρόμηση. Σύμφωνα με το κριτήριο Wald, η πιο σημαντική μεταβλητή εισόδου ήταν το Σύνολο Χρέους. Είναι αξιοσημείωτο ότι και οι τρεις μέθοδοι, οι οποίες παρείχαν κριτήρια εκτίμησης της σημαντικότητας των ανεξάρτητων μεταβλητών, συμφωνούν ότι η μεταβλητή, η οποία επηρεάζει σε μεγαλύτερο βαθμό το αποτέλεσμα της κατηγοριοποίησης, είναι το σύνολο χρέους.

Και τα τέσσερα μοντέλα επέτυχαν υψηλούς ρυθμούς ακρίβειας έναντι του συνόλου εκπαίδευσης, οι οποίοι κυμαίνονταν από 79% έως 93%. Ωστόσο, αυτές οι επιδόσεις δεν μπορούν να θεωρηθούν ενδεικτικές των πραγματικών δυνατοτήτων των μοντέλων. Ένας υπαρκτός κίνδυνος είναι η υπερπροσαρμογή των μοντέλων, η απομνημόνευση δηλαδή των παρατηρήσεων του συνόλου εκπαίδευσης. Η πραγματική αξία όμως των μοντέλων βρίσκεται στην εφαρμογή τους στην καθημερινή πράξη, όπου και θα συναντήσουν άγνωστες παρατηρήσεις, διαφορετικές από αυτές του συνόλου εκπαίδευσης. Για τον λόγο αυτό, τα μοντέλα υποβλήθηκαν σε διαδικασία επικύρωσης, ώστε να εκτιμηθεί η ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων. Εφαρμόστηκαν δύο διαφορετικές τεχνικές επικύρωσης. Η πρώτη τεχνική ήταν η διασταυρούμενη επικύρωση 10 τμημάτων (10 fold cross validation). Η δεύτερη τεχνική ήταν η διάσπαση του συνόλου δεδομένων σε δύο υποσύνολα, όπου το ένα χρησιμοποιήθηκε για εκπαίδευση και το δεύτερο για επικύρωση. Ειδικότερα, επιλέχθηκαν οι επιχειρήσεις οι οποίες άλλαξαν ελεγκτή τα έτη 2003 και 2004 για την εκπαίδευση των μοντέλων (220 παρατηρήσεις), ενώ οι υπόλοιπες επιχειρήσεις (118 παρατηρήσεις) χρησιμοποιήθηκαν για επικύρωση. Σύμφωνα με τα αποτελέσματα των δύο ελέγχων, και τα τέσσερα μοντέλα αποδείχθηκαν ικανά να κατηγοριοποιούν άγνωστες παρατηρήσεις και επέτυχαν ικανοποιητικούς ρυθμούς ακρίβειας. Και στις δύο περιπτώσεις, το Δένδρο Αποφάσεων επέτυχε τις καλύτερες επιδόσεις, ακολουθούμενο από το Νευρωνικό Δίκτυο. Οι k-Πλησιέστεροι Γείτονες και η Λογιστική Παλινδρόμηση εναλλάσσονται στην τρίτη και τέταρτη θέση ανάλογα με την τεχνική επικύρωσης. Τα αποτελέσματα παρέχουν αποδείξεις ότι οι νέες τεχνικές, οι οποίες προέρχονται από τον χώρο της Μηχανικής Μάθησης, υπερβαίνουν σε επιδόσεις την ευρέως χρησιμοποιούμενη Λογιστική Παλινδρόμηση.

Και οι τέσσερεις τεχνικές επέτυχαν αρκετά υψηλά ποσοστά ακρίβειας. Ωστόσο, η περαιτέρω βελτίωση των επιδόσεων αποτελεί μόνιμη επιδίωξη. Όπως αναφέρθηκε και προηγουμένως, οι σύνθετοι κατηγοριοποιητές μπορούν να υπερβούν τις επιδόσεις των ατομικών τεχνικών. Σε μια απόπειρα αύξησης του ρυθμού ακρίβειας εφαρμόστηκε η τεχνική Bagging, και τα νέα μοντέλα δοκιμάστηκαν με τη μέθοδο της διασταυρούμενης επικύρωσης 10 τμημάτων. Σύμφωνα με τα αποτελέσματα, το Δένδρο Αποφάσεων βελτίωσε τις επιδόσεις του κατά 3,5% περίπου, ενώ μικρότερη βελτίωση υπήρχε στο Νευρωνικό Δίκτυο και τη Λογιστική Παλινδρόμηση. Οι ρυθμοί ακρίβειας των κατηγοριοποιητών παρουσιάζονται αναλυτικά στον Πίνακα 10.1

	C4.5	**Νευρωνικό Δίκτυο**	**k-NN**	**Λογ. Παλινδρόμηση**
10 fold cross validation	82,12	77,27	69,09	76,66
Validation set	82,09	72,88	71,19	63,56
Bagging + 10 fold cross validation	85,45	79.09	69,09	77,88

Πίνακας 10.1 *Ρυθμοί ακρίβειας μοντέλων*

Βιβλιογραφία/Αναφορές

Bergadano, F., Matwin, S., Michalski, R. S., & Zhang, J. (1988). Measuring Quality of Concept Descriptions. In *3rd European Working Session on Learning* (pp. 1-14). Glasgow, SCT: Pittman.

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123–140. doi: 10.1007/bf00058655

Broye, G., & Weill, L. (2008). Does Leverage Influences Auditor Choice? A Cross-country Analysis. A*pplied Financial Economics, 18*(9), 715-731. doi: 10.1080/09603100701222325

Chawla, N. V., Boywer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research, 16*(1), 321-357.

Chen, H. J., Huang, S. Y., & Lin, C. S. (2009). Alternative Diagnosis of Corporate Bankruptcy: A Neuro- Fuzzy Approach. *Expert Systems with Applications, 36*(4), 7710-7720. doi: 10.1016/j.eswa.2008.09.023

Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J., & Neves, C. J. (2011). A Genetic Algorithm-Based Approach to Cost-Sensitive Bankruptcy Prediction. *Expert Systems with Applications*, *38*(10), 12939–12945. doi: 10.1016/j.eswa.2011.04.090

Chow, C., & Rice, S. (1982). Qualified Audit Opinions and Auditor Switching. *The Accounting Review, 57*(2), 326-335.

Citron, D., & Taffler, R. (1992). The Audit Report under Going Concern Uncertainties: An Empirical Analysis. *Accounting and Business Research, 22*(88), 337-345. doi: 10.1080/00014788.1992.9729449

Citron, D., & Manalis, G. (2001). The International Firms as new Entrants to the Statutory Audit Market: An Empirical analysis of auditor selection in Greece, 1993 to 1997. *The European Accounting Review, 10*(3), 439–459. doi: 10.2139/ssrn.233635

DeAngelo, L. (1981). Auditor Size and Auditor Quality. *Journal of Accounting and Economics, 3*(3), 183–199. doi: 10.1016/0165-4101(81)90002-1

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2,* 973-978. San Francisco, CA: Morgan Kaufman.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Francis, J., Maydew, E., & Sparks, H. C. (1999). The Role of Big 6 Auditors in the Credible Reporting of Accruals. *Auditing: A Journal of Practice and Theory, 18*(2), 17-34. doi: 10.2308/aud.1999.18.2.17

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the 13th International Conference on Machine Learning,* 148–156. San Francisco, CA: Morgan Kaufmann.

Icerman, R., & Hillison, W. (1991). Disposition of Auditor-Detected Errors: Some Evidence on Evaluative Materiality. *Auditing: A Journal of Practice and Theory, 10*, 22-34.

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Truck on Inductive Learning,* 111-117. Las Vegas, NV.

Kane, G., & Velury, U. (2004). The Role of Institutional Ownership in the Market for Auditing Services: An empirical Investigation. *Journal of Business Research, 57*(9), 976-983. doi: 10.1016/s0148-2963(02)00499-x

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2010), Audit-Firm Group Appointment: An Artificial Intelligence Approach. *Intelligent Systems in Accounting, Finance and Management, 17*(1), 1-17. doi: 10.1002/isaf.310

Kirkos, E. (2015). Assessing Methodologies for Intelligent Bankruptcy Prediction. *Artificial Intelligence Review, 43*(1), 83-123. doi: 10.1007/s10462-012-9367-6

Knechel, R., Niemi, L., & Sundgren, S. (2008). Determinants of Auditor Choice: Evidence from a Small Client Market. *International Journal of Auditing, 12*(1), 65-88. doi: 1099-1123.2008.00370.x

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2,* 1137-1143. San Francisco, CA: Morgan Kaufmann.

Kreßel, U. (1999). Pairwise Classification and Support Vector Machines. In B. Schoelkopf, C. Burges & A. Smola (Eds.), *Advances in Kernel Methods – Support Vector Learning* (pp. 255-268). Cambridge, MA: MIT Press.

Krishnan, J., Krishnan, J., & Stephens, R. (1996). The Simultaneous Relation between Auditor Switching

and Audit Opinion: An empirical analysis. *Accounting and Business Research, 26*(3), 224–236. doi: 10.1080/00014788.1996.9729513

Kubat, M., & Martin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One Sided Selection. *Proceedings of the 14th International Conference on Machine Learning,* 179-186. Nashville, TN: Morgan Kaufmann.

Lin, W. Y., Hu, Y. H., & Tsai, C. F. (2012). Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems, Man and Cybernetics, 42*(4), 421-436. doi: 10.1109/tsmcc.2011.2170420

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer + Business Media.

Palmrose, Z. (1988). An Analysis of Auditor Litigation and Audit Service Quality. *The Accounting Review, 63*(1), 55-73.

Simonoff, J. (2003). *Analyzing Categorical Data.* New York, NY: Springer-Verlag.

Smola, A. J., & Schoelkopf, B. (2004). A Tutorial on Support Vector Regression. *Statistics and Computing, 14*(3), 199-222. doi: 10.1023/B:STCO.0000035301.49549.88

Steinwart, I. (2003). On the Optimal Parameter Choice for NU-Support Vector Machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(10), 1274-1284. doi: 10.1109/tpami.2003.1233901

Teoh, S. H., & Wong, T. J. (1993). Perceived Auditor Quality and the Earnings Response Coefficient. *The Accounting Review, 68*(2), 346-366.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer Verlag.

Velury, U., Reish, J., & O'Reilly, D. (2003). Institutional Ownership and the Selection of Industry Specialist Auditors. *Review of Quantitative Finance and Accounting, 21*(1), 35–48. doi: 10.1023/A:1024855605207

Weiss, G. M., & Provost, F. (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research, 19*(1), 315-354.

Wilson, R. D., & Martinez, T. R. (1997). Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research, 6*(1), 1-34.

Witten, I. H., & Frank, E. (2000). *Data Mining Practical Machine Learning Tools and Techniques with JAVA Implementations*. San Francisco, CA: Morgan Kaufman.

Wolpert, D. (1992). Stacked Generalization. *Neural Networks*, 5(2), 241-259. doi: 10.1016/S0893-6080(05)80023-1

Κριτήρια Αξιολόγησης

Άσκηση Υπολογισμών 10.1

Στον Πίνακα 10.2 δίνονται οι τιμές των μεταβλητών X και Y. Υπολογίστε τις τιμές των a και b, ώστε να εκφραστεί η μεταβλητή Y ως γραμμική παλινδρόμηση του X (Y=a+b*X). Υπολογίστε την τιμή του Y, εάν το X πάρει την τιμή 10.

X	Y
5	11
4	6
7	16
12	20
3	6
6	11
8	18
8	15
11	21
9	20

Πίνακας 10.2 *Δεδομένα Άσκησης 1*

Λύση

Τα *b* και *a* θα υπολογιστούν σύμφωνα με τις Εξισώσεις 10.18 και 10.19 αντίστοιχα. Αρχικά υπολογίζονται οι μέσες τιμές των *X* και *Y* (X_m=*7,3* και Y_m=*14,4*). Ακολούθως, υπολογίζονται οι τιμές των *a* και *b* (*a=1,085* και *b=1,824*). Το *Y* εκφράζεται ως γραμμική συνάρτηση του *X* σύμφωνα με τη σχέση *Y=1,085+1,824X*.
Εάν το *X* πάρει την τιμή *0*, το *Y* υπολογίζεται ως *Y=1,085+1,824*10=19,325*.

Άσκηση Υπολογισμών 10.2

Ένα μοντέλο κατηγοριοποιητή προβλέπει την πιστοληπτική ικανότητα. Υπάρχουν τρεις δυνατές τιμές κλάσης, η «Υψηλή», η «Μεσαία» και η «Χαμηλή». Τα αποτελέσματα του μοντέλου παρουσιάζονται στον ακόλουθο πίνακα σύγχυσης.

	Υψηλή	Μεσαία	Χαμηλή
Υψηλή	**90**	**6**	**4**
Μεσαία	**7**	**85**	**8**
Χαμηλή	**3**	**5**	**92**

Απαντήστε στις παρακάτω περιπτώσεις:
Πόσες είναι οι παρατηρήσεις συνολικά;
Πόσες περιπτώσεις μεσαίας πιστοληπτικής ικανότητας κατηγοριοποιήθηκαν σωστά;
Ποιο ποσοστό παρατηρήσεων κατηγοριοποιήθηκαν σωστά;
Πόσες περιπτώσεις μεσαίας πιστοληπτικής ικανότητας κατηγοριοποιήθηκαν ως «Υψηλή»;
Πόσες περιπτώσεις χαμηλής πιστοληπτικής ικανότητας κατηγοριοποιήθηκαν λάθος;

Λύση

Οι παρατηρήσεις συνολικά είναι (90+6+4+7+85+8+3+5+92)=300.

Οι περιπτώσεις μεσαίας πιστοληπτικής ικανότητας που κατηγοριοποιήθηκαν σωστά είναι 85.

Το ποσοστό των παρατηρήσεων που κατηγοριοποιήθηκαν σωστά είναι (90+85+92)*100/300 = 89%
Οι περιπτώσεις μεσαίας πιστοληπτικής ικανότητας που κατηγοριοποιήθηκαν ως «Υψηλή» είναι 7.
Οι περιπτώσεις χαμηλής πιστοληπτικής ικανότητας που κατηγοριοποιήθηκαν λάθος είναι (3+5)=8.

Άσκηση Εφαρμογής 10.3

Χρησιμοποιήστε το αρχείο «analcatdata_japansolvent.arff» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή StatLib). Το σύνολο δεδομένων προέρχεται από το βιβλίο του Simonoff (2003), και σχετίζεται με την κατηγοριοποίηση ιαπωνικών επιχειρήσεων σε φερέγγυες (solvent) και αφερέγγυες (insolvent). Υπάρχουν 52 γραμμές, κάθε μια από τις οποίες αναφέρεται σε μια επιχείρηση. Οι 25 επιχειρήσεις χαρακτηρίζονται αφερέγγυες και οι 27 φερέγγυες. Στο σύνολο δεδομένων υπάρχουν 10 πεδία (στήλες). Το πρώτο πεδίο περιέχει τα ονόματα των επιχειρήσεων, το δεύτερο είναι το πεδίο κλάσης, και ακολουθούν οκτώ αριθμοδείκτες. Οι κλάσεις κωδικοποιούνται με τις τιμές «0» για τις αφερέγγυες επιχειρήσεις και «1» για τις φερέγγυες.
Αναπτύξτε μοντέλα πρόβλεψης φερεγγυότητας επιχειρήσεων με χρήση των μεθόδων α)Λογιστικής Παλινδρόμησης, β)Μηχανών Διανυσμάτων Υποστήριξης, γ)k-Πλησιέστερων Γειτόνων. Πειραματιστείτε με τις τιμές των παραμέτρων για τις Μηχανές Διανυσμάτων Υποστήριξης και για τη μέθοδο των k-Πλησιέστερων Γειτόνων, προσπαθώντας να αυξήσετε τις επιδόσεις.

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «analcatdata_japansolvent.arff» πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» εμφανίζονται διάφορες πληροφορίες για τα δεδομένα. Παρατηρήστε τα πεδία (Attributes). Ως πεδίο κλάσης ορίστε το πεδίο «Solvent». Η κατανομή των παρατηρήσεων χρωματίζεται ανάλογα με την τιμή της κλάσης.

Το πεδίο με τα ονόματα των επιχειρήσεων δεν προσφέρει κάτι χρήσιμο στην ανάλυση μας. Το επιλέγετε και το απομακρύνετε πιέζοντας το κουμπί «Remove».

Μελετήστε την κατανομή τιμών στα διάφορα γνωρίσματα, κάνοντας κλικ στο όνομα του γνωρίσματος. Παρατηρήστε ότι για τους περισσότερους αριθμοδείκτες οι αφερέγγυες επιχειρήσεις (μπλε) τείνουν προς το αριστερό άκρο της κατανομής, ενώ οι φερέγγυες (κόκκινες) τείνουν προς το δεξιό άκρο.

Βήμα 2. Μεταβείτε στο tab «Classify».

Βεβαιωθείτε ότι ως πεδίο κλάσης έχει οριστεί το πεδίο «Solvent» και ότι είναι επιλεγμένη η μέθοδος ελέγχου «Cross-validation».

Επιλέξτε μέθοδο κατηγοριοποίησης πιέζοντας το κουμπί «Choose» στο πεδίο «Classifier». Επιλέξτε πρώτα τη μέθοδο weka/classifiers/functions/SimpleLogistic για τη Λογιστική Παλινδρόμηση και πατήστε το κουμπί «Start». Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Το μοντέλο κατηγοριοποιεί σωστά 78.8462% του συνόλου των παρατηρήσεων, 68% της κλάσης «0» και 88.9% της κλάσης «1».

Βήμα 3. Επιλέξτε τη μέθοδο weka/classifiers/functions/SMO για τις Μηχανές Διανυσμάτων Υποστήριξης και πατήστε το κουμπί «Start». Το μοντέλο κατηγοριοποιεί σωστά 76.9231% των συνολικών περιπτώσεων, 60% της κλάσης «0» και 92.6% της κλάσης «1».

Βήμα 4. Επιλέξτε τη μέθοδο weka/classifiers/lazy/IBk για τη μέθοδο των k-Πλησιέστερων Γειτόνων και πατήστε το κουμπί «Start». Το μοντέλο κατηγοριοποιεί σωστά 82.6923% των συνολικών περιπτώσεων, 76% της κλάσης «0» και 88.9% της κλάσης «1».

Βήμα 5. Επιλέξτε ξανά τη μέθοδο SMO και πειραματιστείτε με τις τιμές της παραμέτρου «C». Επίσης, στο πεδίο «kernel», πειραματιστείτε με τον εκθέτη της συνάρτησης πυρήνα. Επιλέξτε ξανά τη μέθοδο IBk και πειραματιστείτε με το πλήθος των γειτόνων (πεδίο «KNN») και με τον τρόπο υπολογισμού της απόστασης (πεδίο «nearestNeighborSearchAlgorithm»). Προσπαθήστε με τη ρύθμιση των παραμέτρων να αυξήσετε τις επιδόσεις. Για παράδειγμα, στη μέθοδο SMO, θέτοντας την τιμή 1,5 στην παράμετρο «C», και ορίζοντας τιμή εκθέτη ίση με τρία για τη συνάρτηση πυρήνα, η ακρίβεια του μοντέλου Μηχανών Διανυσμάτων Υποστήριξης αυξάνεται στο 80.7692%.

Άσκηση Εφαρμογής 10.4

Χρησιμοποιήστε το αρχείο «labor.arff» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://

www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή UCI repository). Τα δεδομένα αποτελούν αποτέλεσμα συλλογικών διαπραγματεύσεων για ζητήματα εργασίας στον Καναδά, προσφέρθηκαν από τον καθηγητή Stan Matwin και χρησιμοποιήθηκαν στην εργασία των Bergadano et al. (1988). Περιλαμβάνονται συνολικά 17 γνωρίσματα, τα οποία αναφέρονται σε αυξήσεις μισθού τον πρώτο, δεύτερο και τρίτο χρόνο, στις ώρες εργασίας, σε συνταξιοδοτικά και ασφαλιστικά πλάνα, σε ημέρες άδειας κλπ. Το τελευταίο γνώρισμα είναι το γνώρισμα κλάσης, και οι δυνατές τιμές κλάσεις είναι «good» και «bad». Τα δεδομένα περιέχουν 57 παρατηρήσεις, εκ των οποίων οι 37 ανήκουν στην κλάση «good» και οι 20 στην κλάση «bad».

Εκτελέστε επιλογή χαρακτηριστικών εφαρμόζοντας τη μέθοδο CFS Subset Evaluator. Αναπτύξτε μοντέλο ικανό να προβλέπει την κλάση των παρατηρήσεων, εφαρμόζοντας τη μέθοδο Δένδρων Αποφάσεων C4.5. Αυξήστε τις επιδόσεις της μεθόδου, αναπτύσσοντας συνδυασμό κατηγοριοποιητών με χρήση της μεθόδου Bagging.

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «labor.arff» πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» εκτελέστε την επιλογή χαρακτηριστικών. Στο πεδίο "Filter" πιέστε το κουμπί "Choose" και επιλέξτε weka/filters/supervised/attribute/AttributeSelection. Αυτομάτως επιλέγεται η μέθοδος CfsSubsetEval. Κάνετε κλικ στο κουμπί «Apply». Μετά την εκτέλεση του αλγορίθμου θα διαπιστώσετε ότι στο πεδίο «Attributes» μειώθηκε το πλήθος των στηλών. Ειδικότερα, παραμένουν επτά στήλες και επιπλέον η στήλη της κλάσης, ενώ οι υπόλοιπες στήλες απομακρύνονται.

Βήμα 2. Μεταβείτε στο tab «Classify» και επιλέξτε κατηγοριοποιητή, κάνοντας κλικ στο κουμπί «Choose» του πεδίου «Classifier». Από τις διαθέσιμες μεθόδους επιλέξτε τη μέθοδο weka/classifiers/trees/J48. Η μέθοδος αυτή δημιουργεί Δένδρα Αποφάσεων C4.5.

Εκπαιδεύστε το μοντέλο και επικυρώστε το με τη μέθοδο «CrossValidation». Για να εκτελέσετε αυτήν την εργασία, βεβαιωθείτε ότι είναι επιλεγμένη η μέθοδος «Cross-validation» στο πεδίο «Test-options» και στη συνέχεια κάντε κλικ στο κουμπί «Start».

Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Μπορείτε να δείτε το Δένδρο Αποφάσεων. Το μοντέλο κατηγοριοποιεί σωστά 77.193% των συνολικών περιπτώσεων, 65% των παρατηρήσεων με κλάση «bad» και 83.8% των παρατηρήσεων με κλάση «good».

Βήμα 3. Μπορείτε να αυξήσετε τις επιδόσεις εφαρμόζοντας τη μέθοδο Bagging. Στο πεδίο «Classifier» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka\classifiers\meta\bagging. Το Bagging είναι μια γενική τεχνική και μπορεί να συνδυαστεί με οποιαδήποτε μέθοδο κατηγοριοποίησης. Για τον λόγο αυτό, πρέπει να ορίσετε τη μέθοδο κατηγοριοποίησης για την οποία θα κατασκευαστούν πολλαπλά μοντέλα. Κάνετε κλικ στα περιεχόμενα του πεδίου «Classifier». Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Στο πεδίο «Classifier» κάντε κλικ στο κουμπί «Choose» και επιλέξτε τη μέθοδο weka/classifiers/trees/J48. Κάντε κλικ στο κουμπί «OK».

Στο σημείο αυτό έχετε ορίσει ότι θέλετε να εφαρμόσετε bagging στη μέθοδο κατηγοριοποίησης C4.5. Κάντε κλικ στο κουμπί «Start» για να εκτελέσετε τον αλγόριθμο.

Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Με τη χρήση της μεθόδου Bagging κατηγοριοποιούνται σωστά 84.2105% των συνολικών περιπτώσεων, 75% των παρατηρήσεων με κλάση «bad» και 89.2% των παρατηρήσεων με κλάση «good». Παρατηρούμε ότι επιτεύχθηκε σημαντική αύξηση των επιδόσεων.

Άσκηση Εφαρμογής 10.5

Χρησιμοποιήστε το αρχείο «analcatdata_bankruptcy.arff» (θα το βρείτε στην ιστοσελίδα δεδομένων του WEKA (http://www.cs.waikato.ac.nz/ml/weka/datasets.html), στη συλλογή StatLib). Το σύνολο δεδομένων προέρχεται από το βιβλίο του Simonoff (2003) και σχετίζεται με τη χρεοκοπία επιχειρήσεων. Υπάρχουν 50 γραμμές, κάθε μια από τις οποίες αναφέρεται σε μια επιχείρηση. Οι μισές επιχειρήσεις έχουν χρεοκοπήσει. Στο σύνολο δεδομένων υπάρχουν 7 πεδία (στήλες). Το πρώτο πεδίο περιέχει τα ονόματα των επιχειρήσεων, και ακολουθούν 5 πεδία με αριθμοδείκτες. Το τελευταίο πεδίο είναι το πεδίο της κλάσης, και περιέχει μια ένδειξη («1» ή «0») για το εάν η επιχείρηση χρεοκόπησε ή εξακολουθεί τη λειτουργία της αντίστοιχα.

Αναπτύξτε μοντέλο πρόβλεψης χρεοκοπίας με χρήση της μεθόδου Νευρωνικό Δίκτυο τύπου Multilayer

Perceptron. Ακολούθως εκτελέστε ευαίσθητη στο κόστος κατηγοριοποίηση, εφαρμόζοντας τη μέθοδο MetaCost, σε συνδυασμό με κατηγοριοποιητή Multilayer Perceptron. Ελέγξτε εάν βελτιώθηκε το ποσοστό ορθών προβλέψεων της ακριβής κλάσης.

Λύση

Βήμα 1. Εκκινήστε το WEKA και ανοίξτε το αρχείο «analcatdata_bankruptcy.arff» πιέζοντας το κουμπί «Open file».

Το πεδίο με τα ονόματα των επιχειρήσεων δεν προσφέρει κάτι χρήσιμο στην ανάλυση μας. Το επιλέγετε και το απομακρύνετε πιέζοντας το κουμπί «Remove».

Μεταβείτε στο tab «Classify».

Επιλέξτε τη μέθοδο κατηγοριοποίησης weka/classifiers/functions/MultilayerPerceptron και πατήστε το κουμπί «Start». Το μοντέλο κατηγοριοποιεί σωστά 90% των συνολικών περιπτώσεων, 88% της κλάσης «0» (μη χρεοκοπία) και 92% της κλάσης «1» (χρεοκοπία).

Βήμα 2. Στο πεδίο «Classifier» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka\classifiers\meta\ MetaCost. Η MetaCost είναι μια γενική μέθοδος, η οποία συνδυάζεται με οποιαδήποτε μέθοδο κατηγοριοποίησης, ώστε να επιτευχθεί ευαίσθητη στο κόστος κατηγοριοποίηση.

Κάντε κλικ στα περιεχόμενα του πεδίου «Classifier» και στο όνομα «MetaCost», ώστε να ανοίξει το παράθυρο ρύθμισης των παραμέτρων.

Σε αυτό το παράθυρο κάντε κλικ στο κουμπί «Choose» του πεδίου «classifier» και επιλέξτε weka/classifiers/functions/MultilayerPerceptron. Με τον τρόπο αυτό ορίζετε ότι θα εφαρμόσετε τη μέθοδο MetaCost σε συνδυασμό με νευρωνικό δίκτυο Multilayer Perceptron.

Στο ίδιο παράθυρο κάντε κλικ στα περιεχόμενα του πεδίου «costMatrix». Ανοίγει ένα νέο παράθυρο όπου καθορίζονται οι τιμές του πίνακα κόστους. Στο πεδίο «Classes» γράψτε την τιμή «2» και κάντε κλικ στο κουμπί «Resize». Στο άνω και αριστερά μέρος του παραθύρου εμφανίζεται ένας πίνακας 2X2.

Στον πίνακα 2X2 καταχωρίστε τις τιμές κόστους όπως φαίνεται κατωτέρω.

0.0	1.0
10.0	0.0

Με τον τρόπο αυτό, ορίζετε ότι το κόστος εσφαλμένης κατηγοριοποίησης χρεοκοπημένων επιχειρήσεων είναι δεκαπλάσιο από το κόστος εσφαλμένης κατηγοριοποίησης μη χρεοκοπημένων επιχειρήσεων.

Κλείστε το παράθυρο με τον πίνακα κόστους και κάντε κλικ στο κουμπί «OK» του παραθύρου ρύθμισης παραμέτρων. Κάντε κλικ στο κουμπί «Start».

Στο πεδίο «Classifier output» εμφανίζονται τα αποτελέσματα. Παρατηρούμε ότι το ποσοστό ορθών προβλέψεων έπεσε στο 82%. Επίσης το ποσοστό των ορθών προβλέψεων των παρατηρήσεων κλάσης «0» (μη χρεοκοπία) έπεσε στο 64%. Όμως το ποσοστό των ορθών προβλέψεων των παρατηρήσεων κλάσης «1» (χρεοκοπία) αυξήθηκε στο 100%. Με τη χρήση της μεθόδου MetaCost επιτεύχθηκε αύξηση των ορθών προβλέψεων της ακριβής κλάσης, με αντίτιμο την πτώση των ορθών προβλέψεων των μη χρεοκοπημένων επιχειρήσεων. Με δεδομένο ότι σε πραγματικές συνθήκες το κόστος εσφαλμένης πρόβλεψης μιας χρεοκοπημένης επιχείρησης είναι πολύ μεγαλύτερο, μπορούμε να ισχυριστούμε ότι πρακτικά το μοντέλο βελτιώθηκε.

11 Ανάλυση Συστάδων

Σύνοψη

Η Ανάλυση Συστάδων (ΑΣ) (Clustering) είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Στόχος της ΑΣ είναι ο επιμερισμός ενός συνόλου παραδειγμάτων σε συστάδες. Οι συστάδες συγκροτούνται στη βάση της ομοιότητας των μελών τους. Το γεγονός ότι δεν υπάρχει εκ των προτέρων γνώση σχετικά με την ύπαρξη ομάδων χαρακτηρίζει την ΑΣ ως μη επιβλεπόμενη μάθηση. Ένας τρόπος καθορισμού του βαθμού ομοιότητας δύο παρατηρήσεων είναι με τη χρήση της απόστασης τους. Οι παρατηρήσεις θεωρούνται ως σημεία σε έναν πολυδιάστατο χώρο. Η απόσταση τους σε αυτόν τον χώρο αποτελεί το μέτρο της ομοιότητας τους. Εάν όλα τα γνωρίσματα είναι αριθμητικά, τότε για τον υπολογισμό της ανομοιότητας χρησιμοποιείται η Ευκλείδεια απόσταση ή κάποια παραλλαγή της, όπως η απόσταση Manhattan ή η απόσταση Minkowski. Συνήθως, σε μια βάση δεδομένων, εκτός από τα αριθμητικά πεδία, υπάρχουν και δυαδικά, ονομαστικά και διατακτικά πεδία. Για κάθε έναν από αυτούς τους τύπους των γνωρισμάτων έχουν καθοριστεί τρόποι υπολογισμού της απόστασης. Οι τρόποι αυτοί παρουσιάζονται αναλυτικά. Επίσης, ορίζεται ο υπολογισμός της απόστασης για περιπτώσεις, όπου οι παρατηρήσεις περιλαμβάνουν γνωρίσματα διαφορετικών τύπων, δηλαδή αριθμητικά, δυαδικά, ονομαστικά κλπ.
Υπάρχει μια μεγάλη ποικιλία μεθόδων ΑΣ. Οι μέθοδοι αυτές χωρίζονται σε Ιεραρχικές, Διαχωριστικές, μεθόδους βασισμένες στην πυκνότητα, μεθόδους πλέγματος και μεθόδους βασισμένες σε μοντέλα. Οι Ιεραρχικές μέθοδοι δημιουργούν μια ιεραρχία επιπέδων, κάθε ένα από τα οποία περιλαμβάνει ένα σύνολο συστάδων. Η επιλογή του κατάλληλου συνόλου συστάδων εναπόκειται στον χρήστη. Η ιεραρχία των επιπέδων και οι αντίστοιχες συστάδες αναπαριστώνται γραφικά με τη χρήση δενδρογραμμάτων. Οι Ιεραρχικές μέθοδοι υποδιαιρούνται σε συσσωρευτικές, οι οποίες δημιουργούν την ιεραρχία μέσα από μια διαδικασία διαδοχικών συγχωνεύσεων, και σε διαιρετικές, οι οποίες δημιουργούν την ιεραρχία μέσω διαδοχικών διασπάσεων. Για τη συγχώνευση ή διάσπαση συστάδων απαιτείται καθορισμός της απόστασης τους. Έχουν προταθεί διάφοροι τρόποι μέτρησης της απόστασης των συστάδων. Οι βασικότεροι από αυτούς είναι η μέθοδος της Απλής Σύνδεσης, η μέθοδος της Πλήρους Σύνδεσης, η Σύνδεση Μέσου Όρου, η μέθοδος Ward κλπ. Οι μέθοδοι αυτές παρουσιάζονται αναλυτικά. Στη Διαχωριστική ΑΣ, τα αντικείμενα επιμερίζονται σε k συστάδες. Τυπικά, ο αριθμός των συστάδων προκαθορίζεται από τον χρήστη. Στη συνέχεια εφαρμόζεται μια επαναληπτική διαδικασία, κατά την οποία τα αντικείμενα μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετράται με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη, και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. Ο πιο γνωστός αλγόριθμος Διαχωριστικής ΑΣ είναι ο k-Means. Στις βασισμένες στην πυκνότητα μεθόδους, ελέγχεται η πυκνότητα των αντικειμένων, και η συστάδα επεκτείνεται, όσο η γειτονιά των παρακείμενων σημείων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι πλέγματος επιμερίζουν τον χώρο των δεδομένων σε διακριτά κελιά, τα οποία συγκροτούν ένα πλέγμα, και η αναζήτηση των συστάδων γίνεται στα κελιά του πλέγματος. Τέλος, στις βασισμένες στα μοντέλα μεθόδους, γίνεται χρήση μοντέλων, με στόχο τη βελτιστοποίηση της προσαρμογής ανάμεσα στα δεδομένα και τα μοντέλα. Μια πολύ διαδεδομένη μέθοδος αυτής της κατηγορίας είναι οι Αυτοοργανούμενοι Χάρτες. Οι Αυτοοργανούμενοι Χάρτες είναι ένας ειδικός τύπος Νευρωνικού Δικτύου με ένα επίπεδο. Οι νευρώνες είναι χωρικά διατεταγμένοι σε ένα πλέγμα, και περιέχουν ένα διάνυσμα ίδιων διαστάσεων με τα δεδομένα εισόδου. Με κατάλληλη, μη επιβλεπόμενη μάθηση, όμοια δεδομένα εισόδου αντιστοιχίζονται με γειτονικούς νευρώνες του πλέγματος. Οι Αυτοοργανούμενοι Χάρτες παρουσιάζονται αναλυτικά στο υποκεφάλαιο 11.6. Στο τέλος του παρόντος κεφαλαίου γίνεται μια σύντομη αναφορά στις εφαρμογές της ΑΣ στη σύγχρονη επιχείρηση

Προηγούμενη Γνώση

Το παρόν Κεφάλαιο εισάγει τον αναγνώστη στη θεματική ενότητα της Ανάλυσης Συστάδων, η οποία είναι αυτόνομη και για τον λόγο αυτό δεν απαιτούνται ιδιαίτερες προηγούμενες γνώσεις. Ωστόσο, για την καλύτερη κατανόηση των περιεχομένων θα συνιστούσαμε την προηγούμενη ανάγνωση του Κεφαλαίου 6, το οποίο αποτελεί εισαγωγή στην Εξόρυξη Δεδομένων. Η θεματική ενότητα της Ανάλυσης Συστάδων είναι πολύ εκτεταμένη και φυσικά είναι αδύνατο να καλυφθεί στα πλαίσια ενός κεφαλαίου. Υπάρχουν πολλά εξειδικευμένα βιβλία, τα οποία ασχολούνται αποκλειστικά με την Ανάλυση Συστάδων. Στα βιβλία αυτά, ο αναγνώστης μπορεί να βρει πρόσθετες μεθόδους ΑΣ, καθώς και παρουσίαση τεχνικών για την ομαδοποίηση δεδομένων ειδικού τύπου, όπως κειμένων, δικτύων, πολυμέσων κλπ. Ενδεικτικά αναφέρουμε το βιβλίο των Aggarwal and Reddy (2014).

11.1 Εισαγωγή

Η Ανάλυση Συστάδων (ΑΣ) (Clustering) είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Σε γενικές γραμμές, η ΑΣ αφορά την ένταξη οντοτήτων σε ομοειδείς ομάδες. Η δραστηριότητα αυτή είναι εγγενής στους ανθρώπους, και εκτελείται αυθόρμητα από την παιδική τους ηλικία. Ένας άνθρωπος σε πρωτόγονες συνθήκες, αλλά με σχετική εμπειρία, κατανοεί αυθόρμητα ομάδες, όπως δένδρα, πουλιά κλπ. (Kruskal, 1977). Στην επιστημονική ΑΣ, οι ομάδες εξάγονται βάσει αλγορίθμων από τα δεδομένα.

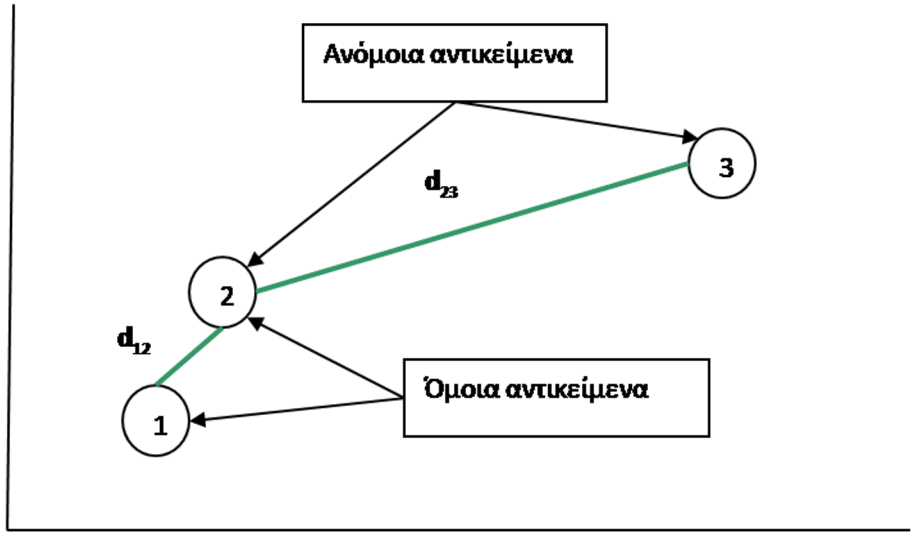
Στόχος της ΑΣ είναι ο επιμερισμός ενός συνόλου παραδειγμάτων σε υποσύνολα. Τα υποσύνολα καλούνται συστάδες. Για τον επιμερισμό, καθοριστικό ρόλο παίζει η **ομοιότητα**. Τα παραδείγματα μιας συστάδας «μοιάζουν» μεταξύ τους, ενώ «δεν μοιάζουν» με τα παραδείγματα των άλλων συστάδων. Ένα σχετικά συγγενές αντικείμενο είναι η Κατηγοριοποίηση, η οποία στοχεύει στην πρόβλεψη της κατηγορίας κάθε παρατήρησης. Όμως οι διαφορές ανάμεσα στην ΑΣ και την Κατηγοριοποίηση είναι πολλές. Στην Κατηγοριοποίηση, οι κατηγορίες είναι γνωστές εκ των προτέρων. Στα δεδομένα υπάρχει ένα γνώρισμα, το γνώρισμα της κλάσης, στο οποίο καταγράφεται η κατηγορία της εκάστοτε παρατήρησης. Οι αλγόριθμοι μοντελοποιούν τις σχέσεις ανάμεσα στο γνώρισμα της κλάσης και στα υπόλοιπα γνωρίσματα. Η Κατηγοριοποίηση είναι μια μορφή **εκπαίδευσης μέσω παραδειγμάτων** (learning by examples). Το γεγονός ότι υπάρχει εκ των προτέρων γνώση σχετικά με τις κατηγορίες, και ότι η γνώση αυτή καθοδηγεί τη διαδικασία εκπαίδευσης, χαρακτηρίζει την Κατηγοριοποίηση ως επιβλεπόμενη μάθηση (supervised learning). Στην ΑΣ δεν υπάρχει κάποιο γνώρισμα στο οποίο να καταγράφεται η κλάση των παραδειγμάτων, και οι συστάδες δεν είναι γνωστές εκ των προτέρων. Αντιθέτως, το ζητούμενο είναι να εντοπιστούν συστάδες και να ενταχθούν τα παραδείγματα στην κατάλληλη συστάδα. Οι συστάδες συγκροτούνται στη βάση της ομοιότητας των μελών τους. Για τον λόγο αυτό, η ΑΣ θεωρείται μια μορφή **εκπαίδευσης μέσω παρατήρησης** (learning by observation). Επίσης, το γεγονός ότι δεν υπάρχει εκ των προτέρων γνώση χαρακτηρίζει την ΑΣ ως **μη επιβλεπόμενη μάθηση** (unsupervised learning). Ο κύριος σκοπός της Κατηγοριοποίησης είναι η διατύπωση προβλέψεων (predictive), ενώ ο κύριος σκοπός της ΑΣ είναι περιγραφικός (descriptive). Σε διαδικαστικό επίπεδο, η ΑΣ αντιμετωπίζει όλα τα γνωρίσματα ισότιμα και τα χρησιμοποιεί για τον υπολογισμό της ομοιότητας των παρατηρήσεων. Αντιθέτως, η Κατηγοριοποίηση χρησιμοποιεί τα υπόλοιπα γνωρίσματα για να προβλέψει τις τιμές του γνωρίσματος της κλάσης.

Στα πλαίσια της Εξόρυξης Δεδομένων, η ΑΣ έχει πολλαπλή χρησιμότητα. Ως αυτόνομη αναλυτική εργασία, επιτρέπει στον αναλυτή να επιμερίσει τα δεδομένα σε ομάδες ομοειδών παρατηρήσεων. Ακολούθως, ο αναλυτής μπορεί να επικεντρωθεί στην εκάστοτε ομάδα, να αναγνωρίσει τα κοινά χαρακτηριστικά της, και να εξάγει γνώση χρήσιμη για τη λήψη αποφάσεων. Η πιο γνωστή εφαρμογή της ΑΣ στις επιχειρηματικές διαδικασίες είναι στη διαφήμιση, και ειδικότερα για την τμηματοποίηση της αγοράς. Ο όρος τμηματοποίηση της αγοράς περιγράφει τον επιμερισμό των καταναλωτών σε ομάδες με όμοια καταναλωτική συμπεριφορά. Η τμηματοποίηση της αγοράς είναι κεφαλαιώδους σημασίας για το μάρκετινγκ. Οι διαφημίσεις μαζικής απεύθυνσης έχουν υψηλό κόστος και μικρό ποσοστό ανταπόκρισης. Με τον εντοπισμό ομάδων όμοιων καταναλωτών μπορούν να σχεδιαστούν διαφημιστικές εκστρατείες προσαρμοσμένες στα ιδιαίτερα χαρακτηριστικά της κάθε ομάδας. Η στοχευμένη σε συγκεκριμένες ομάδες διαφήμιση κοστίζει λιγότερο και επιτυγχάνει σημαντικά υψηλότερα ποσοστά ανταπόκρισης των καταναλωτών. Πέρα από την αξία της ως αυτόνομο εργαλείο ανάλυσης, η ΑΣ μπορεί να συνδυαστεί με άλλες εργασίες Εξόρυξης Δεδομένων και να αποτελέσει στάδιο προεπεξεργασίας. Χάρη στην ικανότητα των αλγορίθμων της να ομαδοποιούν τις παρατηρήσεις σύμφωνα με την ομοιότητα τους, μπορεί να χρησιμοποιηθεί για τον εντοπισμό παρατηρήσεων με ακραίες τιμές (outliers) (Ng & Han, 1994; Shekhar & Chawla, 2003). Οι ακραίες παρατηρήσεις θα απομακρυνθούν από το σύνολο δεδομένων, ώστε να προκύψει ένα βελτιωμένο σύνολο εκπαίδευσης. Επίσης, οι συστάδες, οι οποίες θα προκύψουν, μπορούν να θεωρηθούν κατηγορίες. Σε ακόλουθο στάδιο, μπορεί να εκτελεστεί κατηγοριοποίηση για την ανάπτυξη μοντέλων ικανών να προβλέπουν την κατηγορία. Συνδυασμός μεθόδων ΑΣ και Κατηγοριοποίησης μπορεί να αποφέρει υβριδικούς κατηγοριοποιητές.

11.2 Ομοιότητα και απόσταση

Εφόσον στην ΑΣ οι παρατηρήσεις ομαδοποιούνται σύμφωνα με την ομοιότητα τους, είναι φανερό ότι ένα από τα βασικότερα ζητήματα είναι ο καθορισμός μέτρων ομοιότητας. Ένας τρόπος καθορισμού του βαθμού ομοιότητας δύο παρατηρήσεων είναι με τη χρήση της **απόστασης** τους. Ας θεωρήσουμε αρχικά μια απλή περίπτωση, όπου οι παρατηρήσεις αποτελούνται από δύο μόνο γνωρίσματα X και Y και ότι και τα δύο γνωρίσματα παίρνουν αριθμητικές τιμές. Κάθε παρατήρηση μπορεί να αναπαρασταθεί στον δισδιάστατο χώρο X,Y ως ένα σημείο. Δύο σημεία, τα οποία βρίσκονται κοντά στον δισδιάστατο χώρο, θεωρούνται όμοια, ενώ δύο σημεία,

τα οποία βρίσκονται μακριά στον δισδιάστατο χώρο, θεωρούνται ανόμοια. Στο Σχήμα 11.1 απεικονίζονται τρία σημεία στον δισδιάστατο χώρο. Τα σημεία 1 και 2 θεωρούνται όμοια, ενώ τα σημεία 2 και 3 θεωρούνται ανόμοια.



Σχήμα 11.1 *Ομοιότητα με χρήση απόστασης*

Εάν οι παρατηρήσεις έχουν *n* γνωρίσματα, τότε θεωρούνται σημεία στον χώρο των *n* διαστάσεων, και η ομοιότητα τους υπολογίζεται από την απόσταση τους σε αυτόν τον χώρο. Για τον υπολογισμό της απόστασης υπάρχει διαφοροποίηση ανάλογα με το εάν τα γνωρίσματα περιέχουν αριθμητικές, δυαδικές, ή ονομαστικές τιμές.

11.2.1 Απόσταση με αριθμητικά γνωρίσματα

Εάν όλα τα γνωρίσματα των παρατηρήσεων έχουν αριθμητικές τιμές, τότε ως μέτρο ομοιότητας δύο παρατηρήσεων x_a και x_b μπορεί να χρησιμοποιηθεί η Ευκλείδεια απόσταση. Θεωρούμε ότι οι παρατηρήσεις έχουν *n* γνωρίσματα. Η απόσταση μεταξύ των σημείων x_a και x_b συμβολίζεται ως $d(x_a,x_b)$. Η Ευκλείδεια απόσταση των σημείων x_a και x_b δίνεται από την Εξίσωση 11.1

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} (x_{aj} - x_{bj})^2}$$

(11.1)

όπου xaj είναι η τιμή της μεταβλητής *j* της παρατήρησης x_a.

Η Ευκλείδεια απόσταση είναι η πιο διαδεδομένη, ωστόσο δεν είναι η μοναδική. Μια παραλλαγή της, η οποία χρησιμοποιείται συχνά, είναι η απόσταση Manhattan. Η απόσταση Manhattan ορίζεται από την Εξίσωση 11.2

$$d(x_a, x_b) = \sum_{j=1}^{n} |x_{aj} - x_{bj}|$$

(11.2)

Γενίκευση της Ευκλείδειας απόστασης και της απόστασης Manhattan είναι η απόσταση Minkowski, η

οποία ορίζεται από την Εξίσωση 11.3.

$$d(x_a, x_b) = \sqrt[q]{\sum_{j=1}^{n} |x_{aj} - x_{bj}|^q}$$

(11.3)

Η Ευκλείδεια απόσταση ταυτίζεται με την Minkowski για *q=2*, ενώ η απόσταση Manhattan ταυτίζεται με την Minkowski για *q=1*.

Από τους παραπάνω ορισμούς προκύπτει ότι η απόσταση ενός σημείου από τον εαυτό του είναι ίση με μηδέν, και ότι η απόσταση είναι ένας θετικός αριθμός. Επίσης, η απόσταση μεταξύ δύο σημείων είναι συντομότερη από οποιαδήποτε άλλη διαδρομή, η οποία συνδέει τα δύο αυτά σημεία μέσω ενός τρίτου σημείου.

Η Ευκλείδεια απόσταση δεν επηρεάζεται από την προσθήκη νέων παρατηρήσεων και αποδίδει καλά όταν στα δεδομένα υπάρχουν συμπαγείς ή απομονωμένες συστάδες (Mao & Jain, 1996). Ένα μειονέκτημα της Ευκλείδειας απόστασης είναι ότι μεταβολή των μονάδων μέτρησης μιας μεταβλητής (πχ έκφραση μιας απόστασης από χιλιόμετρα σε μέτρα ή μετατροπή ενός χρηματικού ποσού από ευρώ σε γιεν) επηρεάζει σημαντικά την απόσταση, και μπορεί να οδηγήσει σε σημαντικά διαφορετικές συστάδες. Επίσης, οι μεταβλητές, οι οποίες παίρνουν μεγαλύτερες τιμές ή που παρουσιάζουν μεγάλες διαφορές τιμών μεταξύ των παρατηρήσεων, επηρεάζουν δυσανάλογα την απόσταση. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η κανονικοποίηση των τιμών, με χρήση κάποιας τεχνικής όπως η Z-Score. Η τεχνική Z-Score και άλλες τεχνικές κανονικοποίησης παρουσιάζονται στο Κεφάλαιο 7.

Στον υπολογισμό της Ευκλείδειας απόστασης όλα τα γνωρίσματα θεωρούνται ισότιμα. Ωστόσο, ο αναλυτής πιθανόν να επιθυμεί να προσδώσει ιδιαίτερη βαρύτητα σε ορισμένα από αυτά. Αυτό μπορεί να επιτευχθεί με εκχώρηση συντελεστών βαρύτητας στα γνωρίσματα, έτσι ώστε η διαφορά των τιμών δυο αντικειμένων για το συγκεκριμένο γνώρισμα να πολλαπλασιάζεται με τον συντελεστή βαρύτητας του γνωρίσματος. Με τη χρήση συντελεστών βαρύτητας, η Ευκλείδεια απόσταση υπολογίζεται σύμφωνα με την Εξίσωση 11.4

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^{n} w_j * (x_{aj} - x_{bj})^2}$$

(11.4)

όπου w_j ο συντελεστής βαρύτητας του γνωρίσματος j.

Η ύπαρξη γραμμικής συσχέτισης μεταξύ των γνωρισμάτων μπορεί να προκαλέσει στρεβλώσεις στον υπολογισμό της απόστασης. Το πρόβλημα αυτό μπορεί να μετριαστεί με τη χρήση της απόστασης Mahalanobis. Η απόσταση Mahalanobis λαμβάνει υπόψη της τις συσχετίσεις μεταξύ των γνωρισμάτων, και δεν επηρεάζεται από την κλίμακα μέτρησης των μεταβλητών. Ο υπολογισμός της γίνεται σύμφωνα με την Εξίσωση 11.5

$$d(x_a, x_b) = (x_a - x_b)^T S^{-1} (x_a - x_b)$$

(11.5)

όπου *S* ο πίνακας συνδιασποράς.

11.2.2 Απόσταση με δυαδικά γνωρίσματα

Η Ευκλείδεια απόσταση είναι ένα κατάλληλο μέτρο ομοιότητας, όταν οι παρατηρήσεις αποτελούνται από γνωρίσματα αριθμητικών τιμών. Σε μια βάση δεομένων όμως υπάρχουν και πεδία άλλων τύπων, όπως δυαδικά πεδία και ονομαστικά πεδία. Για παρατηρήσεις με γνωρίσματα άλλων τύπων έχουν προταθεί άλλα, πιο κατάλληλα μέτρα ομοιότητας.

Τα δυαδικά (binary) γνωρίσματα δέχονται δύο δυνατές τιμές, την τιμή 0 και την τιμή 1. Ένα δυαδικό γνώρισμα μπορεί να αναπαριστά μια πληροφορία, όπου οι δύο δυνατές τιμές κωδικοποιούν δύο καταστάσεις ίσης αξίας ή σημασίας. Παράδειγμα τέτοιου γνωρίσματος είναι το πεδίο «Φύλο», όπου το 1 συμβολίζει το «άρρεν»

και το 0 συμβολίζει το «θήλυ». Μια τέτοια δυαδική μεταβλητή ονομάζεται **συμμετρική**. Υπάρχουν όμως δυαδικές μεταβλητές, όπου οι δύο καταστάσεις τις οποίες συμβολίζουν οι τιμές 0 και 1 δεν είναι ισότιμες. Συνήθως τέτοιες δυαδικές μεταβλητές καταγράφουν την ύπαρξη ή την ανυπαρξία συμβάντος (πχ χρεοκοπία επιχείρησης ή μη χρεοκοπία). Κατά κανόνα η ύπαρξη συμβάντος είναι πιο σπάνια και κωδικοποιείται με την τιμή 1. Σε τέτοιου τύπου δυαδικές μεταβλητές, οι δύο καταστάσεις δεν είναι ίσης αξίας. Οι μεταβλητές αυτές καλούνται μη **συμμετρικές**.

Θεωρούμε ένα σύνολο δεδομένων που περιέχει μόνο δυαδικές μεταβλητές. Τα x_a και x_b είναι δύο αντικείμενα αυτού του συνόλου. Τα x_a και x_b μπορούν να έχουν σε μια μεταβλητή ίδιες τιμές (1 ή 0) ή διαφορετικές τιμές (1 και 0 ή 0 και 1). Η συμφωνία τιμών των δύο παρατηρήσεων στις διάφορες μεταβλητές τους έχει ως ακολούθως:

- *k* είναι το πλήθος των μεταβλητών όπου και το x_a και το x_b έχουν την τιμή 1,
- *l* είναι το πλήθος των μεταβλητών όπου το x_a έχει την τιμή 1, ενώ το x_b έχει την τιμή 0,
- *m* είναι το πλήθος των μεταβλητών όπου το x_a έχει την τιμή 0, ενώ το x_b έχει την τιμή 1,
- *n* είναι το πλήθος των μεταβλητών όπου και το x_a και το x_b έχουν την τιμή 0.

Τα παραπάνω συνοψίζονται στον πίνακα συνάφειας που παρουσιάζεται στο Σχήμα 11.2

		Αντικείμενο x_b		
		1	0	Άθροισμα
Αντικείμενο x_a	1	k	l	k+l
	0	m	n	m+n
	Άθροισμα	k+m	l+n	k+l+m+n

Σχήμα 11.2 *Πίνακας συνάφειας με δυαδικές μεταβλητές*

Εάν οι δυαδικές μεταβλητές είναι συμμετρικές, τότε η απόσταση των αντικειμένων x_a και x_b δίνεται από τον συντελεστή **simple matching** (simple matching coefficient), ο οποίος ορίζεται με την Εξίσωση 11.6.

$$d(x_a x_b) = \frac{l + m}{k + l + m + n} \tag{11.6}$$

Εάν οι δυαδικές μεταβλητές δεν είναι συμμετρικές, η σύμπτωση τιμών ίσων με 0 είναι μικρότερης σημασίας. Για τον λόγο αυτό, έχουν προταθεί πιο κατάλληλα μέτρα. Το πιο γνωστό μέτρο είναι ο συντελεστής **Jaccard**, ο οποίος ορίζεται με την Εξίσωση 11.7.

$$d(x_a x_b) = \frac{l + m}{k + l + m} \tag{11.7}$$

11.2.3 Απόσταση με ονομαστικά γνωρίσματα

Ονομαστικά (nominal) καλούνται τα γνωρίσματα τα οποία δέχονται ονομαστικές τιμές, δηλαδή λέξεις. Ένα ονομαστικό γνώρισμα μπορεί να λάβει ένα πεπερασμένο πλήθος τιμών. Οι τιμές αυτές δεν υποδηλώνουν κάποια μορφή εσωτερικής ιεράρχησης, όπως συμβαίνει στα διατακτικά ονομαστικά γνωρίσματα. Παράδειγμα ονομαστικού γνωρίσματος είναι η κατηγορία προϊόντος (τρόφιμα, είδη ένδυσης, είδη καλλωπισμού κλπ.).

Θεωρούμε δύο αντικείμενα x_a και x_b με *n* ονομαστικά γνωρίσματα. Τα δύο αντικείμενα έχουν σε *m* γνωρίσματα τις ίδιες τιμές. Η απόσταση των δύο αντικειμένων μπορεί να υπολογιστεί με τρόπο αντίστοιχο με τον συντελεστή simple matching (Εξίσωση 11.8)

$$d(x_a x_b) = \frac{n-m}{n}$$

(11.8)

Ένας εναλλακτικός τρόπος υπολογισμού της απόστασης είναι με την εισαγωγή ψευδομεταβλητών (dummy variables). Για μια ονομαστική μεταβλητή, η οποία μπορεί να δεχτεί *k* διαφορετικές τιμές, δημιουργούμε *k* ψευδομεταβλητές, μία για κάθε δυνατή τιμή. Για παράδειγμα, αν η κατηγορία προϊόντος μπορεί να δεχτεί μόνο τις τρείς τιμές που αναφέρθηκαν προηγουμένως, τότε δημιουργούμε μια μεταβλητή «Τρόφιμα», μια μεταβλητή «Είδη Ένδυσης» και μια μεταβλητή «Είδη καλλωπισμού». Εάν ένα αντικείμενο έχει μια συγκεκριμένη τιμή στην ονομαστική μεταβλητή, τότε η αντίστοιχη ψευδομεταβλητή παίρνει την τιμή 1, ενώ οι υπόλοιπες ψευδομεταβλητές την τιμή 0. Για παράδειγμα, αν ένα αντικείμενο έχει στην ονομαστική μεταβλητή «Κατηγορία Προϊόντος» την τιμή «Τρόφιμα», τότε η ψευδομεταβλητή «Τρόφιμα» θα πάρει την τιμή 1, ενώ οι ψευδομεταβλητές «Είδη Ένδυσης» και «Είδη καλλωπισμού» θα πάρουν την τιμή 0. Με τον τρόπο αυτό, μια ονομαστική μεταβλητή μετατρέπεται σε πολλές δυαδικές μεταβλητές. Μετά τη μετατροπή, ο υπολογισμός της απόστασης μπορεί να γίνει με τον τρόπο που περιγράφτηκε για τα γνωρίσματα δυαδικών τιμών.

11.2.4 Απόσταση με διατακτικά γνωρίσματα

Τα διατακτικά (ordinal) γνωρίσματα δέχονται τιμές, οι οποίες υποδηλώνουν μια θέση σε μια διάταξη ή σειρά. Ένα παράδειγμα διατακτικών τιμών είναι οι δείκτες αξιολόγησης της πιστοληπτικής ικανότητας, τους οποίους εκδίδουν οι οίκοι αξιολόγησης. Η πιστοληπτική ικανότητα ενός φορέα βαθμολογείται με μια τιμή της μορφής AAA, AA, A, BBB, BB, B, CCC, CC, C, D(edault) ή κάποια παραλλαγή της. Οι τιμές είναι λεκτικές, δηλώνουν όμως μια θέση σε μια ποιοτική διαβάθμιση.

Το γεγονός ότι οι διατακτικές τιμές υποδηλώνουν μια σειρά μας επιτρέπει να τις χειριστούμε σαν αριθμητικές τιμές. Αν ένα διατακτικό γνώρισμα δέχεται n τιμές, τότε η τιμή που δηλώνει τη χαμηλότερη θέση στη σειρά μπορεί να αντικατασταθεί με τον αριθμό 1, η επόμενη τιμή με τον αριθμό 2, μέχρι την τελευταία, η οποία θα αντικατασταθεί με τον αριθμό *n*. Η προσέγγιση αυτή έχει το μειονέκτημα ότι μεταβλητές με πολλές διατακτικές τιμές μπορούν να δημιουργήσουν μεγάλες διαφορές μεταξύ αντικειμένων, και οι διαφορές αυτές να επηρεάσουν δυσανάλογα την απόσταση. Για τον λόγο αυτό, οι αριθμητικές τιμές κανονικοποιούνται και ανάγονται στο διάστημα [0,0..1,0]. Ο μετασχηματισμός των τιμών γίνεται σύμφωνα με την Εξίσωση 11.9

$$m_{new} = \frac{m-1}{n-1}$$

(11.9)

όπου m_{new} είναι η νέα τιμή, *m* είναι η τιμή πριν την κανονικοποίηση και *n* είναι το πλήθος δυνατών τιμών της διατακτικής μεταβλητής. Στον πίνακα 11.1 παρουσιάζεται ο μετασχηματισμός των τιμών για τους δείκτες πιστοληπτικής ικανότητας

Διατακτικές τιμές	Αριθμητικές τιμές	Κανονικοποιημένες τιμές
AAA	10	1,00
AA	9	0,89
A	8	0,78
BBB	7	0,67
BB	6	0,56
B	5	0,44
CCC	4	0,33
CC	3	0,22
C	2	0,11
D	1	0,00

Πίνακας 11.1 *Μετασχηματισμός διατακτικών τιμών δείκτη πιστοληπτικής ικανότητας*

Μετά τον μετασχηματισμό των διατακτικών τιμών και την αντιστοίχιση τους σε αριθμητικές τιμές της περιοχής [0,0..1,0], μπορεί να γίνει υπολογισμός της απόστασης δύο αντικειμένων με τη χρήση της Ευκλείδειας απόστασης ή κάποιας παραλλαγής της.

11.2.5 Απόσταση με μεικτών τύπων γνωρίσματα

Όλοι οι υπολογισμοί αποστάσεων, οι οποίοι αναφέρθηκαν μέχρι αυτό το σημείο, θεωρούν ότι όλα τα γνωρίσματα είναι του ιδίου τύπου. Σε πραγματικές βάσεις δεδομένων όμως τα πεδία είναι διαφόρων τύπων, δηλαδή αριθμητικά, δυαδικά, ονομαστικά κλπ. Για τον υπολογισμό της απόστασης αντικειμένων με διάφορους τύπους γνωρισμάτων μπορεί να γίνει συνδυασμός των προηγούμενων τεχνικών.

Η απόσταση δύο αντικειμένων x_a και x_b με n γνωρίσματα διαφόρων τύπων μπορεί να υπολογιστεί σύμφωνα με την Εξίσωση 11.10.

$$d(x_a, x_b) = \frac{\sum_{j=1}^{n} \delta_{abj} \Delta_{abj}}{\sum_{j=1}^{n} \delta_{abj}}$$

(11.10)

Το δ_{abj} παίρνει τιμές ως ακολούθως:

- Τιμή = 0 εάν η τιμή του x_a (x_{aj}) ή του x_b (x_{bj}) στη μεταβλητή j λείπει.
- Τιμή = 0 εάν η μεταβλητή j είναι μη συμμετρική και η τιμή των x_a και x_b στη μεταβλητή j είναι ίση με 0 ($x_{aj} = x_{bj} = 0$).
- Τιμή = 1 σε οποιαδήποτε άλλη περίπτωση.

Ο υπολογισμός της τιμής του Δ_{abj} εξαρτάται από τον τύπο της μεταβλητής j:

- Εάν η μεταβλητή j είναι δυαδική ή ονομαστική το Δ_{abj} παίρνει την τιμή 0 εάν $x_{aj} = x_{bj.}$ Διαφορετικά παίρνει την τιμή 1.
- Εάν η μεταβλητή j είναι αριθμητική, τότε το Δ_{abj} υπολογίζεται σύμφωνα με την Εξίσωση 11.11, όπου maxj είναι η μέγιστη τιμή του γνωρίσματος j και min_j είναι η ελάχιστη τιμή του γνωρίσματος j.

$$\Delta_{abj} = \frac{|x_{aj} - x_{bj}|}{max_j - min_j}$$

(11.11)

- Εάν η μεταβλητή j είναι διατακτική, τότε οι τιμές της μετασχηματίζονται και ανάγονται στην περιοχή [0,0..1,0] και ακολούθως το Δ_{abj} υπολογίζεται με τρόπο αντίστοιχο των αριθμητικών μεταβλητών.

11.3 Κατηγορίες Μεθόδων ΑΣ

Η επιστημονική βιβλιογραφία περιλαμβάνει έναν μεγάλο αριθμό διαφορετικών μεθόδων Ανάλυσης Συστάδων. Οι μέθοδοι αυτές παρουσιάζουν σημαντικές διαφορές στις επαγωγικές αρχές τους και στον τρόπο σχηματισμού των συστάδων. Ένας από τους λόγους ύπαρξης αυτής της ποικιλίας μεθόδων είναι το γεγονός ότι δεν υπάρχει ένας αυστηρός ορισμός της έννοιας της συστάδας (Estivill-Castro & Yang, 2000). Οι Han, Kamber and Pei (2011) ορίζουν πέντε κατηγορίες μεθόδων ΑΣ:

- **Ιεραρχικές μέθοδοι**. Οι ιεραρχικές μέθοδοι (hierarchical methods) δημιουργούν μια ιεραρχία από συστάδες. Στο κατώτατο επίπεδο της ιεραρχίας βρίσκονται τα μεμονωμένα αντικείμενα. Στο ανώτατο επίπεδο βρίσκεται μια υπερσυστάδα, η οποία περιλαμβάνει όλα τα αντικείμενα. Κάθε ενδιάμεσο επίπεδο ορίζει ένα σύνολο συστάδων. Η ιεραρχία προκύπτει από μια διαδικασία διαδοχικών διασπάσεων ή συγχωνεύσεων συστάδων. Η επιλογή του κατάλληλου συνόλου συστάδων εναπόκειται στον χρήστη. Αναλυτική παρουσίαση των ιεραρχικών μεθόδων γίνεται στο υποκεφάλαιο 11.4.
- **Διαχωριστικές μέθοδοι**. Οι διαχωριστικές μέθοδοι (partitioning methods) επιμερίζουν τα αντικείμε-

να σε *k* συστάδες. Τυπικά το πλήθος των συστάδων προκαθορίζεται από τον χρήστη. Στις μεθόδους αυτής της κατηγορίας εφαρμόζεται μια επαναληπτική διαδικασία, κατά την οποία τα αντικείμενα μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετράται με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. Ο πιο γνωστός αλγόριθμος διαχωριστικής ΑΣ είναι ο k-Means. Μέθοδοι και ζητήματα διαχωριστικής ΑΣ παρουσιάζονται στο υποκεφάλαιο 11.5.

- **Μέθοδοι βασισμένες στην πυκνότητα**. Στις βασισμένες στην πυκνότητα μεθόδους (density based methods) ελέγχεται η πυκνότητα των αντικειμένων στον χώρο και δημιουργούνται συστάδες, οι οποίες καλύπτουν τις πυκνές περιοχές. Για κάθε παρατήρηση που ανήκει σε μια συστάδα, η γειτονιά της, η οποία είναι καθορισμένης διαμέτρου, πρέπει να περιλαμβάνει έναν ελάχιστο αριθμό παρατηρήσεων. Η συστάδα συνεχίζει να επεκτείνεται όσο η γειτονιά των παρακείμενων σημείων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι αυτές μπορούν να δημιουργήσουν συστάδες με μη κυρτά και περίπλοκα σχήματα. Επίσης, είναι ιδιαίτερα ικανές να απομονώνουν τις εξαιρέσεις.
- **Μέθοδοι πλέγματος**. Οι μέθοδοι πλέγματος (grid based methods) επιμερίζουν τον χώρο των δεδομένων σε διακριτά κελιά, τα οποία συγκροτούν ένα πλέγμα. Τα αντικείμενα πλέον αντιπροσωπεύονται από τα κελιά στα οποία ανήκουν. Η αναζήτηση των συστάδων γίνεται στα κελιά του πλέγματος και όχι στα αντικείμενα. Στις μεθόδους πλέγματος ο χρόνος επεξεργασίας εξαρτάται από το πλήθος των κελιών και όχι από το πλήθος των αντικειμένων. Επειδή κατά κανόνα ο αριθμός των κελιών είναι πολύ μικρότερος από τον αριθμό των αντικειμένων, οι μέθοδοι αυτές είναι σημαντικά ταχύτερες. Ένα σημαντικό ζήτημα είναι ο καθορισμός κελιών κατάλληλου μεγέθους.
- **Μέθοδοι βασισμένες σε μοντέλα**. Στις βασισμένες σε μοντέλα μεθόδους (model based methods), όπως υπονοεί το όνομα τους, γίνεται χρήση μοντέλων. Στόχος τους είναι να βελτιστοποιηθεί η προσαρμογή ανάμεσα στα δεδομένα και στα μοντέλα. Το μοντέλο εκπαιδεύεται με μη επιβλεπόμενη μάθηση σχετικά με τη συμμετοχή των παρατηρήσεων σε συστάδες. Μια πολύ διαδεδομένη μέθοδος αυτής της κατηγορίας είναι ένα ειδικός τύπος νευρωνικών δικτύων, που ονομάζονται Αυτοοργανούμενοι Χάρτες (Self Organizing Maps). Οι Αυτοοργανούμενοι Χάρτες παρουσιάζονται στο υποκεφάλαιο 11.7.

11.4 Ιεραρχική Ανάλυση Συστάδων

Η Ιεραρχική ΑΣ συνίσταται σε μια διαδικασία διαδοχικών συγχωνεύσεων ή διασπάσεων συστάδων. Οι σχετικές τεχνικές αντιστοίχως χωρίζονται σε συσσωρευτικές και διαιρετικές.

Οι **συσσωρευτικές** (agglomerative) μέθοδοι αρχικά θεωρούν κάθε ξεχωριστό αντικείμενο ως μια συστάδα. Τα πιο όμοια αντικείμενα επιλέγονται και συγχωνεύονται, δημιουργώντας μια νέα συστάδα. Από τις συστάδες που προκύπτουν, επιλέγονται οι πιο όμοιες και συγχωνεύονται. Η διαδικασία επαναλαμβάνεται μέχρι να ενταχθούν όλα τα αντικείμενα σε μια ενιαία συστάδα. Οι συσσωρευτικές μέθοδοι έχουν ως αφετηριακό σημείο το κατώτερο επίπεδο της ιεραρχίας των διαδοχικών συγχωνεύσεων, και σταδιακά ανέρχονται τα επίπεδα. Υιοθετούν δηλαδή μια προσέγγιση «από κάτω προς τα επάνω» (bottom up).

Οι **διαιρετικές** (divisive) μέθοδοι αρχικά θεωρούν όλα τα αντικείμενα ως μέλη μιας ενιαίας συστάδας. Η αρχική αυτή συστάδα διαιρείται σε δύο υποομάδες. Η διάσπαση γίνεται με τέτοιον τρόπο, ώστε οι υποομάδες οι οποίες θα προκύψουν θα έχουν τη μεγαλύτερη ανομοιότητα. Η διαδικασία των διαδοχικών διασπάσεων επαναλαμβάνεται μέχρι κάθε αντικείμενο να αποτελεί μια ξεχωριστή υποομάδα. Οι διαιρετικές μέθοδοι έχουν αφετηριακό σημείο το ανώτατο επίπεδο της ιεραρχίας και ακολουθούν μια προσέγγιση «από επάνω προς τα κάτω» (top down).

Για την επιλογή των συστάδων δημιουργείται ένας πίνακας ανομοιότητας. Εάν τα δεδομένα περιέχουν *N* σημεία, τότε ο πίνακας είναι διαστάσεων *NXN*. Κάθε εγγραφή του πίνακα είναι ένα μέτρο ανομοιότητας ή απόστασης μεταξύ δύο σημείων. Ο πίνακας ανομοιότητας έχει την ακόλουθη μορφή:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \dots & \dots & \dots & 0 & \\ d(N,1) & \dots & \dots & d(N,N-1) & 0 \end{bmatrix} \quad (11.12)$$

όπου $d(x_1,x_2)$ είναι η απόσταση μεταξύ των σημείων x_1 και x_2. Εφόσον η απόσταση κάθε σημείου από τον εαυτό του είναι μηδενική ($d(x_i,x_i)=0$) οι εγγραφές της διαγωνίου από επάνω και αριστερά προς κάτω και δεξιά έχουν μηδενικές τιμές. Επειδή η απόσταση μεταξύ δύο σημείων είναι συμμετρική ($d(x_i,x_j)= d(x_j,x_i)$), η διαγώνιος χωρίζει τον πίνακα σε δύο κατοπτρικά μέρη, οπότε διατηρούνται μόνο οι εγγραφές οι οποίες βρίσκονται κάτω από τη διαγώνιο.

Στην Ιεραρχική ΑΣ δημιουργείται μια ιεραρχία, η οποία περιλαμβάνει ένα σύνολο από δυνατές συστάδες. Κάθε επίπεδο της ιεραρχίας περιγράφει ένα συγκεκριμένο τρόπο διαμοιρασμού των αντικειμένων σε συστάδες. Αποτελεί αρμοδιότητα του χρήστη να αποφασίσει πιο είναι το κατάλληλο επίπεδο, το οποίο περιγράφει έναν φυσικό τρόπο διαμοιρασμού των αντικειμένων, δηλαδή ποιες είναι οι συστάδες, οι οποίες είναι επαρκώς όμοιες μεταξύ τους. Εάν στα δεδομένα μας υπάρχουν *N* σημεία, τότε και στις δύο κατηγορίες μεθόδων υπάρχουν *N-1* επίπεδα.

Τα βασικά **πλεονεκτήματα** των Ιεραρχικών Μεθόδων είναι τα ακόλουθα:

- Οι ιεραρχικές μέθοδοι παρουσιάζουν καλή προσαρμοστικότητα. Μπορούν να εντοπίσουν καλά διαχωρισμένες, επιμήκεις και ομόκεντρες συστάδες.
- Δημιουργούν πολλαπλά επίπεδα φωλιασμένων συστάδων και επιτρέπουν στον χρήστη να επιλέξει το επίπεδο που αυτός επιθυμεί.

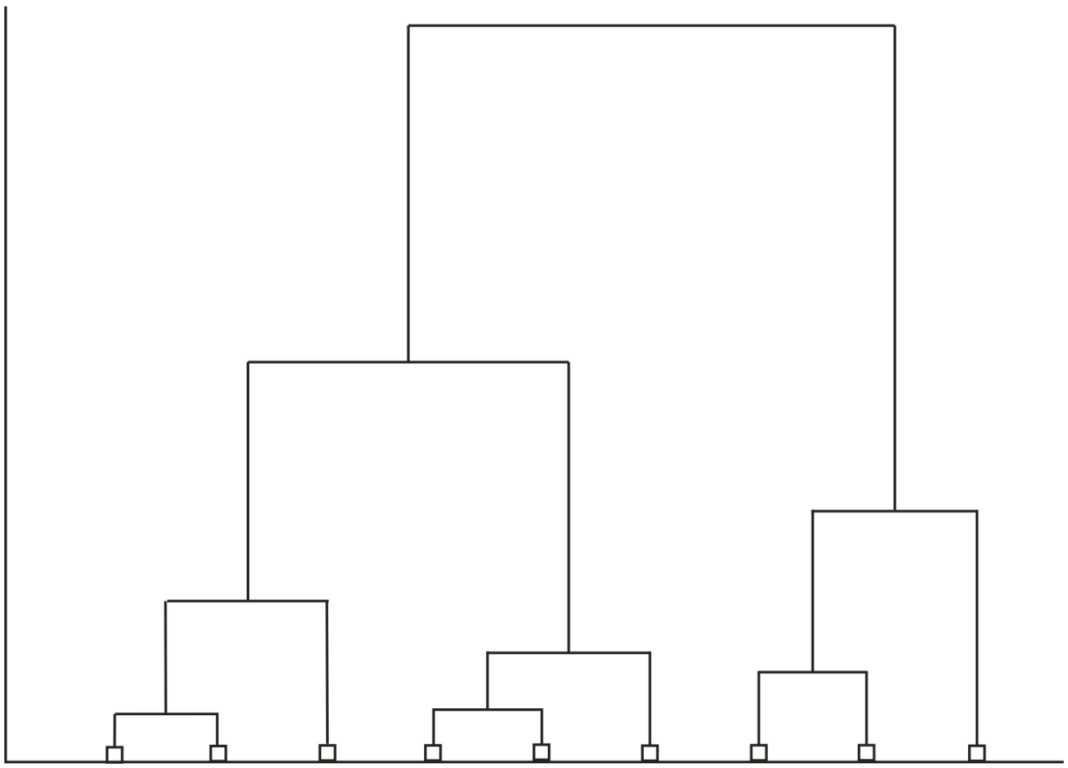
Μειονεκτήματα των Ιεραρχικών μεθόδων είναι τα εξής:

- Κάθε ενέργεια, η οποία πραγματοποιείται σε ένα στάδιο, δεν είναι αντιστρέψιμη. Από τη στιγμή που δύο αντικείμενα ενταχθούν στην ίδια ομάδα, θα παραμείνουν στην ίδια ομάδα, και δεν υπάρχει δυνατότητα να διαχωριστούν αργότερα και να ενταχθούν σε διαφορετικές ομάδες.
- Οι ιεραρχικές μέθοδοι χρειάζεται να ελέγξουν πολλές αποστάσεις, και για τον λόγο αυτό καθυστερούν όταν χρειάζεται να επεξεργαστούν μεγάλο αριθμό αντικειμένων. Το υπολογιστικό κόστος είναι τουλάχιστον $O(N^2)$ όπου *N* το πλήθος των αντικειμένων.

11.4.1 Δενδρογράμματα

Τα Δενδρογράμματα είναι ένας γραφικός τρόπος αναπαράστασης της διαδικασίας των διαδοχικών συγχωνεύσεων ή διασπάσεων. Το Δενδρόγραμμα έχει τη μορφή ανεστραμμένου δένδρου. Στα φύλλα του δένδρου, δηλαδή στο κατώτερο επίπεδο, βρίσκονται τα μεμονωμένα αντικείμενα. Κάθε κόμβος του δένδρου αντιπροσωπεύει μια συστάδα. Επίσης, κάθε κόμβος αποτελεί αφετηρία δύο κλάδων. Στη συσσωρευτική ομαδοποίηση, ένας κόμβος με τους κλάδους και τα τέκνα του συμβολίζει τη συγχώνευση των συστάδων-τέκνων, και τη δημιουργία της συστάδας-γονέα. Στη διαιρετική ομαδοποίηση, ένας κόμβος με τους κλάδους και τα τέκνα του συμβολίζει τη διάσπαση του κόμβου-γονέα, και τη δημιουργία των συστάδων-τέκνων.

Σε όλες τις συσωρευτικές μεθόδους και ορισμένες διαιρετικές μεθόδους, ο βαθμός ανομοιότητας αυξάνεται μονότονα με το επίπεδο. Ο σχεδιασμός του δένδρου γίνεται με τέτοιον τρόπο, ώστε η διαφορά ύψους των επιπέδων να αποτυπώνει την αύξηση της ανομοιότητας. Ο χρήστης μπορεί να χρησιμοποιήσει το Δενδρόγραμμα για να επιλέξει ένα επίπεδο και να αποφασίσει ένα συγκεκριμένο τρόπο διαμοιρασμού των αντικειμένων σε συστάδες.. Ωστόσο, ο χρήστης πρέπει να γνωρίζει ότι διαφορετικές μέθοδοι ιεραρχικής ομαδοποίησης ή και μικρές αλλαγές στα δεδομένα μπορούν να δημιουργήσουν σημαντικά διαφορετικά Δενδρογράμματα. Στο Σχήμα 11.3 παρουσιάζεται ένα Δενδρόγραμμα.



Σχήμα 11.3 *Δενδρόγραμμα Ιεραρχικής Ομαδοποίησης*

11.4.2 Ιεραρχική Συσσωρευτική Ανάλυση Συστάδων

*Ο*ι συσσωρευτικές μέθοδοι εκτελούν διαδοχικές συγχωνεύσεις συστάδων. Σε κάθε επανάληψη οι δύο πλησιέστερες συστάδες συνενώνονται. Ο γενικός αλγόριθμος της Ιεραρχικής Συσσωρευτικής ΑΣ έχει ως ακολούθως:

- Αρχικά, κάθε ένα από τα *N* σημεία θεωρείται ως μια ξεχωριστή συστάδα. Στον πίνακα αποστάσεων καταγράφονται οι αποστάσεις μεταξύ των σημείων.
- Εντοπίζεται στον πίνακα αποστάσεων η μικρότερη τιμή. Η τιμή αυτή είναι η απόσταση των δύο πιο όμοιων συστάδων *U* και *V (d(U,V))*.
- Οι συστάδες *U* και *V* συνενώνονται σε μια ενιαία συστάδα *UV*. Στον πίνακα αποστάσεων, διαγράφονται οι γραμμές και οι στήλες που αντιστοιχούν στις συστάδες *U* και *V*, και προστίθεται μια γραμμή και μια στήλη για τη νέα συστάδα *UV*. Επαναϋπολογίζονται οι αποστάσεις μεταξύ των συστάδων.
- Επαναλαμβάνονται τα βήματα 2 και 3, *N-1* φορές. Σε κάθε επανάληψη καταγράφονται οι συστάδες που συγχωνεύονται καθώς και οι αποστάσεις τους.

Για τον υπολογισμό της εγγύτητας των συστάδων είναι απαραίτητο ένα μέτρο. Έχουν προταθεί διάφοροι τρόποι μέτρησης της απόστασης μεταξύ των συστάδων. Εναλλακτικές μέθοδοι συσωρευτικής ΑΣ διαφοροποιούνται μεταξύ τους, ανάλογα με το μέτρο απόστασης το οποίο εφαρμόζουν. Οι κυριότεροι τρόποι μέτρησης της απόστασης είναι οι ακόλουθοι:

11.4.3 Απλή Σύνδεση

Η μέθοδος της Απλής Σύνδεσης (Simple Linkage) ονομάζεται και μέθοδος του κοντινότερου γείτονα. Σύμφωνα με αυτήν τη μέθοδο, η απόσταση ανάμεσα σε δύο συστάδες είναι η μικρότερη απόσταση από οποιοδήποτε μέλος της πρώτης συστάδας προς οποιοδήποτε μέλος της δεύτερης συστάδας (Sneath & Sokal, 1973). Με απλούστερα λόγια, η απόσταση των συστάδων είναι η απόσταση μεταξύ των δύο πλησιέστερων σημείων τους. Με μαθηματικό τρόπο η απόσταση αυτή ορίζεται από την Εξίσωση 11.13

$$d(C_1, C_2) = \min_{x_a \in C_1, x_b \in C_2} d(x_a, x_b)$$

(11.13)

όπου C_1, C_2 είναι οι δύο συστάδες, x_a, x_b είναι σημεία των συστάδων και $d(x_a,x_b)$ είναι η απόσταση μεταξύ των σημείων x_a και x_b.

Ένα σύνηθες πρόβλημα της απλής σύνδεσης είναι ότι συνενώνει συστάδες, οι οποίες έχουν δύο κοντινά σημεία και πολλά άλλα σημεία που βρίσκονται σε μεγάλες αποστάσεις. Ένα άλλο πρόβλημα είναι ότι μπορεί να προκληθεί η δημιουργία μιας επιμήκους συστάδας, και να προστίθενται διαρκώς νέα σημεία στην «ουρά» της συστάδας. Επίσης, εάν μεταξύ δύο πραγματικών συστάδων υπάρχουν μεμονωμένα σημεία που δημιουργούν μια «γέφυρα», τότε οι συστάδες αυτές θα ενωθούν. Το αποτέλεσμα αυτής της διαδικασίας είναι ότι τα σημεία που βρίσκονται στα δύο άκρα της συστάδας θα απέχουν πολύ μεταξύ τους. Το πρόβλημα αυτό είναι γνωστό ως φαινόμενο της αλυσίδας (chaining phenomenon). Πλεονέκτημα της απλής σύνδεσης είναι ότι μπορεί να εντοπίσει μη ελλειψοειδείς συστάδες.

11.4.4 Πλήρης Σύνδεση

Η μέθοδος της Πλήρους Σύνδεσης (Complete Linkage) ονομάζεται και μέθοδος του μακρινότερου γείτονα. Στη μέθοδο αυτή η λογική υπολογισμού της απόστασης των συστάδων είναι αντίστροφη από αυτήν της Απλής Σύνδεσης. Πιο συγκεκριμένα, η απόσταση μεταξύ δύο συστάδων C_1 και C_2 είναι η μεγαλύτερη απόσταση από οποιοδήποτε μέλος της C_1 προς οποιοδήποτε μέλος της C_2 (King, 1967). Με απλούστερα λόγια, η απόσταση μεταξύ δύο συστάδων είναι η απόσταση ανάμεσα στα δύο πιο απομακρυσμένα σημεία τους. Ο μαθηματικός ορισμός της πλήρους σύνδεσης δίνεται από την Εξίσωση 11.14

$$d(C_1, C_2) = \max_{x_a \in C_1, x_b \in C_2} d(x_a, x_b)$$

(11.14)

Με τη μέθοδο της πλήρους σύνδεσης αποφεύγονται προβλήματα που παρουσιάζονται με την απλή σύνδεση, όπως η δημιουργία επιμηκών συστάδων. Αντιθέτως, η πλήρης σύνδεση τείνει να δημιουργήσει συμπαγείς και σφαιρικές συστάδες με συγκρίσιμη διάμετρο. Αυτό συμβαίνει, γιατί από όλες τις υποψήφιες για συνένωση συστάδες, επιλέγει εκείνες τις δύο, οι οποίες θα δημιουργήσουν τη νέα συστάδα με τη μικρότερη διάμετρο. Το κριτήριο της μεθόδου δεν είναι τοπικό. Ολόκληρη η δομή της εκάστοτε συστάδας θα επηρεάσει την απόφαση για τη συνένωση. Η μέθοδος της πλήρους σύνδεσης ενδείκνυται, όταν γνωρίζουμε ότι αντικείμενα της ίδιας συστάδας είναι δυνατόν να βρίσκονται σε μεγάλες αποστάσεις μεταξύ τους. Ένα μειονέκτημα της μεθόδου είναι η ευαισθησία της στην ύπαρξη αντικειμένων με ακραίες τιμές. Εάν υπάρχει ένα αντικείμενο με ακραίες τιμές σε μια συστάδα, τότε δύσκολα αυτή η συστάδα θα συγχωνευθεί με κάποιαν άλλη.

11.4.5 Σύνδεση Μέσου Όρου

Η μέθοδος της Σύνδεσης Μέσου Όρου (Average Link). Σύμφωνα με αυτήν την προσέγγιση, η απόσταση δύο συστάδων είναι ίση με τη μέση απόσταση όλων των ζευγών αντικειμένων, όπου το πρώτο αντικείμενο ανήκει στην πρώτη συστάδα και το δεύτερο αντικείμενο ανήκει στη δεύτερη συστάδα (Murtagh, 1984). Πρόκειται δηλαδή για τη μέση απόσταση μεταξύ των αντικειμένων των συστάδων. Ο μαθηματικός ορισμός της απόστασης Μέσου Όρου δίνεται από την Εξίσωση 11.15

$$d(C_1, C_2) = \frac{\sum_{x_a \in C_1} \sum_{x_b \in C_2} d(x_a, x_b)}{N_{C_1} N_{C_2}}$$

(11.15)

όπου C_1 και C_2 είναι οι δύο συστάδες, $d(x_a,x_b)$ είναι η απόσταση μεταξύ των αντικειμένων x_a και x_b, και N_{C1}, N_{C2} είναι το πλήθος των συστάδων C_1 και C_2 αντίστοιχα.

Η απόσταση μέσου όρου αποτελεί ενδιάμεση λύση ανάμεσα στην ευαισθησία στα αντικείμενα με ακραίες τιμές της μεθόδου πλήρους σύνδεσης, και στην τάση δημιουργίας επιμηκών συστάδων της απλής σύνδεσης. Χάρη στον υπολογισμό της μέσης απόστασης μεταξύ των ζευγών, δεν δημιουργείται το φαινόμενο της αλυσίδας. Επίσης, εξομαλύνεται η επιρροή των αντικειμένων με ακραίες τιμές. Από άποψη υπολογιστικού κόστους

η μέθοδος είναι ακριβή, καθώς υπολογίζει τις αποστάσεις όλων των δυνατών ζευγών. Ένα άλλο μειονέκτημα είναι ότι μπορεί να διασπάσει υπαρκτές επιμήκεις συστάδες.

11.4.6 Απόσταση Μέσων Σημείων (centroids)

Σύμφωνα με την προσέγγιση αυτή, η απόσταση μεταξύ δύο συστάδων είναι η απόσταση ανάμεσα στα μέσα σημεία των δύο συστάδων. Ο μαθηματικός ορισμός της απόστασης μέσων σημείων δίνεται από την Εξίσωση 11.16

$$d(C_1, C_2) = d(m_1, m_2)$$

(11.16)

όπου m_1, m_2 είναι τα μέσα σημεία των συστάδων C_1 και C_2.

Η απόσταση μέσων σημείων έχει το πλεονέκτημα ότι δεν επηρεάζεται σημαντικά από την ύπαρξη αντικειμένων με ακραίες τιμές.

11.4.7 Μέθοδος Ward

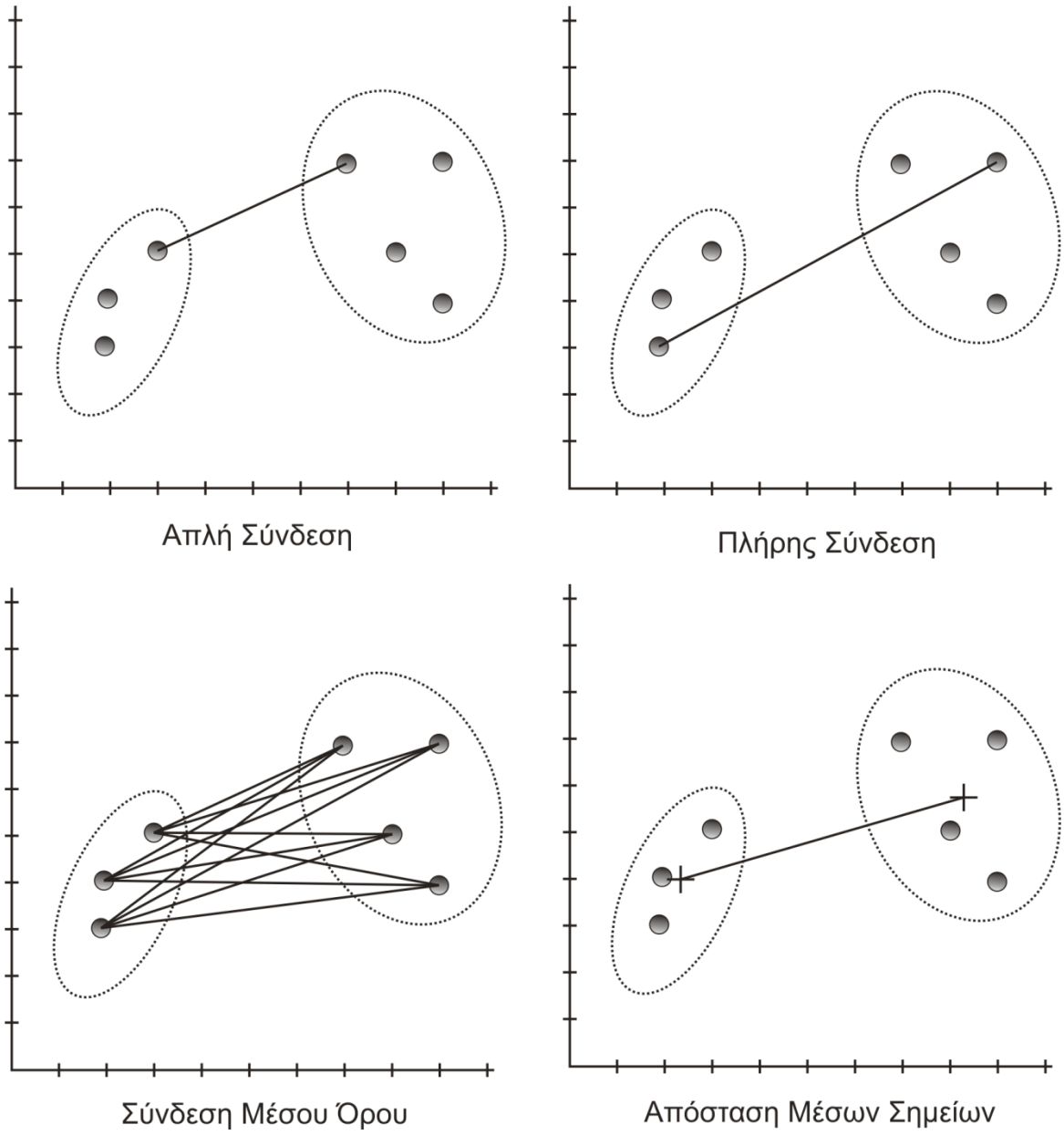
Η μέθοδος του Ward (1963) διαφέρει σημαντικά από τις προηγούμενες μεθόδους, καθώς δεν υπολογίζει κάποια «απόσταση» μεταξύ των συστάδων. Κριτήριο για τη δημιουργία συστάδων είναι η μεγιστοποίηση της ομοιογένειας στο εσωτερικό των συστάδων. Το μέτρο που εφαρμόζεται είναι το άθροισμα του τετραγωνικού σφάλματος, και επιδίωξη της μεθόδου είναι η ελαχιστοποίηση του. Το ίδιο κριτήριο χρησιμοποιείται και από τον αλγόριθμο k-Means, οπότε η μέθοδος Ward μπορεί να θεωρηθεί το ιεραρχικό ανάλογο του k-Means.

Το τετραγωνικό σφάλμα δίνεται από τη Σχέση 11.17

$$E = \sum_{x \in C_i} (x - m_i)^2$$

(11.17)

όπου C_i είναι μια κλάση και mi είναι το μέσο σημείο της. Η μέθοδος, για να συνενώσει δύο συστάδες από συνολικό πλήθος *k* συστάδων, ελέγχει τα δυνατά *k(k-1)/2* ζεύγη συστάδων τα οποία μπορούν να δημιουργηθούν, και επιλέγει το ζεύγος, το οποίο όταν ενωθεί θα μας δώσει τη συστάδα με το ελάχιστο τετραγωνικό σφάλμα. Η μέθοδος του Ward έχει την τάση να παράγει ισοπληθείς ομάδες.



Σχήμα 11.4 *Απόσταση μεταξύ συστάδων*

11.5 Διαχωριστική Ανάλυση Συστάδων

Οι διαχωριστικές μέθοδοι θεωρούν ένα πλήθος *N* σημείων και ένα πλήθος *k* συστάδων, και διαμερίζουν τα σημεία στις συστάδες. Τυπικά, το πλήθος των συστάδων *k* προκαθορίζεται από τον χρήστη. Ξεκινώντας από έναν αρχικό διαχωρισμό, με μια επαναληπτική διαδικασία, τα σημεία μετακινούνται από μια συστάδα σε μιαν άλλη. Ο σχηματισμός των συστάδων γίνεται με τρόπο τέτοιο, ώστε να βελτιστοποιείται ένα κριτήριο διαχωρισμού. Στόχος είναι να δημιουργηθούν συστάδες, οι οποίες να περιέχουν όμοια αντικείμενα, ενώ τα αντικείμενα διαφορετικών συστάδων να είναι ανόμοια.

Οι διαχωριστικές μέθοδοι παρουσιάζουν ευαισθησία στις αρχικές τους συνθήκες. Ένα σημαντικό πρόβλημα είναι το πλήθος των συστάδων *k*. Η εργασία του Dubes (1987) παρέχει καθοδήγηση για τον καθορισμό του πλήθους των συστάδων. Επίσης, για την εύρεση της καθολικά βέλτιστης λύσης θα έπρεπε να δοκιμαστούν όλοι οι δυνατοί διαχωρισμοί. Ωστόσο, λόγω υπολογιστικού κόστους, αυτό δεν είναι εφικτό. Στην πράξη εφαρμόζεται μια διαδικασία αρχικοποίησης του διαχωρισμού, και στη συνέχεια, μετακίνησης των σημείων.

Οι διαχωριστικές μέθοδοι δημιουργούν ένα σύνολο συστάδων, σε αντίθεση με τις ιεραρχικές μεθόδους, οι οποίες δημιουργούν μια ιεραρχική δομή διαδοχικών επιπέδων, όπου κάθε επίπεδο ορίζει ένα σύνολο συστάδων. Επίσης, είναι υπολογιστικά λιγότερο ακριβές από τις ιεραρχικές μεθόδους, και για τον λόγο αυτό

μπορούν να εφαρμοστούν σε μεγαλύτερα σύνολα δεδομένων. Η πιο γνωστή μέθοδος διαχωριστικής ανάλυσης συστάδων είναι ο αλγόριθμός k-Means.

11.5.1 Η μέθοδος k-Means

Η μέθοδος k-Means προτάθηκε από τον MacQueen (1967), και είναι η πιο γνωστή και διαδεδομένη διαιρετική μέθοδος ΑΣ. Στόχος της είναι να κατανείμει ένα σύνολο αντικειμένων σε έναν προκαθορισμένο αριθμό συστάδων, με τρόπο τέτοιο που να αυξάνει την ομοιότητα εντός των συστάδων. Ο αλγόριθμος περιλαμβάνει μια επαναληπτική διαδικασία, όπου σε κάθε επανάληψη υπολογίζεται το κέντρο της συστάδας (centroid). Τα αντικείμενα εντάσσονται στη συστάδα με το πλησιέστερο κέντρο.

Αναλυτικότερα, ο αλγόριθμος της μεθόδου k-Means έχει ως ακολούθως:

- Αρχικά επιλέγονται τυχαία *k* αντικείμενα. Ο αριθμός *k* είναι το πλήθος των συστάδων που θα προκύψουν και προκαθορίζεται από τον χρήστη. Τα επιλεγμένα σημεία θεωρούνται κέντρα συστάδων.
- Κάθε αντικείμενο κατατάσσεται στη συστάδα, της οποίας το κέντρο είναι πλησιέστερα του. Για τον υπολογισμό της απόστασης συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση.
- Τα κέντρα της κάθε συστάδας επαναϋπολογίζονται. Για κάθε διάσταση το κέντρο έχει τιμή ίση με τη μέση τιμή όλων των αντικειμένων, τα οποία ανήκουν στη συστάδα.

$$m_i = \frac{1}{M_i} \sum_{j=1}^{M_i} x_j$$

(11.18)

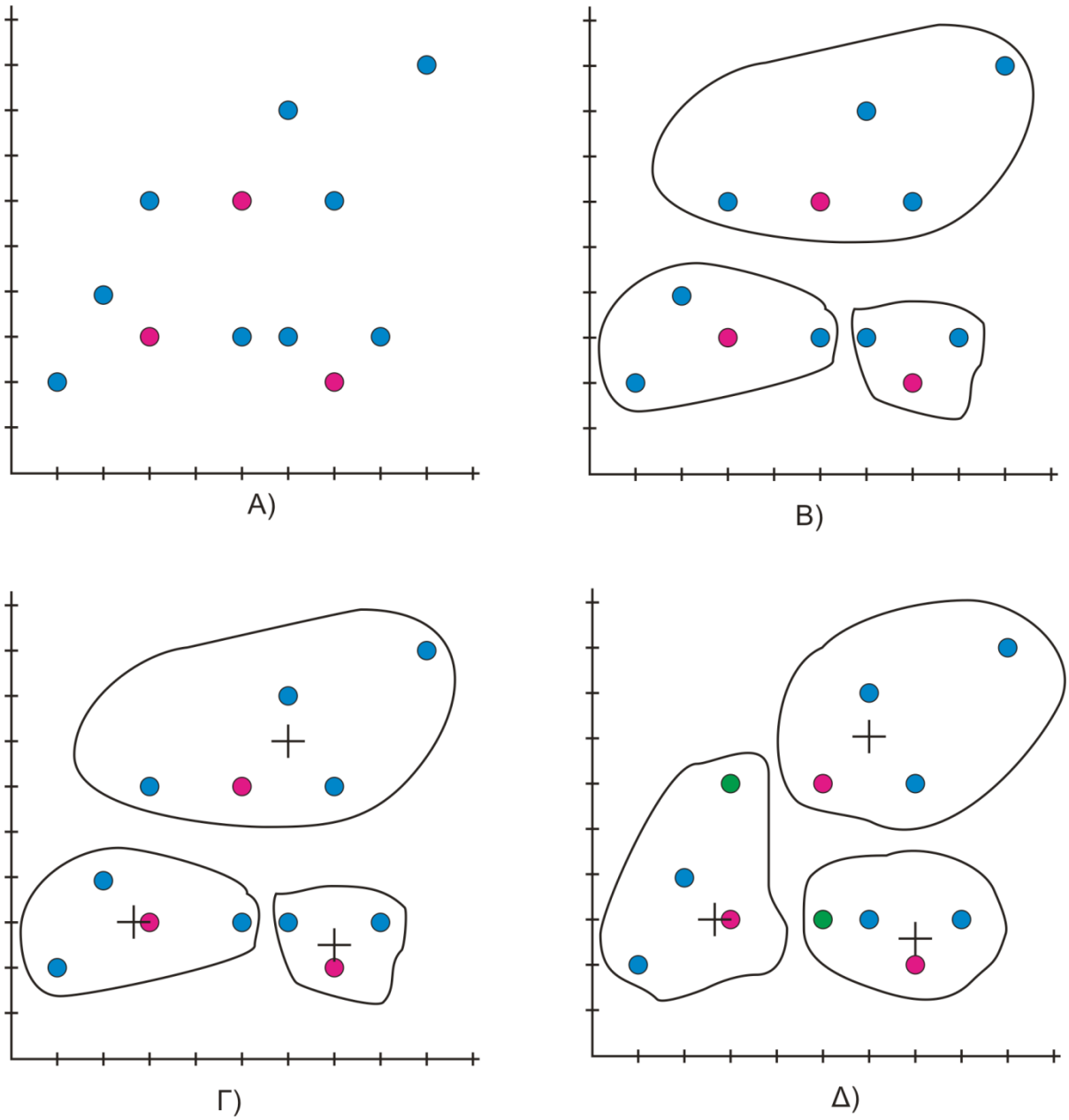
- όπου M_i είναι το πλήθος των αντικειμένων της συστάδας *i*, και m_i είναι το υπολογιζόμενο κέντρο.
- Τα προηγούμενα δύο βήματα επαναλαμβάνονται μέχρι να ικανοποιηθεί η συνθήκη εξόδου. Τυπικά, συνθήκη εξόδου είναι η ελαχιστοποίηση του τετραγωνικού σφάλματος, το οποίο ορίζεται από την Εξίσωση 11.19.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2$$

(11.19)

- όπου C_i είναι οι συστάδες, *x* είναι τα αντικείμενα και m_i είναι το κέντρο της συστάδας C_i.

Στο Σχήμα 11.5 παρουσιάζεται ο σχηματισμός των συστάδων με τη μέθοδο k-Means. Στο τμήμα Α) παρουσιάζονται τα σημεία. Τα κόκκινα σημεία συμβολίζουν τα αρχικώς επιλεγμένα κέντρα. Στο τμήμα Β) σχηματίζονται οι συστάδες. Κάθε σημείο εντάσσεται στη συστάδα, στης οποίας το κέντρο βρίσκεται πλησιέστερα. Στο τμήμα Γ) υπολογίζονται τα νέα κέντρα των υφιστάμενων συστάδων. Τα νέα κέντρα συμβολίζονται με το σχήμα του σταυρού. Στο τμήμα Δ) επαναϋπολογίζεται η απόσταση των σημείων από τα νέα κέντρα, και τα σημεία επανεντάσσονται στις συστάδες. Τα δύο πράσινα σημεία αλλάζουν συστάδα.



Σχήμα 11.5 *Δημιουργία συστάδων με k-Means*

Ο αλγόριθμος k-Means διαθέτει τα παρακάτω **πλεονεκτήματα**:

- Είναι απλός και κατανοητός.
- Τα αντικείμενα μοιράζονται σε συστάδες με αυτόματο τρόπο.
- Είναι αρκετά γρήγορος, τουλάχιστον σε σχέση με τις ιεραρχικές μεθόδους. Ο χρόνος εκτέλεσης του αλγορίθμου εξαρτάται γραμμικά από τα στοιχεία του προβλήματος, όπως το πλήθος των συστάδων *k*, το πλήθος των αντικειμένων n και το πλήθος των επαναλήψεων *l*. Η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι *O(nkl)*. Για τον λόγο αυτό, είναι πιο κατάλληλος από άλλες μεθόδους για την ομαδοποίηση μεγάλων συνόλων αντικειμένων.

Τα βασικά **μειονεκτήματα** του k-Means είναι τα ακόλουθα:

- Ο αριθμός των συστάδων πρέπει να προκαθοριστεί από τον χρήστη.
- Το τελικό αποτέλεσμα εξαρτάται σε σημαντικό βαθμό από την επιλογή των αρχικών κέντρων. Επιλογή διαφορετικών κέντρων μπορεί να οδηγήσει σε σημαντικά διαφορετικές συστάδες.
- Είναι πολύ ευαίσθητος στην ύπαρξη αντικειμένων με ακραίες τιμές (outliers). Λίγα αντικείμενα με πολύ μεγάλες τιμές μπορούν να επηρεάσουν σημαντικά τον υπολογισμό των νέων

κέντρων και κατά συνέπεια τη διαμόρφωση των τελικών συστάδων.
- Έχει την τάση να δημιουργεί σφαιρικές και ίσου μεγέθους συστάδες. Για τον λόγο αυτό, δεν είναι κατάλληλος για συστάδες με περίπλοκα σχήματα ή με πολύ διαφορετικά μεγέθη.

Για την αντιμετώπιση των προβλημάτων του k-Means έχουν προταθεί διάφορες λύσεις. Ένα βασικό πρόβλημα είναι ο προκαθορισμός του αριθμού των συστάδων. Μια δυνατή λύση σε αυτό το πρόβλημα είναι να εφαρμοστεί αρχικά ιεραρχική ΑΣ. Η ιεραρχική ΑΣ συνίσταται σε μια διαδικασία διαδοχικών συνενώσεων ή διασπάσεων των συστάδων. Με τον τρόπο αυτόν, ο χρήστης μπορεί να εκτιμήσει το πλήθος των συστάδων, και στη συνέχεια να εκτελέσει τον k-Means. Ένα άλλο σημαντικό πρόβλημα είναι ότι ο αλγόριθμος μπορεί να συγκλίνει σε τοπικά βέλτιστα, και δεν υπάρχει εγγύηση για την εύρεση ενός καθολικού βέλτιστου. Το τελικό αποτέλεσμα επηρεάζεται σημαντικά από την επιλογή των αρχικών κέντρων. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι οι διαδοχικές, πολλαπλές εκτελέσεις του αλγορίθμου, με διαφορετικά αρχικά κέντρα κάθε φορά. Πρόσθετες τεχνικές επιδιώκουν τη σύγκλιση σε καθολικό βέλτιστο. Οι Likas, Vlassis and Verbeek (2003) εφαρμόζουν μια αιτιοκρατική διαδικασία καθολικής αναζήτησης. Στη διαδικασία αυτή εκτελούνται πολλαπλές τοπικές αναζητήσεις με τον k-Means για διαρκώς αυξανόμενο πλήθος συστάδων, μέχρι το τελικό επιθυμητό πλήθος συστάδων *M*.

11.5.2 Λοιποί αλγόριθμοι Διαιρετικής Ανάλυσης Συστάδων

11.5.2.1 k-Medoids

Όπως αναφέρθηκε και προηγουμένως, ο αλγόριθμος k-Means είναι ευαίσθητος στην ύπαρξη εξαιρέσεων. Ένας τρόπος αντιμετώπισης αυτού του προβλήματος είναι η χρήση ως κέντρου, όχι ενός υπολογιζόμενου μέσου σημείου, αλλά ενός υπαρκτού σημείου δεδομένων. Ο αλγόριθμος k-Medoids ακολουθεί αυτήν την προσέγγιση. Μια από τις πρώτες εκδοχές του k-Medoids ήταν η μέθοδος Partitioning Around Medoids – PAM (Kaufman & Rousseeuw, 1990).

Οι αλγόριθμοι k-Means και k-Medoids παρουσιάζουν αρκετές ομοιότητες:

- Αρχικά επιλέγονται αυθαίρετα τα κέντρα των συστάδων.
- Σε μια επαναληπτική διαδικασία τα κέντρα επαναπροσδιορίζονται.
- Σε κάθε επανάληψη μειώνεται το κριτήριο.
- Επιλογή διαφορετικών αρχικών κέντρων μπορεί να δώσει διαφορετικά αποτελέσματα.
- Δεν επιτυγχάνουν καθολικά βέλτιστα.

Αναλυτικότερα, στον αλγόριθμο k-Medoids επιλέγονται αρχικά k σημεία ως κέντρα (medoids). Τα υπόλοιπα σημεία κατατάσσονται στη συστάδα του πλησιέστερου κέντρου. Μια συνάρτηση κόστους μετρά το άθροισμα των αποστάσεων όλων των σημείων από το κέντρο της συστάδας τους. Σε μια επαναληπτική διαδικασία, σημεία τα οποία δεν είναι κέντρα δοκιμάζονται ως πιθανά κέντρα. Εάν για ένα σημείο το κόστος γίνεται μικρότερο, τότε το σημείο αυτό γίνεται το νέο κέντρο στη θέση του προηγούμενου.

Ο αλγόριθμος k-Medoids λειτουργεί πιο αποτελεσματικά από τον k-Means, όταν στα δεδομένα υπάρχουν αντικείμενα με ακραίες τιμές. Ωστόσο, το κόστος υπολογισμού των medoids είναι σημαντικά μεγαλύτερο από το κόστος υπολογισμού των μέσων τιμών. Για τον λόγο αυτό, ο k-Medoids υπολείπεται του k-Means ως προς τον χρόνο επεξεργασίας μεγάλων συνόλων δεδομένων.

11.5.2.2 CLARA

Ο αλγόριθμος k-Medoids δεν αποδίδει καλά με μεγάλα σύνολα δεδομένων, λόγω υπολογιστικού κόστους. Μια βελτίωση του αλγορίθμου, η οποία αντιμετωπίζει αυτό το πρόβλημα, είναι η μέθοδος CLARA (Clustering LARge Applcations) (Kaufman & Rousseeuw, 1990). Η μέθοδος CLARA δεν χρησιμοποιεί ολόκληρο το σύνολο δεδομένων. Αντιθέτως, εκτελεί τυχαία δειγματοληψία και επιλέγει ένα υποσύνολο του. Το υποσύνολο δεδομένων υπόκειται σε ανάλυση συστάδων, σύμφωνα με τη μέθοδο PAM. Λόγω της τυχαίας δειγματοληψίας, είναι αρκετά πιθανό, ότι τα medoids που θα υπολογιστούν, θα είναι όμοια με αυτά που θα προέκυπταν από την επεξεργασία ολόκληρου του συνόλου δεδομένων. Ο αλγόριθμος επιλέγει πολλά υποσύνολα δεδομένων και επιστρέφει το καλύτερο αποτέλεσμα.

11.6 Αυτοοργανούμενοι Χάρτες

Οι **Αυτοοργανούμενοι Χάρτες (ΑΟΧ)** (Self Organizing Maps (SOMs)) είναι ένας τύπος Νευρωνικού Δικτύου, ο οποίος προτάθηκε από τον Φιλανδό καθηγητή Kohonen (1982). Λόγω των ιδιαίτερων χαρακτηριστικών και δυνατοτήτων τους, οι Αυτοοργανούμενοι Χάρτες προσέλκυσαν το ενδιαφέρον του επιστημονικού κόσμου, και πλήθος βιβλίων και ερευνητικών εργασιών αναφέρθηκαν σε αυτούς. Ωστόσο, το βιβλίο το οποίο θεωρείται σημείο αναφοράς είναι το Kohonen (2001).

Ο τρόπος εκπαίδευσης των ΑΟΧ είναι μη επιβλεπόμενος. Στη μη-επιβλεπόμενη μάθηση, οι απαντήσεις δεν είναι γνωστές εκ των προτέρων. Οι αλγόριθμοι χρησιμοποιούν για την εκπαίδευση μόνο τα δεδομένα εισόδου, και όχι τις απαντήσεις του δικτύου. Στην περίπτωση των ΑΟΧ, εκτελείται μια επαναληπτική διαδικασία, όπου το μοντέλο τροφοδοτείται με παραδείγματα εκπαίδευσης. Οι νευρώνες του δικτύου προσαρμόζονται, έτσι ώστε να «μοιάζουν» με τα δεδομένα εκπαίδευσης. Ένα πολύ σημαντικό χαρακτηριστικό είναι ότι παρεμφερή παραδείγματα αντιστοιχίζονται σε περιοχές γειτονικών νευρώνων. Με τον τρόπο αυτό, διατηρούνται οι τοπολογικές σχέσεις των δεδομένων εισόδου. Οι ΑΟΧ παρουσιάζουν σημαντικές αναλογίες με τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Ο εγκέφαλος είναι χωρικά οργανωμένος, και συγκεκριμένα τμήματα του είναι υπεύθυνα για συγκεκριμένες εργασίες, όπως πχ τη μνήμη, την ομιλία κλπ. Με αντίστοιχο τρόπο, οι ΑΟΧ αντιστοιχούν παρεμφερείς έννοιες σε γειτονικές περιοχές. Για τον λόγο αυτό, οι ΑΟΧ θεωρούνται ένα από τα πιο ρεαλιστικά μοντέλα του ανθρώπινου εγκεφάλου.

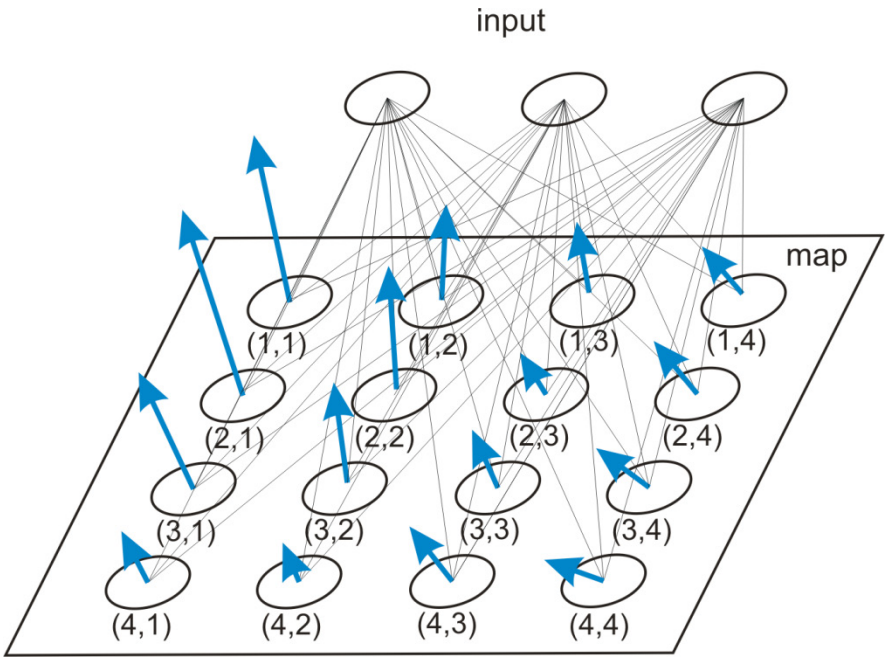
Οι ΑΟΧ παρέχουν έναν τρόπο απεικόνισης πολυδιάστατων δεδομένων, σε έναν χώρο πολύ λιγότερων, τυπικά μίας ή δύο, διαστάσεων. Με τον τρόπο αυτό, οπτικοποιούν σύνθετα δεδομένα, και τα παρουσιάζουν με τρόπο κατανοητό στους ανθρώπους. Ο Kohonen περιγράφει τους ΑΟΧ σαν ένα εργαλείο οπτικοποίησης και ανάλυσης πολυδιάστατων δεδομένων. Ωστόσο, χάρη στα ιδιαίτερα χαρακτηριστικά τους, οι ΑΟΧ μπορούν να χρησιμοποιηθούν για διάφορες εργασίες, όπως για την ανάλυση συστάδων, τη μείωση των διαστάσεων, ακόμα και για κατηγοριοποίηση. Μπορεί κανείς να φανταστεί τους ΑΟΧ σαν μια ελαστική επιφάνεια που διακυμαίνεται, έτσι ώστε να είναι όσο το δυνατό πλησιέστερα στα πρότυπα εκπαίδευσης,

11.6.1 Δομή ΑΟΧ

Ένας ΑΟΧ είναι ένα νευρωνικό δίκτυο ενός επιπέδου και αποτελείται από νευρώνες. Οι νευρώνες είναι διατεταγμένοι σε ένα πλέγμα *n* διαστάσεων. Κατά κανόνα, το πλέγμα είναι 2 διαστάσεων και ορθογώνιο ή εξαγωνικό.

Στο Σχήμα 11.6 απεικονίζεται ένας Αυτοοργανούμενος Χάρτης, ο οποίος διαθέτει ορθογώνιο πλέγμα 2 διαστάσεων. Το δίκτυο αποτελείται από 16 νευρώνες σε διάταξη *4X4*. Επίσης, υπάρχουν 3 νευρώνες εισόδου. Αυτό σημαίνει ότι τα πρότυπα εισόδου έχουν τρεις διαστάσεις, και ότι γίνεται προβολή ενός χώρου 3 διαστάσεων σε έναν χώρο 2 διαστάσεων. Κάθε νευρώνας εισόδου είναι συνδεδεμένος με κάθε νευρώνα του χάρτη. Με τον τρόπο αυτό, οι τιμές των προτύπων εισόδου διαβιβάζονται σε όλους τους νευρώνες του δικτύου. Ένα σημαντικό στοιχείο είναι ότι οι νευρώνες του χάρτη δεν είναι συνδεδεμένοι μεταξύ τους. Επίσης, κάθε νευρώνας του χάρτη διαθέτει δύο χωρικές συντεταγμένες *(i,j)*, οι οποίες καθορίζουν τη θέση του. Οι συντεταγμένες χρησιμεύουν για τον προσδιορισμό του εκάστοτε νευρώνα, καθώς και για τον υπολογισμό των αποστάσεων μεταξύ των νευρώνων.

Τα δεδομένα, με τα οποία θα εκπαιδευτεί και θα χρησιμοποιηθεί το δίκτυο, έχουν N διαστάσεις, είναι δηλαδή της μορφής $x=(x_1, x_2,...,x_N)$. Κάθε πρότυπο εισόδου μπορεί να θεωρηθεί ως ένα διάνυσμα N διαστάσεων. Επίσης, κάθε νευρώνας του δικτύου w περιέχει ένα σύνολο αριθμητικών τιμών $(w_1, w_2,...w_N)$, οι οποίες καλούνται και βάρη του νευρώνα. Το πλήθος των βαρών είναι ίσο με *N*, όσο δηλαδή είναι και το πλήθος των διαστάσεων των διανυσμάτων εισόδου. Επομένως, ο νευρώνας μπορεί να θεωρηθεί ως ένα διάνυσμα στον ίδιο χώρο με τα διανύσματα εισόδου. Τα διανύσματα των νευρώνων συμβολίζονται στο Σχήμα 11.6 με μπλε βέλη.



Σχήμα 11.6 *Αυτοοργανούμενος Χάρτης*

11.6.2 Εκπαίδευση ΑΟΧ

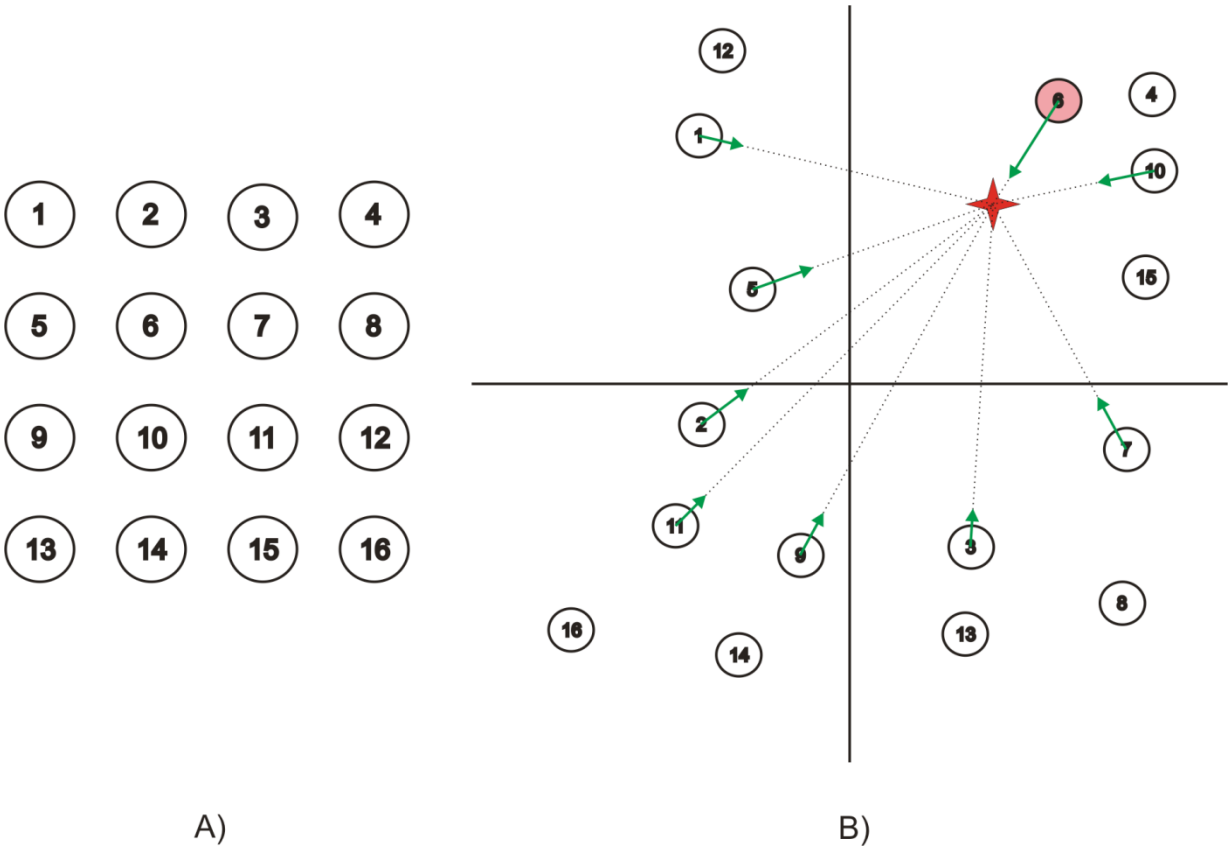
Η εκπαίδευση ενός ΑΟΧ συνίσταται στη ρύθμιση των βαρών των νευρώνων. Εκτελείται μια επαναλαμβανόμενη διαδικασία, όπου σε κάθε επανάληψη το δίκτυο τροφοδοτείται με ένα διάνυσμα εισόδου. Στη συνέχεια, υπολογίζεται η απόσταση του διανύσματος εισόδου από το διάνυσμα του κάθε νευρώνα. Ο νευρώνας, ο οποίος βρίσκεται πλησιέστερα στο διάνυσμα εισόδου, θεωρείται «νικητής», και είναι αυτός που αντιπροσωπεύει το διάνυσμα εισόδου. Ο νικητής νευρώνας ονομάζεται και Best Matching Unit (BMU). Τα βάρη του BMU τροποποιούνται κατάλληλα, έτσι ώστε το διάνυσμα του να μετακινηθεί προς το διάνυσμα εισόδου, και να βελτιωθεί η αντιπροσώπευση. Το μέγεθος της μεταβολής των βαρών εξαρτάται από μια παράμετρο, που ονομάζεται ρυθμός εκπαίδευσης.

Σκοπός των ΑΟΧ είναι η προβολή παρεμφερών προτύπων πολλών διαστάσεων σε μια ορισμένη περιοχή, ενός χώρου λιγότερων διαστάσεων. Για την επίτευξη αυτού του σκοπού, κατά τη διάρκεια της εκπαίδευσης, δεν ρυθμίζονται μόνο τα βάρη του νευρώνα νικητή, αλλά και τα βάρη των τοπογραφικά γειτονικών του νευρώνων. Τα διανύσματα των γειτονικών νευρώνων μετακινούνται και αυτά προς το διάνυσμα εισόδου. Η διαδικασία παρουσίασης ενός διανύσματος εισόδου, και ρύθμισης των βαρών του νικητή νευρώνα και των γειτονικών του νευρώνων επαναλαμβάνεται πολλές φορές. Το πλήθος των επαναλήψεων εξαρτάται από τον αριθμό των εποχών. Μια εποχή είναι η επεξεργασία όλων των διανυσμάτων εισόδου μια φορά. Με τον τρόπο αυτό, κάθε διάνυσμα εισόδου χρησιμοποιείται για εκπαίδευση τόσες φορές όσες είναι το πλήθος των εποχών.

Στο Σχήμα 11.7 παρουσιάζεται σχηματικά η ρύθμιση των βαρών των νευρώνων με τη χρήση ενός διανύσματος εισόδου. Στο αριστερό μέρος του σχήματος (τμήμα Α) απεικονίζεται η τοπολογική διάταξη των νευρώνων στο πλέγμα. Το δίκτυο διαθέτει 16 νευρώνες, σε επίπεδη διάταξη *4X4*. Για καλύτερη οπτική απεικόνιση, θεωρούμε διανύσματα εισόδου 2 διαστάσεων. Κατά συνέπεια, και οι νευρώνες του δικτύου θα είναι δύο διαστάσεων. Ανάλογα με τις τιμές των βαρών τους, οι νευρώνες απεικονίζονται ως σημεία σε έναν δισδιάστατο χώρο, στο δεξιό μέρος του σχήματος (τμήμα Β). Στο δίκτυο παρουσιάζεται ένα διάνυσμα εισόδου, το οποίο συμβολίζεται με το σχήμα του κόκκινου άστρου. Ο πλησιέστερος, από άποψη βαρών, νευρώνας, είναι ο νευρώνας 6. Ο νευρώνας αυτός είναι ο νικητής. Ο νικητής νευρώνας μετακινείται προς το διάνυσμα εισόδου, ώστε να βελτιωθεί η αντιπροσώπευση. Οι πλησιέστεροι τοπολογικά νευρώνες στον νευρώνα 6 είναι οι 1, 2, 3, 5, 7, 9, 10 και 11 (τμήμα Α). Οι νευρώνες αυτοί μετακινούνται επίσης προς το διάνυσμα εισόδου. Η μετακίνηση των νευρώνων γίνεται κατά μήκος των διακεκομμένων γραμμών. Η μεταβολή των βαρών και η αντίστοιχη μετακίνηση συμβολίζεται με τα πράσινα βέλη. Το μέγεθος του βέλους υποδηλώνει το ποσοστό της μεταβολής. Όσο τοπολογικά μακρύτερα βρίσκεται ο εκάστοτε νευρώνας από τον BMU, τόσο λιγότερο μετακινείται.

Πριν την έναρξη της εκπαίδευσης, τα βάρη των νευρώνων αρχικοποιούνται με τυχαίες τιμές. Κατά τη διάρκεια της εκπαίδευσης, το μέγεθος της γειτονιάς του BMU και ο ρυθμός εκπαίδευσης δεν παραμένουν στα-

θεροί, αλλά μειώνονται με τον αριθμό των εποχών. Ο καθορισμός του ρυθμού εκπαίδευσης και της ακτίνας της γειτονιάς αποτέλεσε αντικείμενο μελέτης πολλών ερευνητών. Σύμφωνα με τον Kohonen, η εκπαίδευση χωρίζεται σε δύο στάδια. Κατά το πρώτο στάδιο, το οποίο είναι γνωστό και ως unfolding phase, ο ρυθμός εκπαίδευσης μειώνεται από 0,9 σε 0,1. Η ακτίνα της γειτονιάς αρχικά ισούται με το ήμισυ της διαμέτρου του πλέγματος, και σταδιακά περιορίζεται, με τρόπο που να περιλαμβάνει στο τέλος τους άμεσους γείτονες του BMU. Κατά το δεύτερο στάδιο, γνωστό και ως fine tuning phase, ο ρυθμός εκπαίδευσης μειώνεται σταδιακά από την τιμή 0,1 στην τιμή 0,0, ενώ η ακτίνα της γειτονιάς διατηρεί σταθερή τιμή ίση με 1 και περιλαμβάνει μόνο τον νευρώνα BMU. Πρακτικά, κατά το πρώτο στάδιο καθορίζεται το γενικό σχήμα του δικτύου, και στο δεύτερο στάδιο προσαρμόζονται καλύτερα οι νευρώνες στα διανύσματα εισόδου.



Σχήμα 11.7 *Εκπαίδευση ΑΟΧ*

Μετά την ολοκλήρωση της εκπαίδευσης, κάθε διάνυσμα εισόδου θα αντιστοιχείται σε έναν νευρώνα, δεν θα ταυτίζεται όμως με αυτόν, δηλαδή μεταξύ τους θα υπάρχει μια διαφορά. Η διαφορά αυτή καλείται **σφάλμα κβάντωσης** (quantization error). Αθροίζοντας τη διαφορά κάθε διανύσματος εισόδου με τον νευρώνα στον οποίο αντιστοιχίζεται, λαμβάνουμε το συνολικό σφάλμα κβάντωσης. Το συνολικό σφάλμα κβάντωσης αποτελεί μέτρο της ποιότητας της αντιπροσώπευσης των διανυσμάτων εισόδου από το εκπαιδευμένο δίκτυο.

Ένα άλλο ζήτημα, το οποίο αφορά την ποιότητα του εκπαιδευμένου δικτύου, είναι η διατήρηση της τοπολογίας του χώρου εισόδου, δηλαδή η διατήρηση των σχέσεων γειτονίας που υφίστανται στον χώρο εισόδου. Για τη μέτρηση αυτού του ποιοτικού στοιχείου, ο Kohonen (2001) προτείνει ένα **τοπογραφικό σφάλμα**. Το σφάλμα αυτό μετρά το ποσοστό των διανυσμάτων εισόδου, για τα οποία ο πρώτος και ο δεύτερος νευρώνας αντιπροσώπευσης δεν είναι γειτονικοί. Πρόσθετα μέτρα για την εκτίμηση της διατήρησης της τοπολογίας έχουν προταθεί από τους Bauer and Pawelzik (1992), τους Bezdek and Pal (1995) και τους Uriarte and Martin (2005).

Το τελικό αποτέλεσμα της εκπαίδευσης ενός ΑΟΧ είναι η ρύθμιση των βαρών των νευρώνων. Ωστόσο, ο πίνακας των βαρών δεν θεωρείται ως ο πλέον κατάλληλος, για την οπτική αναπαράσταση και την αναγνώριση των συστάδων. Διάφορες τεχνικές έχουν προταθεί για την καλύτερη οπτική αναπαράσταση του χάρτη. Μια πολύ διαδεδομένη τεχνική είναι η U-Matrix (Ultsch & Siemon, 1990). Σύμφωνα με την τεχνική U-Matrix, υπολογίζονται οι αποστάσεις μεταξύ των διανυσμάτων γειτονικών νευρώνων. Οι αποστάσεις αυτές είναι μια εκτίμηση του βαθμού ανομοιότητας ομάδων των διανυσμάτων εισόδου. Αν οι αποστάσεις των διανυσμάτων μιας ομάδας γειτονικών νευρώνων είναι μικρές, τότε αυτό περιγράφει μια συστάδα δεδομένων εισόδου. Για

την καλύτερη οπτική αναπαράσταση της πληροφορίας, η διαφορά των αποστάσεων συμβολίζεται με διαβαθμίσεις του χρώματος γκρι. Ανοιχτόχρωμες περιοχές υποδηλώνουν την ύπαρξη συστάδας, ενώ σκουρόχρωμες περιοχές θεωρούνται ότι διαχωρίζουν συστάδες. Ορισμένες υλοποιήσεις του αλγορίθμου σε πακέτα λογισμικού χρησιμοποιούν πολλαπλά χρώματα αντί για διαβαθμίσεις του γκρι. Άλλη μέθοδος για την οπτικοποίηση των ΑΟΧ είναι η προβολή Sammon (Sammon, 1969).

11.7 Επιχειρηματικές Εφαρμογές της Ανάλυσης Συστάδων

Η Ανάλυση Συστάδων έχει βρει σημαντικές εφαρμογές στις σύγχρονες επιχειρήσεις. Το πιο γνωστό πεδίο εφαρμογής είναι το **μάρκετινγκ**. Οι διαφημιστικές εκστρατείες, οι οποίες απευθύνονται στο σύνολο του πληθυσμού, είναι ακριβές και έχουν μικρό ποσοστό ανταπόκρισης. Είναι προφανές ότι κάθε προσφορά δεν είναι χρήσιμη για κάθε πελάτη, και κάθε πελάτης δεν ανταποκρίνεται με τον ίδιο τρόπο στο ίδιο διαφημιστικό μήνυμα. Επιθυμία των διαφημιστών είναι να επιμερίσουν τον πληθυσμό σε ομάδες με ομοειδή χαρακτηριστικά. Οι επιχειρήσεις διατηρούν στα μηχανογραφικά τους συστήματα πληροφορίες για τους πελάτες τους. Τα στοιχεία αυτά μπορούν να αναλυθούν, ώστε να εξαχθούν συμπεράσματα χρήσιμα για διαφημιστικούς σκοπούς. Ένας νέος όρος, που περιγράφει αυτήν την πρακτική, είναι ο όρος «data base marketing». Εάν η επιχείρηση δεν διαθέτει στοιχεία, τότε μπορεί να διεξάγει μια έρευνα σε ένα αντιπροσωπευτικό δείγμα του πληθυσμού. Οι πελάτες μπορούν να ομαδοποιηθούν σύμφωνα με διάφορα κριτήρια, όπως γεωγραφικά (πόλη, περιοχή, χώρα κλπ.), ψυχογραφικά (τρόπος ζωής, προσωπικότητα, ηθικές αξίες), δημογραφικά (φύλο, ηλικία κλπ.) και καταναλωτικού προφίλ (προηγούμενες αγορές, τρόπος χρήσης του προϊόντος). Ο επιμερισμός του καταναλωτικού κοινού σε ομογενείς ομάδες είναι γνωστός με τον όρο «**τμηματοποίηση αγοράς**» (market segmentation). Οι διαφημιστές, έχοντας γνώση των τμημάτων που απαρτίζουν την αγορά, μπορούν να οργανώσουν στοχευμένες διαφημιστικές εκστρατείες, εξειδικευμένες για κάθε τμήμα πελατών. Αυτές οι διαφημιστικές εκστρατείες έχουν χαμηλότερο κόστος και καλύτερο ποσοστό ανταπόκρισης. Ο ακριβής και ουσιαστικός καθορισμός της ομάδας αυξάνει περαιτέρω τον βαθμό ανταπόκρισης. Η τμηματοποίηση της αγοράς και η βαθύτερη γνώση του καταναλωτικού προφίλ των υπαρκτών και υποψήφιων πελατών δεν αφορά μόνο τη διαφήμιση. Η εξυπηρέτηση των πελατών, μετά την πώληση, μπορεί να βελτιωθεί, με την ομαδοποίηση των απόψεων των πελατών, τα παράπονα τους, τις αναφορές για τεχνικά προβλήματα και βλάβες κλπ.

Το μάρκετινγκ είναι το πιο γνωστό παράδειγμα εφαρμογής της ΑΣ, δεν είναι όμως το μοναδικό. Κάθε τομέας της επιχειρηματικής δράσης μπορεί να ορίσει πρόσωπα, αντικείμενα ή ενέργειες σε σχέση με γνωρίσματα, και να ωφεληθεί, ανακαλύπτοντας συστάδες. Στη **διαχείριση ανθρωπίνων πόρων**, οι εργαζόμενοι ομαδοποιούνται σύμφωνα με τις αξιολογήσεις τους, την επαγγελματική τους εκπαίδευση, τη συμμετοχή τους σε ομάδες εργασίες, τις ειδικές δεξιότητες τους κλπ. Τα στοιχεία αυτά χρησιμοποιούνται για τον καθορισμό της μισθολογικής πολιτικής, τις προαγωγές των εργαζομένων, τη στελέχωση ομάδων εργασίας κλπ.

Εφαρμογή της ΑΣ γίνεται και στην **παραγωγή**. Κάθε προϊόν ανήκει σε κάποια κατηγορία προϊόντων. Η ένταξη προϊόντων σε κατηγορίες είναι μια παγιωμένη τακτική, ωστόσο με την πάροδο του χρόνου παρουσιάζονται προβλήματα, όπως απαρχαιωμένες κατηγορίες, ένταξη προϊόντων σε λάθος κατηγορίες κλπ. Με τη χρήση της ΑΣ, μπορούν να επανακαθοριστούν οι κατηγορίες με ουσιαστικότερο τρόπο και να επανενταχθούν προϊόντα στην κατάλληλη κατηγορία. Επίσης, η γνώση των ειδικών χαρακτηριστικών τμημάτων του καταναλωτικού κοινού αξιοποιείται και στην παραγωγή, με τον σχεδιασμό εξειδικευμένων προϊόντων για συγκεκριμένες ομάδες, και με τη δημιουργία αντίστοιχων γραμμών παραγωγής. Τέλος, ολόκληρες γραμμές παραγωγής και εργοστάσια μπορούν να ομαδοποιηθούν σύμφωνα με την ταχύτητα, την ποιότητα και το κόστος.

Η ΑΣ συστάδων μπορεί να χρησιμοποιηθεί για έλεγχο και **εντοπισμό απάτης**. Οι περιπτώσεις απάτης είναι λίγες και έχουν ιδιαίτερα χαρακτηριστικά. Με την εφαρμογή της ΑΣ μπορούν να εντοπιστούν μικρές και απομονωμένες συστάδες, οι οποίες είναι ύποπτες για απάτη. Παράδειγμα τέτοιας ανάλυσης είναι η εργασία των Thiprungsri and Vasarhelyi (2011). Οι ερευνητές μελετούν το πρόβλημα των ομαδικών ασφαλίσεων ζωής. Στις ομαδικές ασφαλίσεις, ο πελάτης είναι μια εταιρεία, η οποία ασφαλίζει το προσωπικό της, και όχι ένας μεμονωμένος ιδιώτης. Οι ασφαλιστικές εταιρείες, που παρέχουν τέτοια ασφαλιστικά συμβόλαια, δεν τηρούν πληροφορίες για τα ασφαλισμένα άτομα. Σε αυτόν τον τύπο ασφάλειας έχουν εντοπιστεί περιπτώσεις απάτης. Εφαρμόζοντας τη μέθοδο k-Means, οι ερευνητές εντόπισαν ολιγομελείς συστάδες, οι οποίες χρήζουν ειδικού περαιτέρω ελέγχου.

Η ΑΣ βρίσκει εφαρμογή στη **διαχείριση των Επιχειρηματικών Διαδικασιών**. Οι σύγχρονες επιχειρήσεις οργανώνουν τη λειτουργία τους με τον καθορισμό επιχειρηματικών διαδικασιών. Πρόσφατα, οι επιχειρηματικές διαδικασίες τυποποιούνται, και αυτοματοποιούνται σε εξειδικευμένα πληροφοριακά συστήματα. Οι Jung, Bae and Liu (2009) μετασχηματίζουν τις επιχειρηματικές διαδικασίες σε διανυσματικά μοντέλα, και τις ομα-

δοποιούν, εφαρμόζοντας Συσσωρευτική Ανάλυση Συστάδων. Τα αποτελέσματα μπορούν να χρησιμοποιηθούν για τον καθορισμό νέων ή τον ανασχεδιασμό των υπαρχουσών επιχειρηματικών διαδικασιών.

Η ΑΣ αξιοποιείται και στο **στρατηγικό μάνατζμεντ**. Τα διοικητικά στελέχη αναλύουν στοιχεία του επιχειρηματικού κλάδου, και ομαδοποιούν τις επιχειρήσεις ανάλογα με τα ιδιαίτερα χαρακτηριστικά τους, τα πλεονεκτήματα τους και τις αδυναμίες τους. Με τον τρόπο αυτό, κατανοούν καλύτερα τις συνθήκες του ανταγωνισμού, και μπορούν να σχεδιάσουν δράσεις, που θα τους εξασφαλίσουν το ανταγωνιστικό πλεονέκτημα. Ένα παράδειγμα ανάλυσης στοιχείων επιχειρηματικού κλάδου είναι η εργασία των Mosleh, Nosratabadi and Bahrami (2015), οι οποίοι αναλύουν τα επιχειρηματικά μοντέλα που εφαρμόζονται στα πρακτορεία τουρισμού του Ιράν. Εφαρμόζοντας ιεραρχική ΑΣ, εντοπίζουν ότι εφαρμόζονται τρία διαφορετικά επιχειρηματικά μοντέλα, με δημοφιλέστερο το μοντέλο που βασίζεται στα χρηματοοικονομικά, και ακολουθούμενο από το μοντέλο που βασίζεται στους πελάτες, και το μοντέλο που βασίζεται στις υπηρεσίες.

Πολύ ενδιαφέροντα πεδία εφαρμογής βρίσκει η ΑΣ σε ζητήματα που αφορούν την **ανάλυση χωρικής** πληροφορίας. Για παράδειγμα, μπορούν να ομαδοποιηθούν οι κατοικίες ανάλογα με τη γεωγραφική τους θέση, τον τύπο τους και την αξία τους, και να καθοριστούν εξειδικευμένες υπηρεσίες για συγκεκριμένες ομάδες. Επίσης, μπορούν να σχεδιαστούν θεματικοί χάρτες, οι οποίοι αναγνωρίζουν περιοχές με παρόμοια χρήση γης, όπως αγροτικές περιοχές, αστικές περιοχές, βιομηχανικές περιοχές κλπ. Τα στοιχεία αυτά είναι χρήσιμα σε κρατικούς φορείς για την άσκηση πολιτικής και τη δημιουργία υποδομών. Ένα πολύ σημαντικό σχετικό αντικείμενο είναι ο εντοπισμός **συστάδων επιχειρήσεων**. Μια συστάδα επιχειρήσεων είναι ένα σύνολο επιχειρήσεων, οι οποίες βρίσκονται στον ίδιο γεωγραφικό χώρο, και επιπλέον ενώνονται στη βάση κοινών στοιχείων, και λειτουργούν ανταγωνιστικά ή/και συμπληρωματικά. Η πιο γνωστή περίπτωση επιχειρηματικής συστάδας είναι η διαβόητη Silicon Valley. Η γεωγραφική συγκέντρωση παρόμοιων και αλληλοσχετιζόμενων επιχειρήσεων μπορεί να εξασφαλίσει μια σειρά από οφέλη όπως:

- Μείωση εξόδων μεταφοράς.
- Ανάπτυξη εξειδικευμένων υποδομών, χρήσιμων για τον κλάδο.
- Επίτευξη οικονομία κλίμακας.
- Αυξημένη έκθεση στον ανταγωνισμό, η οποία ενισχύει τη βελτίωση της ποιότητας.
- Κοινή αξιοποίηση του τοπικού εργατικού δυναμικού, το οποίο σταδιακά αποκτά εξειδικευμένες γνώσεις και δεξιότητες.
- Διάχυση βέλτιστων πρακτικών και νέων τεχνολογιών.
- Επίτευξη ανταγωνιστικού πλεονεκτήματος.

Ο εντοπισμός επιχειρηματικών συστάδων, ακόμα και εν τη γενέσει τους, είναι ιδιαίτερα χρήσιμος, καθώς οι αρμόδιοι φορείς μπορούν να λάβουν ειδικά μέτρα για την ενίσχυση του φαινομένου. Τέτοια μέτρα περιλαμβάνουν την ενίσχυση των υποδομών, την εκπαίδευση του πληθυσμού, την εξασφάλιση επιδοτήσεων, την προσαρμογή της φορολογικής πολιτικής, τη θέσπιση απαγορευτικών διατάξεων κλπ.

Όπως καταδείχτηκε ανωτέρω, η ΑΣ βρίσκει πεδία εφαρμογής σε πλήθος επιχειρηματικών δραστηριοτήτων και ζητημάτων. Σε κάθε περίπτωση, οι εμπλεκόμενοι αναλυτές και τα διοικητικά στελέχη οφείλουν να γνωρίζουν ότι στην ΑΣ συχνά απαιτείται πειραματισμός με διάφορα γνωρίσματα και μεθόδους, μέχρι να εξαχθούν χρήσιμα αποτελέσματα. Η ΑΣ έχει αρκετά έντονα περιγραφικό και διερευνητικό χαρακτήρα. Τα αποτελέσματα συνδυάζονται με τη γνώμη στελεχών, τα οποία διαθέτουν εξειδικευμένη γνώση στο συγκεκριμένο πεδίο. Τα στελέχη αυτά αξιολογούν τα αποτελέσματα των αλγορίθμων και αποφαίνονται εάν είναι λογικά, και εάν προσφέρουν χρήσιμη πληροφορία.

Βιβλιογραφία/Αναφορές

Aggarwal, C., & Reddy, C. (Eds.). (2014). *Data Clustering: Algorithms and Applications*. Hoboken: CRC Press.

Bauer, H. U., & Pawelzik, K. R. (1992). Quantifying the Neighborhood Preservation of Self-Orrganizing Maps. *IEEE Transactions on Neural Networks, 3*(4), 570-579. doi: 10.1109/72.143371

Bezdek, J. C., & Pal, N. R. (1995). An Index of Topological Preservation for Feature Extraction. *Pattern Recognition, 28*(3), 381-391. doi: 10.1016/0031-3203(94)00111-x

Dubes, R. C. (1987). How many Clusters are Best: An Experiment. *Pattern Recognition, 20*(6), 645-663. doi: 10.1016/0031-3203(87)90034-3

Estivill-Castro, V., & Yang, J.A. (2000). Fast and Robust General Purpose Clustering Algorithm. *Proceedings of the Pacific Rim International Conference on Artificial Intelligence,* 208-218. doi: 10.1007/3-540-44533-1_24

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques*. Waltham, MA: Morgan Kaufmann Publishers.

Jung, J. Y., Bae, J., & Liu, L. (2009). Hierarchical Clustering of Business Process Models. *International Journal of Innovative Computing, Information and Control, 5*(12), 1-10.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: John Wiley and Sons.

King, B. (1967). Step-Wise Clustering Procedures. *Journal of the American Statistical Association, 62*(317), 86-101. doi: 10.2307/2282912

Kohonen, T. (1982). Self Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics, 43*(1), 59-69. doi: 10.1007/bf00337288

Kohonen, T. (2001). *Self-Organizing Maps*. Berlin: Springer-Verlag.

Kruskal, J. (1977). The Relationship between Multidimensional Scaling and Clustering. In J. Van Ryzin (Ed.), *Classification and Clustering* (pp 17-45). New York, NY: Academic Press Inc.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The Global k-Means Clustering Algorithm. *Pattern Recognition, 36*(2), 451-461. doi: 10.1016/S0031-3203(02)00060-2

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* 281-297. Berkeley, CA: University of California Press.

Mao, J., & Jain, A. K. (1996). A self-organizing network for hyper ellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks, 7*(1), 16–29. doi: 10.1109/72.478389

Mosleh, A., Nosratabadi, S., & Bahrami, P. (2015). Recognizing the Business Models types in Tourism Agencies: Utilizing the Cluster Analysis. *International Business Research, 8*(2), 173-180. doi: 10.5539/ibr.v8n2p173

Murtagh, F. A. (1984). A Survey of Recent Advances in Hierarchical Clustering Algorithms Which Use Cluster Centers. *The Computer Journal, 26*(4), 354-359. doi: 10.1093/comjnl/26.4.354

Ng, R. T., & Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. *Proceedings of the 20th International conference on Very Large Data Bases,* 144-155. San Francisco, CA: Morgan Kaufmann Publishers.

Sammon, J. (1969). A Nonlinear Mapping for data Structure Analysis. *IEEE Transactions on Computers, C-18*(5), 401-409. doi: 10.1109/t-c.1969.222678

Shekhar, S., & Chawla, S. W. (2003). *Spatial Databases: A Tour*. Upper Saddle River, NJ: Prentice Hall.

Sneath, P., & Sokal, R. (1973*). Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco, CA: WH Freeman Co.

Thiprungsri, S., & Vasarhelyi, M. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *The International Journal of Digital Accounting Research, 11*, 69-84. doi: 10.4192/1577-8517-v11_4

Ultsch, A., & Siemon, H. P. (1990). Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. *Proceedings of the ICNN'90 International Neural Network Conference,* 305-308.

Uriarte, A., & Martin, D. (2005). Topology Preservation in SOM. *International Journal of Applied Mathematics and Computer Sciences, 1*(1), 19-22.

Ward, J. (1963). Hierarchical Grouping to Optimize and Objective Function. *Journal of the American*

Statistical Association, 58(301), 236-244. doi: 10.2307/2282967

Κριτήρια Αξιολόγησης

Άσκηση Υπολογισμών 11.1

Δίνονται τα ακόλουθα οκτώ αντικείμενα, κάθε ένα από τα οποία αποτελείται από δύο αριθμητικές τιμές: A(1,3), B(1,7), C(4,6), D(5,6), E(6,8), F(7,5), G(7,7), H(7,9). Εφαρμόστε τη μέθοδο k-Means για τον διαμοιρασμό των αντικειμένων σε δύο συστάδες. Χρησιμοποιήστε την Ευκλείδεια απόσταση. Υλοποιήστε τα ακόλουθα βήματα:
Θεωρήστε ως αρχικά κέντρα τα σημεία C και G.
Υπολογίστε τις αποστάσεις των υπόλοιπων έξι σημείων από τα σημεία C και G.
Κατανείμετε τα σημεία σε συστάδες ανάλογα με τις αποστάσεις τους από τα σημεία C και G.
Υπολογίστε τα νέα κέντρα των συστάδων C1 και C2.
Υπολογίστε τις αποστάσεις όλων των σημείων από τα νέα κέντρα C1 και C2.
Κατανείμετε τα σημεία σε συστάδες ανάλογα με τις αποστάσεις τους από τα νέα κέντρα C1 Και C2.

Λύση

Για τον υπολογισμό της απόστασης των έξι αντικειμένων από τα αντικείμενα C και G χρησιμοποιήστε την Εξίσωση 11.1. Οι αποστάσεις δίνονται στον Πίνακα 11.2.

Αντικείμενο	X	Y	Dist to C	Dist to G
A	1	3	4,2426	7,211103
B	1	7	3,1623	6
D	5	6	1	2,236068
E	6	8	2,8284	1,414214
F	7	5	3,1623	2
H	7	9	4,2426	2

Πίνακας 11.2 *Αποστάσεις αντικειμένων από τα αντικείμενα C και G*

Σύμφωνα με τα στοιχεία του πίνακα, τα αντικείμενα A, B, D βρίσκονται πλησιέστερα στο C, ενώ τα αντικείμενα E,F,H βρίσκονται πλησιέστερα στο G. Η πρώτη συστάδα θα αποτελείται από τα αντικείμενα A, B, C, D και η δεύτερη συστάδα από τα αντικείμενα E, F, G, H.

Υπολογίστε τα νέα κέντρα των δύο συστάδων C1 και C2 σύμφωνα με την Εξίσωση 11.18. Τα νέα κέντρα είναι τα σημεία C1(2,75, 5,5) και C2(6,75, 7,25).

Υπολογίστε τις αποστάσεις των οκτώ σημείων από τα νέα κέντρα C1 και C2 σύμφωνα με την Εξίσωση 11.1. Οι αποστάσεις παρατίθενται στον Πίνακα 11.3.

Αντικείμενο	X	Y	Dist to C1	Dist to C2
A	1	3	3,0516	7,150175
B	1	7	2,3049	5,755432
C	4	6	1,3463	3,020761
D	5	6	2,3049	2,150581
E	6	8	4,1003	1,06066
F	7	5	4,2793	2,263846
G	7	7	4,5069	0,353553
H	7	9	5,5057	1,767767

Πίνακας 11.3 *Αποστάσεις σημείων από τα νέα κέντρα C1 και C2*

Σύμφωνα με τα στοιχεία του πίνακα, τα αντικείμενα A, B, C βρίσκονται πλησιέστερα στο C1, ενώ τα αντικείμενα D,E,F,G,H βρίσκονται πλησιέστερα στο C2. Η πρώτη συστάδα θα αποτελείται από τα αντικείμενα A,

B, C και η δεύτερη συστάδα από τα αντικείμενα D, E, F, G, H. Το αντικείμενο D αλλάζει συστάδα.

Άσκηση Εφαρμογής 11.2

Χρησιμοποιήστε το αρχείο «Wholesale Customers Data.csv». Θα το βρείτε στην ιστοσελίδα UCI Machine Learning Repository με το όνομα «Wholesale Customers». Το αρχείο περιέχει στοιχεία πωλήσεων σε πελάτες ενός διανομέα χονδρικής. Αποτελείται από οκτώ στήλες και 440 γραμμές. Η πρώτη στήλη αφορά το κανάλι διανομής, η δεύτερη τις περιοχές, και οι υπόλοιπες έξι στήλες αφορούν τις κατηγορίες προϊόντων. Οι τιμές αναφέρονται σε συνολικές ετήσιες πωλήσεις ανά κατηγορία προϊόντος. Σύμφωνα με την ιστοσελίδα UCI Machine Learning Repository, το συγκεκριμένο σύνολο δεδομένων είναι κατάλληλο για Ανάλυση Συστάδων και Κατηγοριοποίηση.
Πραγματοποιήστε με το WEKA Ιεραρχική Ανάλυση Συστάδων. Χρησιμοποιήστε διαδοχικά τύπο σύνδεσης, Single, Complete, Mean, Centroid και Ward. Για κάθε περίπτωση ορίστε να δημιουργηθούν τρεις συστάδες. Συγκρίνετε τα αποτελέσματα και τα δενδρογράμματα της κάθε περίπτωσης.

Λύση

Βήμα 1. Προμηθευτείτε το αρχείο από την ιστοσελίδα UCI Machine Learning Repository.

Το αρχείο είναι σε μορφότυπο Comma Separated Values (CSV). Πρέπει να μετατραπεί σε μορφότυπο ARFF. Ανοίξτε το αρχικό αρχείο στο Excel και αποθηκεύστε σε μορφότυπο XLS. Χρησιμοποιήστε την εφαρμογή μετατροπής αρχείων Excel σε αρχεία ARFF. Τη συγκεκριμένη εφαρμογή θα τη βρείτε στην ιστοσελίδα Excel to Arff Converter download. Μετατρέψτε το αρχείο, μεταφέρετε τα αποτελέσματα με αντιγραφή και επικόλληση σε κάποιον κειμενογράφο και αποθηκεύστε τα σε αρχείο κειμένου. Αλλάξτε την κατάληξη του αρχείου από TXT σε ARFF.

Βήμα 2. Εκκινήστε το WEKA και ανοίξτε το αρχείο «Wholesale Customers Data.csv», πιέζοντας το κουμπί «Open file».

Στο tab «Preprocess» μελετήστε τα γνωρίσματα και τις κατανομές τιμών. Μπορείτε να δείτε τα δεδομένα αναλυτικά κάνοντας κλικ στο κουμπί «Edit».

Βήμα 3. Μεταβείτε στο tab «Cluster».

Στο πεδίο «Clusterer» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka/clusterer/HierarchicalClusterer.

Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Βεβαιωθείτε ότι ως συνάρτηση απόστασης έχει οριστεί η Ευκλείδεια απόσταση και ότι ο τρόπος σύνδεσης είναι τύπου Single. Ορίστε να εξαχθούν τρεις συστάδες (πεδίο «numClusters»). Κάντε κλικ στο κουμπί «OK» για να επιστέψετε στο παράθυρο του WEKA και στη συνέχεια κάντε κλικ στο κουμπί «Start» για να εκτελέσετε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Η κατανομή των παρατηρήσεων σε συστάδες δεν είναι τόσο επιτυχημένη, καθώς η πρώτη συστάδα περιέχει 420 παρατηρήσεις, η δεύτερη μια παρατήρηση και η τρίτη δεκαεννέα παρατηρήσεις. Στο πεδίο «Results list», κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize tree» για να δείτε το δενδρόγραμμα.

Βήμα 4. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Αλλάξτε τον τρόπο σύνδεσης σε Complete. Ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 138, 298 και 4 παρατηρήσεις. Στο πεδίο «Results list» κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize tree» για να δείτε το δενδρόγραμμα.

Βήμα 5. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Αλλάξτε τον τρόπο σύνδεση σε Mean. Ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 120, 259 και 61 παρατηρήσεις. Η κατανομή των παρατηρήσεων σε συστάδες έγινε λιγότερο ανισομερής. Στο πεδίο «Results list» κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize tree» για να δείτε το δενδρόγραμμα.

Βήμα 6. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Αλλάξτε τον τρόπο σύνδεσης σε Centroid. Ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 142, 298 και 1

παρατηρήσεις. Δείτε με γραφικό τρόπο το δενδρόγραμμα.

Βήμα 7. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα HierarchicalClusterer. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Αλλάξτε τον τρόπο σύνδεσης σε Ward. Ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 142, 239 και 59 παρατηρήσεις. Δείτε με γραφικό τρόπο το δενδρόγραμμα.

Άσκηση Εφαρμογής 11.3

Χρησιμοποιήστε το αρχείο «Wholesale Customers Data.csv». Θα το βρείτε στην ιστοσελίδα UCI Machine Learning Repository με το όνομα «Wholesale Customers». Περισσότερες πληροφορίες για τα δεδομένα υπάρχουν στην εκφώνηση της Άσκησης 11.2
Πραγματοποιήστε με το WEKA Ανάλυση Συστάδων με τη μέθοδο k-Means. Δημιουργήστε τρεις συστάδες χρησιμοποιώντας την Ευκλείδεια απόσταση και την απόσταση Manhattan. Δείτε με γραφικό τρόπο την κατανομή των παρατηρήσεων σε συστάδες ανάλογα με τις τιμές του κάθε γνωρίσματος. Δημιουργήστε στα δεδομένα μια καινούργια στήλη, στην οποία θα αναφέρεται η συστάδα στην οποία ανήκει, σύμφωνα με τη μέθοδο k-Means με απόσταση Manhattan και πλήθος συστάδων ίσο με τρία.

Λύση

Βήμα 1. Προμηθευτείτε το αρχείο από την ιστοσελίδα UCI Machine Learning Repository.

Το αρχείο είναι σε μορφή Comma Separated Values (CSV). Πρέπει να μετατραπεί σε μορφή ARFF. Οδηγίες για τη μετατροπή θα βρείτε στη λύση της Άσκησης 11.2.

Εκκινήστε το WEKA και ανοίξτε το αρχείο «Wholesale Customers Data.arff» πιέζοντας το κουμπί «Open file».

Βήμα 2. Μεταβείτε στο tab «Cluster».

Στο πεδίο «Clusterer» κάντε κλικ στο κουμπί «Choose» και επιλέξτε weka/clusterer/SimpleKMeans.

Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα SimpleKMeans. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Βεβαιωθείτε ότι ως συνάρτηση απόστασης έχει οριστεί η Ευκλείδεια απόσταση. Ορίστε να εξαχθούν τρεις συστάδες (πεδίο «numClusters»). Κάντε κλικ στο κουμπί «OK» για να επιστέψετε στο παράθυρο του WEKA και στη συνέχεια κάντε κλικ στο κουμπί «Start» για να εκτελέσετε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 298, 132 και 10 παρατηρήσεις.

Βήμα 3. Στο πεδίο «Results list» κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize cluster assignments». Ανοίγει το παράθυρο οπτικοποίησης του WEKA, όπου οι παρατηρήσεις απεικονίζονται σε δισδιάστατο χώρο. Στον άξονα *x* είναι διατεταγμένες οι παρατηρήσεις, ενώ ο άξονας *y* αντιστοιχεί σε ένα από τα γνωρίσματα. Επιλέξτε διαδοχικά τα γνωρίσματα και παρατηρήστε την κατανομή των παρατηρήσεων και την ένταξη τους σε κλάσεις. Ο διαχωρισμός των κλάσεων είναι ιδιαίτερα εμφανής σύμφωνα με τις τιμές του πεδίου «Detergents_Paper».

Βήμα 4. Κάντε κλικ στα περιεχόμενα του πεδίου «Clusterer» και στο όνομα SimpleKMeans. Θα ανοίξει το παράθυρο ρύθμισης παραμέτρων. Επιλέξτε ως συνάρτηση απόστασης την απόσταση Manhattan και ορίστε να εξαχθούν τρεις συστάδες. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Clusterer output» εμφανίζονται τα αποτελέσματα. Οι τρεις συστάδες περιέχουν 298, 89 και 53 παρατηρήσεις.

Βήμα 5. Στο πεδίο «Results list» κάντε δεξί κλικ στην εγγραφή και επιλέξτε «Visualize cluster assignments». Ανοίγει το παράθυρο οπτικοποίησης του WEKA. Επιλέξτε διαδοχικά τα γνωρίσματα για τον άξονα *y* και παρατηρήστε την κατανομή των παρατηρήσεων και την ένταξη τους σε κλάσεις. Ξανά ο διαχωρισμός των κλάσεων είναι ιδιαίτερα εμφανής σύμφωνα με τις τιμές του πεδίου «Detergents_Paper» και σε μικρότερο βαθμό για τις τιμές των πεδίων «Grocery» και «Milk».

Βήμα 6. Μεταβείτε στο tab «Preprocess». Κάντε κλικ στο κουμπί «Choose» του πεδίου «Filter» και επιλέξτε weka/filters/unsupervised/attribute/AddCluster. Κάντε κλικ στα περιεχόμενα του πεδίου «Filter» και στο όνομα «AddCluster». Ανοίγει το παράθυρο ρύθμισης παραμέτρων. Κάντε κλικ στο κουμπί «Choose» του πεδίου «Clusterer» και επιλέξτε weka/clusterer/SimpleKMeans. Κάντε κλικ στα περιεχόμενα του πεδίου «clusterer» και στο όνομα «SimpleKMeans». Ανοίγει το παράθυρο ρύθμισης παραμέτρων του SimpleKMeans. Στο πεδίο «distanceFunction» επιλέξτε την απόσταση Manhattan και εισάγετε στο πεδίο «numClusters» την

τιμή 3. Κλείστε τα παράθυρα ρύθμισης παραμέτρων πατώντας το πλήκτρο «OK» και εκτελέστε τον αλγόριθμο κάνοντας κλικ στο κουμπί «Apply». Θα προστεθεί στα δεδομένα μια νέα στήλη, όπου θα αναγράφεται η συστάδα στην οποία ανήκει η εκάστοτε παρατήρηση.

Άσκηση Εφαρμογής 11.4

Χρησιμοποιήστε το αρχείο «Wholesale Customers Data.csv». Θα το βρείτε στην ιστοσελίδα UCI Machine Learning Repository με το όνομα «Wholesale Customers». Περισσότερες πληροφορίες για τα δεδομένα υπάρχουν στην εκφώνηση της Άσκησης 11.2
Πραγματοποιήστε με το WEKA Ανάλυση Συστάδων με τη μέθοδο k-Means. Δημιουργήστε τρεις συστάδες χρησιμοποιώντας την απόσταση Manhattan. Δημιουργήστε στα δεδομένα μια καινούργια στήλη, στην οποία θα αναφέρεται η συστάδα στην οποία ανήκει, σύμφωνα με τη μέθοδο k-Means με απόσταση Manhattan και πλήθος συστάδων ίσο με τρία.
Χρησιμοποιώντας ως τιμή κλάσης τη συστάδα στην οποία ανήκουν οι παρατηρήσεις, εκτελέστε κατηγοριοποίηση με Δένδρα Αποφάσεων C4.5.

Λύση

Βήμα 1. Προμηθευτείτε το αρχείο από την ιστοσελίδα UCI Machine Learning Repository.

Το αρχείο είναι σε μορφή Comma Separated Values (CSV). Πρέπει να μετατραπεί σε μορφή ARFF. Οδηγίες για τη μετατροπή θα βρείτε στη λύση της Άσκησης 11.2.

Εκκινήστε το WEKA και ανοίξτε το αρχείο «Wholesale Customers Data.arff» πιέζοντας το κουμπί «Open file».

Βήμα 2. Μεταβείτε στο tab «Cluster» και εκτελέστε την Ανάλυση Συστάδων σύμφωνα με τα στοιχεία της εκφώνησης και τις οδηγίες της Άσκησης 11.3. Προσθέστε τη στήλη με το όνομα της συστάδας σύμφωνα με τα στοιχεία της εκφώνησης και τις οδηγίες της Άσκησης 11.3.

Βήμα 3. Μεταβείτε στο tab «Classify».

Επιλέξτε κατηγοριοποιητή J48 και επικύρωση τύπου cross-validation. Εκτελέστε τον αλγόριθμο.

Στο πεδίο «Classifier Output» μελετήστε τα αποτελέσματα. Το Δένδρο Αποφάσεων επιτυγχάνει υψηλότατη ακρίβεια της τάξης του 97%. Επίσης, το δένδρο περιγράφει ένα κατανοητό τρόπο κατανομής των παρατηρήσεων στις συστάδες.

12 Διαχείριση Έργων Επιχειρηματικής Ευφυΐας

Σύνοψη

Αντικείμενο του παρόντος Κεφαλαίου είναι η διαχείριση έργων για την ανάπτυξη Συστημάτων Επιχειρηματικής Ευφυΐας (ΣΕΕ). Τα ΣΕΕ παρουσιάζουν σημαντικές διαφορές σε σχέση με άλλα τυπικά πληροφοριακά συστήματα. Οι κυριότερες διαφορές είναι ότι χρησιμοποιούν δεδομένα άλλων συστημάτων και ότι οι απαιτήσεις των χρηστών μεταβάλλονται διαρκώς, και μάλιστα με γρήγορο ρυθμό. Οι διαφορές αυτές καθιστούν τα γενικά μοντέλα ανάπτυξης πληροφοριακών συστημάτων ακατάλληλα για την περίπτωση των ΣΕΕ. Στο κεφάλαιο αυτό αναλύονται τα ειδικά θέματα, τα οποία σχετίζονται με τον κύκλο ζωής των ΣΕΕ. Αρχικά παρουσιάζεται το μοντέλο ανάπτυξης ΣΕΕ, το οποίο αποτελείται από διακριτά στάδια διατεταγμένα σε μια κυκλική αλληλουχία, η οποία αποτυπώνει τη διαρκή μετεξέλιξη των ΣΕΕ. Ακολούθως, για κάθε ένα από τα στάδια αυτά γίνεται παράθεση και σχολιασμός των ειδικών θεμάτων, τα οποία εμπίπτουν σε αυτό. Η αντιμετώπιση του αντικείμενου είναι πολυεπίπεδη. Έμφαση δίνεται στα επιχειρηματικά ζητήματα, τα οποία στα ΣΕΕ είναι βαρύνουσας σημασίας. Μελετώνται θέματα σχετικά με τα ιδιόμορφα υποσυστήματα των ΣΕΕ, όπως είναι οι Αποθήκες Δεδομένων, το υποσύστημα Εξαγωγής, Μετασχηματισμού και Φόρτωσης δεδομένων (Extract Transform Load (ETL)) και το υποσύστημα μεταδεδομένων, το οποίο παρουσιάζει τις δικές του ιδιαιτερότητες. Αναφορά γίνεται και σε τεχνικά θέματα, όπως τα χαρακτηριστικά της υλικοτεχνικής υποδομής. Παρέχονται πρακτικές οδηγίες για τη συγκρότηση της ομάδας έργου και τη διοργάνωση της διαδικασίας αξιολόγησης του έργου. Γενικώς, επιχειρείται μια πολύπλευρη αντιμετώπιση του αντικειμένου, καθίσταται όμως σαφές ότι η έκταση του είναι τέτοια που αποκλείει την εξαντλητική του κάλυψη στα πλαίσια ενός κεφαλαίου.
Σύντομη αναφορά γίνεται και στην Επιχειρηματική Ευφυΐα με υπολογιστική νέφους και λύσεις λογισμικού ως υπηρεσία. Τέλος, στο παρόν κεφάλαιο εξετάζονται οι κρίσιμοι παράγοντες επιτυχίας Συστημάτων Επιχειρηματικής Ευφυΐας, οι οποίοι επίσης διαφέρουν από τους αντίστοιχους άλλων πληροφοριακών συστημάτων. Κοινή εκτίμηση ερευνητών που ασχολήθηκαν με το αντικείμενο, είναι ότι μέχρι σήμερα δεν υπάρχει ένα καθιερωμένο και ευρέως αποδεκτό μοντέλο παραγόντων επιτυχίας για Συστήματα Επιχειρηματικής Ευφυΐας, και ότι απαιτείται περαιτέρω έρευνα. Στο κεφάλαιο παρουσιάζονται οι μέθοδοι και τα ευρήματα δημοσιευμένων εργασιών, κάθε μια από τις οποίες αναδεικνύει ορισμένους παράγοντες που επηρεάζουν την επιτυχία των ΣΕΕ. Οι εργασίες που επιλέχθηκαν δίνουν μια ικανοποιητική εικόνα της τρέχουσας κατάστασης πραγμάτων στον χώρο της εκτίμησης ποιότητας των ΣΕΕ. Στο τέλος του κεφαλαίου γίνεται μια απόπειρα σύνοψης των αποτελεσμάτων αυτών των εργασιών. Κοινή συνισταμένη είναι ότι οι παράγοντες, οι οποίοι σχετίζονται με τον οργανισμό και τις διαδικασίες του, επηρεάζουν περισσότερο την επιτυχία των ΣΕΕ, από τους παράγοντες που σχετίζονται με τεχνικά ζητήματα. Η στάση της ανώτατης διοίκησης απέναντι στο έργο και η αποφασιστική συνδρομή της αναγνωρίζεται ως ο σημαντικότερος παράγοντας επιτυχίας. Καθοριστικός παράγοντας είναι και η ύπαρξη ξεκάθαρου στρατηγικού προσανατολισμού, καθώς και η σύνδεση της επιχειρησιακής στρατηγικής με τη διαχείριση των επιχειρηματικών διαδικασιών.

Προαπαιτούμενη γνώση

Το Κεφάλαιο αυτό καλύπτει τον κύκλο ζωής ανάπτυξης Συστημάτων Επιχειρηματικής Ευφυΐας και τους σχετικούς παράγοντες επιτυχίας. Το αντικείμενο είναι ιδιαίτερα ευρύ. Επιδίωξη του συγγραφέα ήταν να αναπτυχθεί το θέμα με τέτοιον τρόπο, ώστε να είναι κατανοητό σε αναγνώστες που δεν έχουν προηγούμενες γνώσεις σχετικά με την ανάπτυξη και την ποιότητα πληροφοριακών συστημάτων. Ωστόσο, ο αναγνώστης θα λάβει σημαντική βοήθεια και θα κατανοήσει το αντικείμενο σε μεγαλύτερο βάθος, εάν ανατρέξει και σε άλλα συγγράμματα, τα οποία ασχολούνται με θεωρίες και μοντέλα ανάπτυξης πληροφοριακών συστημάτων. Οι θεωρίες αυτές αποτέλεσαν τη βάση για την ανάπτυξη αντίστοιχων μοντέλων, εξειδικευμένων στα ΣΕΕ. Για θέματα ανάπτυξης πληροφοριακών συστημάτων υποδεικνύεται το βιβλίο των Alexander and Maiden (2004), καθώς και τα άρθρα των Boehm (1988), των Larman and Basili (2003), των Davis, Bersoff and Comer (1988) και των Henderson-Sellers and Edwards (1990). Το άρθρο των DeLone and McLean (2003) είναι το πλέον διαδεδομένο για τους παράγοντες επιτυχίας πληροφοριακών συστημάτων. Ενδεικτικά άρθρα που ασχολούνται με το θέμα των παραγόντων επιτυχίας είναι τα Li (1997), Poon and Wagner (2001), και Sumner (1999). Επίσης, το Κεφάλαιο 4 για τις Αποθήκες Δεδομένων και την Πολυδιάστατη Ανάλυση, το Κεφάλαιο 6 για την Εξόρυξη Δεδομένων και το Κεφάλαιο 7 για την προεπεξεργασία των δεδομένων αναφέρονται σε θέματα σημαντικά για την κατανόηση του παρόντος κεφαλαίου.

12.1 Ανάπτυξη Συστημάτων Επιχειρηματικής Ευφυΐας

Η ανάπτυξη ενός συστήματος ΕΕ είναι κάτι πολύ πιο σύνθετο από την αγορά κάποιου εξειδικευμένου λογισμικού και του αντίστοιχου υλικού υπολογιστών. Στην πραγματικότητα, είναι μια περίπλοκη και μακροχρόνια διαδικασία, που απαιτεί κατάλληλες υποδομές και πόρους (Fuchs, 2006; Watson, Abraham, Chen, Preston & Thomas, 2004), καθώς και ανάλυση και βαθιά γνώση των διοικητικών διαδικασιών του οργανισμού, κυρίως αυτών που σχετίζονται με τη λήψη αποφάσεων. Σε έναν πραγματικό οργανισμό οι ανάγκες για πληροφόρηση δεν είναι σταθερές, αλλά μεταβάλλονται με την πάροδο του χρόνου. Αυτό σημαίνει συχνή μεταβολή των απαιτήσεων των χρηστών, μεταβολή στην οποία το σύστημα οφείλει να ανταποκριθεί. Μπορούμε να πούμε ότι η ανάπτυξη ενός συστήματος ΕΕ δεν είναι ένα έργο με τη συνήθη έννοια του όρου, αλλά είναι ένα έργο σε διαρκή εξέλιξη.

Οι ιδιομορφίες των συστημάτων ΕΕ δεν περιορίζονται στη συχνή, ίσως και συνεχή, μεταβολή των απαιτήσεων των χρηστών. Στο επίκεντρο ενός συστήματος ΕΕ βρίσκεται η Αποθήκη Δεδομένων. Η Αποθήκη Δεδομένων είναι μια βάση δεδομένων σημαντικά διαφορετική από αυτές των συστημάτων παρακολούθησης συναλλαγών, τα οποία παράγουν τα δικά τους δεδομένα και λειτουργούν επί αυτών. Η Αποθήκη Δεδομένων συνήθως περιλαμβάνει τεράστιους όγκους δεδομένων, που προέρχονται από πολλά διαφορετικά συστήματα. Επίσης, η δομή των δεδομένων είναι περίπλοκη, ώστε να επιτρέπει την ταχεία πρόσβαση στα δεδομένα, καθώς και τη χρήση τους από διαφορετικά τμήματα του οργανισμού και από στελέχη διαφορετικών επιπέδων.

Όπως αναφέρθηκε στο εισαγωγικό κεφάλαιο, τα τελευταία χρόνια η Επιχειρηματική Ευφυΐα γνωρίζει άνθηση, και οι μεγαλύτερες επιχειρήσεις παγκοσμίως την τοποθετούν στην κορυφή των τεχνολογικών προτεραιοτήτων τους. Οι λόγοι για αυτήν την τάση είναι πολλοί και αναλύονται διεξοδικά στο πρώτο Κεφάλαιο. Πέρα όμως από τους γενικούς λόγους, χρήσιμο είναι να καταγραφούν ορισμένα άμεσα και πρακτικά συμπτώματα και ενδείξεις, που σηματοδοτούν την ανάγκη οικοδόμησης ενός συστήματος ΕΕ σε έναν οργανισμό. Οι Gangadharan and Swami (2004) σταχυολογούν ορισμένα τέτοια ενδεικτικά σημεία:

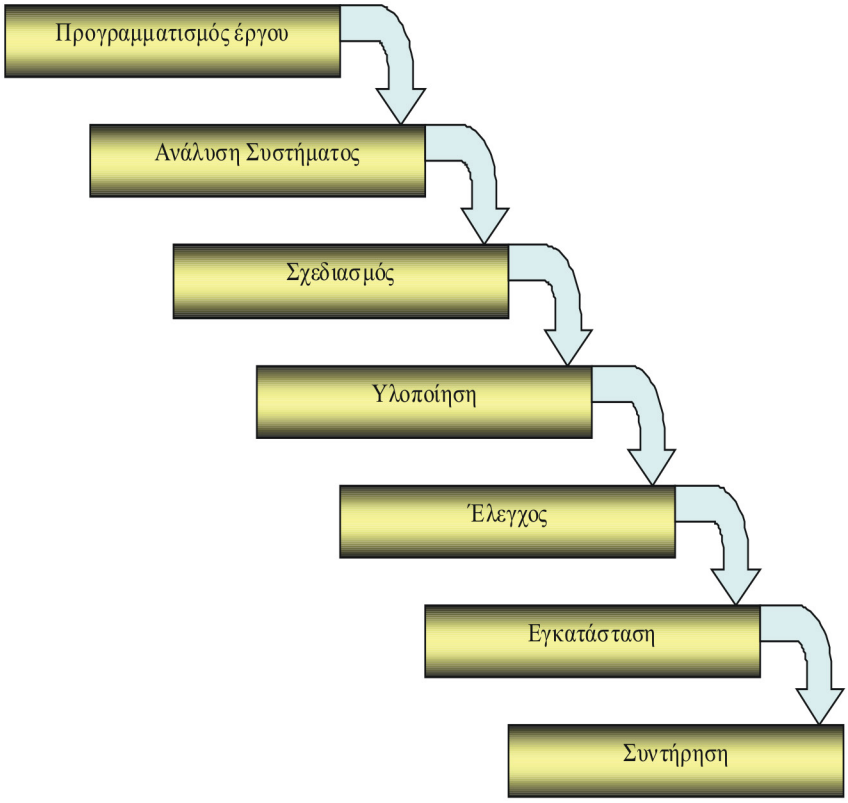
- Ύπαρξη τεράστιων ποσοτήτων δεδομένων, σε αντίθεση με μικρή ποσότητα πληροφορίας.
- Ανάγκη εύρεσης ιστορικής πληροφορίας.
- Υπεραπασχολημένο τμήμα πληροφορικής, που δεν έχει χρόνο για σύνταξη αναφορών.
- Ανάγκη για ενίσχυση των επιχειρηματικών διαδικασιών, ώστε να γίνουν πιο κερδοφόρες.
- Αδυναμία οργάνωσης των δεδομένων με τον επιθυμητό τρόπο.
- Ανάγκη για ταχύτερη λήψη αποφάσεων βασισμένη σε πληροφορία.
- Ανάγκη για δημιουργία ποιοτικών αναφορών, αντίστοιχων με τη δομή του οργανισμού.
- Υπερβολικός χρόνος που δαπανάται για συλλογή και ανάλυση των δεδομένων.

Στο παρόν κεφάλαιο γίνεται συνοπτική παρουσίαση των ζητημάτων που σχετίζονται με την ανάπτυξη αυτών των ιδιόμορφων πληροφοριακών συστημάτων. Μελετάται ο κύκλος ζωής τους, καθώς και οι παράγοντες που συμβάλλουν στην επιτυχία τους.

12.2 Ο Κύκλος Ζωής Ανάπτυξης Συστήματος Επιχειρηματικής Ευφυΐας.

Η ανάπτυξη λογισμικού είναι ένα αντικείμενο που γνώρισε ραγδαία εξέλιξη, όπως και κάθε άλλος τομέας της πληροφορικής. Αρχικά ισοδυναμούσε με την εργασία ενός προγραμματιστή που «έγραφε κώδικα» για να επιλύσει ένα πρόβλημα ή να αυτοματοποιήσει μια διαδικασία. Σύντομα όμως τα προγράμματα έγιναν υπερβολικά μεγάλα και περίπλοκα, και έγινε κατανοητό ότι η ανάπτυξη λογισμικού είναι ένα σύνθετο έργο, στο οποίο εμπλεκόταν στελέχη διαφόρων ειδικοτήτων. Επίσης, έγινε κατανοητό ότι ήταν απαραίτητη μια μεθοδολογία για τη διαχείριση του.

Οι μελέτες σχετικά με τη διαδικασία κατασκευής λογισμικού είχαν καρπό την έννοια του Κύκλου Ζωής Ανάπτυξης Συστήματος (ΚΖΑΣ) (System Development Life Cycle (SDLC)), ο οποίος καθορίζει τα στάδια κατασκευής ενός πληροφοριακού συστήματος. Επίσης, με την πάροδο του χρόνου προτάθηκαν διάφορα μοντέλα που περιγράφουν τον ΚΖΑΣ. Το πρώτο, γνωστότερο και ίσως πιο διαδεδομένο μοντέλο ήταν αυτό της «υδατόπτωσης». Σύμφωνα με το μοντέλο της υδατόπτωσης, η ανάπτυξη ενός πληροφοριακού συστήματος ήταν μια διαδικασία διαδοχικών σταδίων, όπου το αποτέλεσμα του ενός σταδίου τροφοδοτούσε το επόμενο στάδιο. Τα στάδια ήταν διατεταγμένα γραμμικά. Υπήρχε ένα αφετηριακό σημείο του έργου και ένα σημείο τερματισμού του. Κατά καιρούς προτάθηκαν διάφορες εκδοχές ως προς το πόσα και ποια ήταν τα στάδια, χωρίς να θίγεται όμως το βασικό σκεπτικό. Το Σχήμα 12.1 παρουσιάζει το μοντέλο της υδατόπτωσης.



Σχήμα 12.1 *Μοντέλο Υδατόπτωσης*

Το μοντέλο της υδατόπτωσης κατά καιρούς δέχτηκε κριτικές ως ανεπαρκές για την αποτύπωση της πραγματικής διαδικασίας ανάπτυξης ενός συστήματος. Καταρχήν, το μοντέλο θεωρούσε ότι ο αποκλειστικός ρόλος των χρηστών ήταν ο καθορισμός των απαιτήσεων και ότι όλες οι απαιτήσεις μπορούσαν να προκαθοριστούν. Επίσης, σε πραγματικά συστήματα, κάποια στάδια μπορούν να ξεκινήσουν πριν ολοκληρωθεί το προηγούμενο στάδιο. Επιπλέον, είναι δυνατή η επιστροφή σε κάποιο προηγούμενο στάδιο και η αναθεώρηση των αποτελεσμάτων του. Για τους λόγους αυτούς προτάθηκαν και άλλα εναλλακτικά μοντέλα.

Για την περίπτωση των Συστημάτων Επιχειρηματικής Ευφυΐας, το μοντέλο της υδατόπτωσης κρίνεται ακατάλληλο. Το βασικό πρόβλημα είναι ότι τα συστήματα αυτά δεν μηχανοργανώνουν επιχειρηματικές διαδικασίες καθημερινής λειτουργίας, οι οποίες είναι τυποποιημένες και σταθερές για μεγάλα χρονικά διαστήματα. Αντιθέτως, αφορούν διοικητικές διαδικασίες λήψης αποφάσεων, οι οποίες δεν είναι τυποποιημένες και αυστηρά καθορισμένες. Οι απαιτήσεις των λεγόμενων εργατών γνώσης για πληροφορία μεταβάλλονται με μεγάλη συχνότητα. Στη σχετική βιβλιογραφία, ο αναγνώστης μπορεί να βρει και άλλους λόγους, για τους οποίους το κλασσικό μοντέλο της υδατόπτωσης είναι ακατάλληλο για την ανάπτυξη συστημάτων ΕΕ. Οι Moss and Atre (2003) επισημαίνουν ότι το μοντέλο αυτό είναι κατάλληλο για στατικά συστήματα, που επιλύουν απομονωμένα προβλήματα ενός επιχειρησιακού τομέα. Αντιθέτως, τα συστήματα ΕΕ είναι ολοκληρωμένα και διαπερνούν οριζοντίως όλους τους τομείς της επιχείρησης. Επίσης, οι παραδοσιακές μεθοδολογίες δεν καλύπτουν το ζήτημα του στρατηγικού σχεδιασμού, ούτε περιλαμβάνουν την έννοια των διαδοχικών εκδόσεων του συστήματος. Στα συστήματα ΕΕ, τόσο τα δεδομένα όσο και οι λειτουργικότητες τροποποιούνται σε διαδοχικές εκδόσεις, και κάθε έκδοση μπορεί να αναδείξει νέες απαιτήσεις χρηστών. Οι Yeoh and Koronios (2010) τονίζουν ότι το τελικό προϊόν του συστήματος, δηλαδή οι παραγόμενες πληροφορίες, αξιολογούνται διαρκώς από τους χρήστες, και ότι το σύστημα θα επιτύχει μόνον εάν οι χρήστες συνεχώς ανακαλύπτουν και μοντελοποιούν νέα γνώση και τροποποιούν τα δεδομένα. Για τον λόγο αυτό, η όλη διαδικασία ανάπτυξης του συστήματος είναι κυκλική. Σε κυκλικό μοντέλο ανάπτυξης αναφέρονται και οι Gangadharan and Swami (2004), καθώς και οι Bara et al. (2009).

Όλοι οι προαναφερόμενοι συγγραφείς συγκλίνουν στην άποψη ότι το πλέον κατάλληλο μοντέλο για την ανάπτυξη συστημάτων Ε.Ε έχει κυκλική δομή και αποτελείται από διαδοχικά βήματα. Ο κύκλος ζωής ενός συστήματος Ε.Ε. απεικονίζεται στο Σχήμα 12.2. Τα διαδοχικά στάδια που απαρτίζουν τον κύκλο ζωής του συστήματος είναι:

- η αιτιολόγηση του έργου,
- η οργάνωση του έργου,
- η ανάλυση απαιτήσεων του συστήματος,
- ο σχεδιασμός,
- η κατασκευή,
- η εφαρμογή,
- η αξιολόγηση.

Κάθε ένα από τα παραπάνω στάδια παρουσιάζονται αναλυτικά στη συνέχεια του κεφαλαίου.

Σχήμα 12.2 *Μοντέλο ανάπτυξης Συστημάτων Ε.Ε.*

12.2.1 Αιτιολόγηση Έργου

Στο πρώτο στάδιο του κύκλου ζωής ενός συστήματος ΕΕ τεκμηριώνονται οι αιτίες για την υλοποίηση του. Γίνεται καταγραφή των επιχειρηματικών ευκαιριών ή των προβλημάτων που υπάρχουν, και εξηγείται αναλυτικά το πώς το έργο θα συμβάλει στην αντιμετώπιση τους. Έρευνες έχουν αποδείξει ότι η ύπαρξη ενός συγκεκριμένου επιχειρηματικού ζητούμενου είναι σημαντικότατος όρος για την επιτυχία του συστήματος (Yeoh & Koronios, 2010). Έργα, τα οποία έχουν ξεκινήσει με το σκεπτικό ότι απαιτείται γενικώς βελτίωση της πληροφόρησης στον οργανισμό, και δεν είχαν από την αρχή ξεκαθαρίσει ποια επιχειρηματικά ζητήματα θα εξυπηρετηθούν από το έργο και πως θα εξυπηρετηθούν, έχουν πολλές πιθανότητες να οδηγηθούν σε αποτυχία. Εσφαλμένη είναι επίσης μια αντίληψη, η οποία έχει ως αφετηριακό της σημείο την αξιοποίηση της πληροφορικής τεχνολογίας. Υπάρχουν περιπτώσεις στελεχών επιχειρήσεων, ενθουσιωδών οπαδών της πληροφορικής και της τεχνολογίας, που πήραν την πρωτοβουλία για την ανάπτυξη ενός συστήματος ΕΕ, ωθούμενοι από την επιθυμία τους να αξιοποιήσουν στο έπακρο τις νέες τεχνολογίες. Μια τέτοια προσέγγιση μπορεί να οδηγήσει στη δημιουργία ενός περίτεχνου, αλλά και ακριβού συστήματος, το οποίο όμως θα είναι ανίκανο να παράξει αξία για τον οργανισμό. Επαναλαμβάνεται με έμφαση ότι αφετηριακό σημείο πρέπει να είναι η ανάγκη αντιμετώπισης συγκεκριμένων επιχειρηματικών προβλημάτων ή η δυναμική αξιοποίησης επιχειρηματικών

ευκαιριών.

Στο στάδιο αυτό γίνεται και η τεκμηρίωση της στρατηγικής διάστασης του έργου. Τα έργα ΕΕ καλύπτουν ανάγκες διοικητικής πληροφόρησης, θεωρώντας την ολότητα του οργανισμού. Επίσης, φείλουν να εντάσσονται οργανικά στη στρατηγική του, στην εξυπηρέτηση δηλαδή των στρατηγικών του στόχων. Η τεκμηρίωση της αναγκαιότητας εκτέλεσης του έργου πρέπει να περιλαμβάνει και τον σαφή καθορισμό του πώς η παραγόμενη πληροφορία αξιοποιείται για την αντιμετώπιση συγκεκριμένων επιχειρηματικών ζητημάτων, του πώς τα ζητήματα αυτά συναρτώνται με τη στρατηγική του οργανισμού και του πώς το έργο θα εξυπηρετήσει τους στρατηγικούς στόχους της επιχείρησης.

Η αιτιολόγηση του έργου περιλαμβάνει και την ανάλυση κόστους-οφέλους. Καταρχήν εκτιμάται το συνολικό αναγκαίο κόστος για την οικοδόμηση και τη συνεχή λειτουργία του συστήματος. Επιπροσθέτως, γίνεται καταγραφή των οφελών που θα προκύψουν από τη λειτουργία του συστήματος και εκτιμάται η αξία που θα παραχθεί από τη χρήση του. Η αποτίμηση των οφελών δεν είναι πάντα εύκολη ούτε προφανής. Ωστόσο, εάν το σύστημα αποτελεί απάντηση σε συγκεκριμένα προβλήματα με μετρήσιμα αποτελέσματα, πχ σε κάποια δραστηριότητα που προκαλεί ζημίες συγκεκριμένου ύψους, είναι ευκολότερη η ποσοτικοποίηση των οφελών. Ένα σύστημα ΕΕ μπορεί να παράξει όφελος αυξάνοντας τα έσοδα και τα κέρδη, συμπιέζοντας το κόστος, καθώς και αυξάνοντας το μερίδιο της αγοράς και την ικανοποίηση των πελατών.

Στο στάδιο αυτό γίνεται επίσης η εκτίμηση του κινδύνου σχετικά με διάφορες παραμέτρους του έργου. Η εργασία αυτή επιτρέπει την καλύτερη διαχείριση των κινδύνων και καθιστά εφικτή τη ρεαλιστική εκτίμηση των πραγματικών αποτελεσμάτων του έργου. Παράμετροι, σε σχέση με τις οποίες πρέπει να εκτιμηθούν ενδεχόμενοι κίνδυνοι, είναι η τεχνολογία που θα εφαρμοστεί, χρηματοοικονομικά θέματα, ο βαθμός πολυπλοκότητας του έργου, ζητήματα σχετικά με την ενοποίηση και ολοκλήρωση των δεδομένων, καθώς και θέματα σχετικά με τη σύνθεση της ομάδας έργου. Σημειώνεται ότι όλες οι παραπάνω εργασίες πρέπει να εκτελούνται σε κάθε επανάληψη του κύκλου ζωής του συστήματος.

12.2.2 Οργάνωση Έργου

Αντικείμενο της οργάνωσης του έργου είναι η κατάρτιση του σχεδίου, που θα αποτελέσει οδηγό για την εκτέλεση του. Αφετηριακό σημείο ενός τέτοιου σχεδίου είναι η καταγραφή και αποτίμηση των υπαρχουσών υποδομών. Ο όρος υπάρχουσες υποδομές αναφέρεται τόσο σε τεχνικά όσο και σε μη τεχνικά θέματα.

Η υλικοτεχνική υποδομή του οργανισμού περιλαμβάνει το υπάρχων υλικό υπολογιστών και το λογισμικό. Απαιτείται μια καταρχήν καταγραφή του υλικού υπολογιστών που διαθέτει ο οργανισμός, καθώς και η αξιολόγηση του. Το υλικό θα πρέπει να μπορεί να ανταποκριθεί στις απαιτήσεις του νέου συστήματος. Δεδομένου ότι στα συστήματα ΕΕ τείνουν να εξελίσσονται διαρκώς, με αποτέλεσμα να μεγεθύνεται ο όγκος των δεδομένων και να αυξάνονται οι ανάγκες σε υπολογιστική ισχύ, πρέπει να ληφθεί μέριμνα και για την ικανοποίηση και των μελλοντικών απαιτήσεων. Σε περίπτωση που το υπάρχων υλικό κριθεί ανεπαρκές, πρέπει να αγοραστεί νέο υλικό, συμβατό με την υπάρχουσα υποδομή και να συνυπολογιστεί το σχετικό κόστος. Καταγράφεται και αξιολογείται και η υπάρχουσα δικτυακή υποδομή, το διαθέσιμο εύρος ζώνης και οι δυνατότητες αναβάθμισης του.

Σε ότι αφορά το λογισμικό, καταγράφονται τα συστήματα διαχείρισης βάσεων δεδομένων που είναι εγκατεστημένα στον οργανισμό. Εάν ο οργανισμός εκτελεί ήδη κάποιες εργασίες ολοκλήρωσης δεδομένων, ανάλυσης τους και σύνταξης αναφορών, καταγράφεται το λογισμικό που χρησιμοποιείται. Επίσης ελέγχεται η πληροφοριακή υποδομή που παράγει τα πηγαία δεδομένα. Τέτοια υποδομή μπορεί να είναι συστήματα ERP, CRM, SCM κλπ. Η υπάρχουσα υποδομή σε λογισμικό αξιολογείται και καθορίζονται οι ανάγκες αγοράς ή και συγγραφής νέου λογισμικού. Σε αρκετές περιπτώσεις τα συστήματα Ε.Ε. πρέπει να είναι προσβάσιμα με πολλούς εναλλακτικούς τρόπους, που περιλαμβάνουν και τη χρήση κινητών συσκευών. Σε τέτοιες περιπτώσεις ο τρόπος πρόσβασης είναι ένα επιπλέον ζήτημα και πιθανώς να είναι αναγκαία η κατασκευή ενός portal.

Πέραν της υλικοτεχνικής υποδομής, η ύπαρξη και λειτουργία του οργανισμού συνεπάγεται την ύπαρξη μια άλλης, μη τεχνικής υποδομής, που σχετίζεται με οργανωτικές δομές, διαδικασίες, τρόπους λειτουργίας, κανόνες, ρόλους κλπ. Η υποδομή αυτή είναι σημαντική και μπορεί να επηρεάζει την ανάπτυξη του συστήματος. Η εφαρμογή του συστήματος θα επιφέρει αλλαγές, οπότε απαιτείται μια πολιτική διαχείρισης των αλλαγών (change management). Επίσης, ο οργανισμός λογικά θα διαθέτει διαδικασίες ελέγχου, αντιμετώπισης τεχνικών προβλημάτων και διευθέτησης διαφωνιών. Τέτοια ζητήματα πιθανώς θα προκύψουν κατά την εκτέλεση του έργου. Η καταγραφή αυτών των πάγιων διαδικασιών του οργανισμού μπορεί να συμβάλει στην αντιμετώπιση προβλημάτων, όταν αυτά προκύψουν.

Μια σημαντική εργασία, που πρέπει να εκτελεστεί στα πλαίσια της καταγραφής της υπάρχουσας υποδο-

μής, είναι η τυποποίηση της δομής των πηγαίων δεδομένων και η αντιμετώπιση σχετικών ζητημάτων. Θα πρέπει να αναζητηθεί το λογικό μοντέλο των δεδομένων, εάν αυτό υπάρχει. Εάν δεν υπάρχει, θα πρέπει να οριστούν οι υπεύθυνοι γι' αυτήν την εργασία, καθώς και άλλοι για την επικύρωση των αποτελεσμάτων. Ένα συχνό πρόβλημα που παρουσιάζεται είναι η ονοματοδοσία των δεδομένων. Η χρήση διαφορετικών ονομάτων για τα ίδια δεδομένα, όπως και η χρήση διαφορετικών μονάδων μέτρησης (πχ μέτρα, ίντσες κλπ.), καταγράφεται και διευθετείται. Επίσης, εάν ο οργανισμός τηρεί μεταδεδομένα, πρέπει να καταγραφεί το λογισμικό που χρησιμοποιείται γι' αυτήν την εργασία, το περιεχόμενο τους και η διαδικασία ενημέρωσης τους.

Αφού ολοκληρωθεί η καταγραφή των υποδομών, γίνεται η καθαυτό οργάνωση του έργου, η σύνταξη δηλαδή ενός αναλυτικού σχεδίου δράσης για την υλοποίηση του έργου. Το σχέδιο δράσης είναι ιδιαίτερα σημαντικό, γιατί θα καθορίσει τη διαδικασία εκτέλεσης του έργου. Επίσης, ο απολογισμός του έργου, όταν ολοκληρωθεί, θα γίνει σε συνδυασμό με το σχέδιο δράσης. Σε βασικές γραμμές, το σχέδιο θα πρέπει να δίνει απαντήσεις σε τέσσερεις ερωτήσεις:

- Τι είναι αυτό που θα παραδοθεί με την ολοκλήρωση του έργου;
- Πότε θα ολοκληρωθεί το έργο;
- Πόσο θα κοστίσει;
- Ποιοι θα συμμετέχουν και με ποιο ρόλο;

Αρχικά γίνεται καταγραφή των σκοπών και των στόχων του έργου. Συντάσσεται η έκθεση που αναφέρει τους λόγους για την εκπόνηση του έργου και τεκμηριώνεται η ανάγκη αυτή σε συνάρτηση με τη στρατηγική της επιχείρησης. Επίσης, καθορίζεται με σαφήνεια το πεδίο και η έκταση του έργου. Θεωρητικά, οι ανάγκες για πληροφόρηση σε έναν οργανισμό είναι απεριόριστες. Πρέπει λοιπόν να γίνει επιλογή ποιες από αυτές τις ανάγκες θα καλυφθούν και ποιες όχι. Χρήσιμο είναι να αναφέρεται ρητά, όχι μόνον ποια δεδομένα, λειτουργικότητες και δυνατότητες πληροφόρησης θα περιληφθούν, αλλά και ποιες θα αποκλειστούν. Αυξημένη προσοχή πρέπει να δοθεί στην επιλογή των δεδομένων, καθώς τα συστήματα Ε.Ε. είναι περισσότερο προσανατολισμένα προς τα δεδομένα και λιγότερο στις λειτουργικότητες.

Ακολούθως γίνεται ο προγραμματισμός εκτέλεσης του έργου. Στο πλαίσιο αυτού του προγραμματισμού γίνονται οι ακόλουθες εργασίες:

- Το έργο χωρίζεται σε μικρότερα, καλά καθορισμένα τμήματα (υποέργα), τα οποία καταγράφονται.
- Εκτιμάται ο απαραίτητος χρόνος για κάθε ένα από αυτά τα υποέργα.
- Καθορίζονται οι απαραίτητοι πόροι για κάθε ένα από αυτά τα υποέργα.
- Εντοπίζονται και καταγράφονται οι σχέσεις αλληλεξάρτησης των υποέργων.

Κάθε ένα από τα υποέργα κοστολογείται. Οι πόροι οι οποίοι θα διατεθούν, περιλαμβάνουν τα αναγκαία κεφάλαια, αλλά περιλαμβάνουν και τη διαθεσιμότητα ανθρώπινου δυναμικού, και μάλιστα κατάλληλου στελεχιακού επιπέδου. Αυτό μεταφράζεται σε κατανομή καθηκόντων και αρμοδιοτήτων. Ορίζονται τα μέλη της ομάδας έργου και κάθε μέλος αναλαμβάνει τις αρμοδιότητες του, δεσμεύοντας τον αναγκαίο χρόνο. Ως προς τη σχέση αλληλεξάρτησης των υποέργων, ορισμένα έργα δεν μπορούν να αρχίσουν εάν δεν ολοκληρωθούν προηγουμένως κάποια άλλα, ενώ κάποια έργα μπορούν να αρχίσουν παράλληλα. Οι σχέσεις αυτές καταγράφονται για την καλύτερη παρακολούθηση της πορείας εκτέλεσης του έργου, καθώς και για τη διευκόλυνση του απολογισμού σε περίπτωση καθυστερήσεων.

Στο τελικό στάδιο συντάσσεται το χρονοδιάγραμμα εκτέλεσης του έργου. Περιλαμβάνει τους χρόνους εκτέλεσης των υποέργων, προβλέπει καθυστερήσεις που οφείλονται σε αστάθμητους παράγοντες και λαμβάνει υπόψη τις σχέσεις αλληλεξάρτησης των υποέργων. Επίσης, συντάσσεται ο τελικός προϋπολογισμός. Σημαντικό είναι το έργο να έχει εξασφαλισμένη χρηματοδότηση, με την έννοια ότι η διοίκηση του οργανισμού αντιλαμβάνεται την αναγκαιότητα του έργου και είναι αποφασισμένη να λάβει τα απαραίτητα μέτρα για τη χρηματοδότηση του.

12.2.3 Ανάλυση απαιτήσεων του έργου

Η ανάλυση απαιτήσεων είναι το σημαντικότερο στάδιο στην ανάπτυξη του συστήματος. Στο στάδιο αυτό καθορίζονται οι εργασίες που θα εκτελεί το σύστημα, καθώς και ο τρόπος που θα τις εκτελεί. Όμως οι απαιτήσεις ενός συστήματος ΕΕ είναι περισσότερο απαιτήσεις στρατηγικής πληροφορίας και λιγότερο απαιτήσεις για λειτουργικά θέματα. Η ανάλυση απαιτήσεων δίνει έμφαση στην ανάλυση επιχειρηματικών ζητημάτων και όχι στην ανάλυση του συστήματος με την κλασική έννοια. Αφετηριακό σημείο για την ανάλυση είναι καταρχήν ο

καθορισμός των επιχειρηματικών ζητημάτων και των αναγκών για στρατηγική πληροφόρηση που θα καλύψει το σύστημα. Σε επίπεδο συστήματος, οι απαιτήσεις αυτές θα μεταφραστούν σε προσδιορισμό των εργασιών που πρέπει να εκτελούνται, καθώς και των δεδομένων που θα χρησιμοποιηθούν. Οι περισσότερες αποτυχίες παλαιότερων συστημάτων, και όχι μόνο Επιχειρηματικής Ευφυΐας, οφείλονται σε σφάλματα και ελλείψεις κατά την ανάλυση. Επίσης, αβλεψίες κατά την ανάλυση είναι δυνατόν να αποπροσανατολίσουν το σύστημα και γι' αυτό είναι δυσκολότερο να διορθωθούν.

Η εργασία του εντοπισμού και καταγραφής των απαιτήσεων του συστήματος διεξάγεται με οργάνωση συνεντεύξεων και χρήση ερωτηματολογίων. Ο κάθε ένας από τους συμμετέχοντες συνεισφέρει τη δική του ειδική γνώση για την επιτυχή ολοκλήρωση του έργου. Τα ανώτατα διοικητικά στελέχη γνωρίζουν ποιες πληροφορίες πρέπει να παράγει το σύστημα, τι ερωτήσεις πρέπει να απαντηθούν και τι αναφορές πρέπει να συνταχθούν. Επίσης, έχουν αντίληψη της στρατηγικής διάστασης του έργου και κατανοούν τον τρόπο με τον οποίο αυτό θα εξυπηρετήσει την υλοποίηση των στρατηγικών στόχων. Το ζήτημα αυτό είναι από τα κυριότερα και για τον λόγο αυτό η συμμετοχή και ουσιαστική συνέργεια των διοικητικών στελεχών έχει βαρύνουσα σημασία. Τα στελέχη πληροφορικής διαθέτουν τεχνικές γνώσεις, κατανοούν τα ζητήματα της τεχνικής υποδομής και γνωρίζουν την κατάσταση, τις δυνατότητες και τις αδυναμίες των υπαρχόντων δεδομένων. Τέλος, αναλυτές δεδομένων, που ήδη απασχολούνται στον οργανισμό, αντιλαμβάνονται τις ανάγκες ανάλυσης και γνωρίζουν τα προβλήματα των δεδομένων.

Η τελική έκθεση θα περιγράφει τις απαιτήσεις του συστήματος και θα ορίζει το πεδίο του συστήματος, δηλαδή το τι θα περιλαμβάνει το σύστημα και τι θα αποκλείει. Επίσης, πρέπει να είναι διατυπωμένη με επιχειρηματικούς όρους και όχι με όρους λειτουργικότητας. Μια έκθεση που θα παραθέτει τις λειτουργίες που θα εκτελεί το σύστημα δεν είναι η πρώτη προτεραιότητα. Αντιθέτως, πρέπει να δίνονται απαντήσεις στα παρακάτω ερωτήματα:

- Ποιο είναι το πρόβλημα που αντιμετωπίζει η επιχείρηση;
- Τι κόστος επιφέρει αυτό το πρόβλημα στην επιχείρηση;
- Πώς το σύστημα ΕΕ θα επιλύσει το πρόβλημα;
- Πώς η επίλυση του προβλήματος συναρτάται με τον στρατηγικό προσανατολισμό του οργανισμού;
- Ποιοι είναι οι χρήστες του συστήματος και πώς θα διευκολυνθεί η εργασία τους;
- Η κουλτούρα της επιχείρησης πριμοδοτεί την αξιοποίηση της πληροφορίας ως περιουσιακό στοιχείο;
- Ποιες αναφορές θα συντάσσονται, τι ερωτήματα θα εκτελούνται και τι πληροφορίες θα παράγονται;
- Πώς θα παρουσιάζεται και πώς θα διανέμεται η πληροφορία;

Τα συστήματα Ε.Ε. χαρακτηρίζονται «data intensive» ή, όπως θα μπορούσε να μεταφραστεί ελληνικά, «εντάσεως δεδομένων». Αυτό σημαίνει ότι συγκριτικά μεγαλύτερο μέρος των εργασιών τους σχετίζεται με τα δεδομένα και όχι με λειτουργίες. Για τον λόγο αυτό, η έκθεση ανάλυσης απαιτήσεων θα ασχολείται εκτεταμένα με ζητήματα δεδομένων. Ειδικότερα θα δίνονται απαντήσεις σε ερωτήματα όπως:

- Ποια δεδομένα θα χρησιμοποιούνται;
- Ποιο είναι το λογικό μοντέλο δεδομένων που απαιτείται;
- Ποιες θα είναι οι διαστάσεις των κύβων;
- Ποιες θα είναι οι διαδικασίες καθαρισμού και μετασχηματισμού των δεδομένων;
- Σε τι βαθμό λεπτομέρειας θα τηρούνται δεδομένα;
- Σε τι βάθος χρόνου θα τηρούνται δεδομένα;

Τα συστήματα ΕΕ διαπερνούν οριζοντίως τους οργανισμούς, χρησιμοποιούνται από πολλά τμήματα τους και χρησιμοποιούν δεδομένα από αυτά τα τμήματα. Τα διαφορετικά τμήματα πιθανότατα τηρούν τα δικά τους δεδομένα. Τα δεδομένα αυτά ενοποιούνται, ολοκληρώνονται και συγχωνεύονται. Η ανάλυση απαιτήσεων εξετάζει και αυτά τα θέματα. Αναφορικά με τις λειτουργίες που θα εκτελεί το σύστημα, καταγράφονται οι εκθέσεις που θα συντάσσονται, τα ερωτήματα που θα υποβάλλονται στο σύστημα, οι εργασίες ανάλυσης που θα εκτελούνται (πχ εργασίες OLAP, στατιστικής ανάλυσης ή εργασίες προγνωστικής ανάλυσης με τεχνικές εξόρυξης δεδομένων), καθώς και οι τρόποι οπτικοποίησης και διανομής της πληροφορίας. Η έκθεση απαιτήσεων του συστήματος ολοκληρώνεται με την καταγραφή των απαιτήσεων για επέκταση της υλικοτεχνικής υποδομής (υλικό υπολογιστών, βάσεις δεδομένων, εργαλεία εξόρυξης δεδομένων, υποδομή δικτύωσης κλπ.), με την καταγραφή της μη υλικοτεχνικής υποδομής (διαδικασίες, πρότυπα, ρόλοι, οδηγίες), καθώς και με ζητήματα ασφαλείας του συστήματος, όπως ποιοι θα έχουν πρόσβαση σε αυτό, αν η πρόσβαση θα είναι διαβαθμισμένη ανάλογα με τον χρήστη κλπ.

Μια τακτική, που χρησιμοποιείται συχνά για την ανάλυση απαιτήσεων, είναι η δημιουργία ενός πρωτοτύπου του συστήματος (prototyping). Το πρωτότυπο είναι μια περιορισμένη έκδοση του τελικού συστήματος και περιλαμβάνει μόνον κάποιες πλήρεις λειτουργίες ή πιθανώς και μερικές, όχι πλήρεις, λειτουργίες. Το πρωτότυπο δίνεται στους χρήστες και αυτοί το χρησιμοποιούν και το αξιολογούν. Οι χρήστες μπορεί να εντοπίσουν λάθη, αδυναμίες και παραλήψεις. Με αυτόν τον τρόπο, οι αναλυτές του συστήματος δέχονται ανατροφοδότηση από τους χρήστες, από τα αρχικά ακόμα στάδια, όταν οι τροποποιήσεις στο σύστημα είναι ευκολότερες. Η μέθοδος του πρωτοτύπου επιτρέπει επίσης πειραματισμούς με διαφορετικές τεχνολογίες, όπως πχ συστήματα βάσεων δεδομένων. Το πρωτότυπο μπορεί να παραδίδεται σε διαδοχικές εκδόσεις, κάθε μια από τις οποίες διορθώνει τα προβλήματα της προηγούμενης έκδοσης, ενώ ταυτόχρονα προσθέτει και νέες λειτουργίες. Το πρωτότυπο μπορεί επίσης να εξελίσσεται διαρκώς, μέχρι να πάρει τη μορφή ολοκληρωμένου συστήματος.

Πριν κατασκευαστεί ένα πρωτότυπο, ορίζεται η αιτία της δημιουργίας του, ορίζεται δηλαδή το τι θα περιλαμβάνει, σε ποιους απευθύνεται και τι είδους ανατροφοδότηση αναμένεται από τους χρήστες. Επειδή οι εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης των δεδομένων είναι χρονοβόρες και ακριβές, το πρωτότυπο συνήθως λειτουργεί με ένα υποσύνολο των δεδομένων. Επίσης, υλοποιεί ένα υποσύνολο των λειτουργιών. Σκοπός είναι να επιδείξει στους χρήστες κάποιες από τις αναλυτικές δυνατότητες του συστήματος, καθώς και τον τρόπο παρουσίασης της πληροφορίας. Η ανατροφοδότηση από τους χρήστες μπορεί να προκαλέσει μεταβολές στις λειτουργίες, στα δεδομένα ή και στα μεταδεδομένα του συστήματος.

Υπάρχουν διαφόρων ειδών πρωτότυπα. Κάθε ένα από αυτά προσφέρει διαφορετικές δυνατότητες, αλλά έχει και διαφορετικές απαιτήσεις σε χρόνο και κόστος. Στην απλούστερη εκδοχή του το πρωτότυπο περιέχει μόνον τη διεπαφή (interface) με τους χρήστες. Με τον τρόπο αυτό, οι χρήστες αποκτούν αντίληψη των λειτουργιών του συστήματος, των δεδομένων του, καθώς και του τρόπου με τον οποίο θα γίνεται η παρουσίαση των πληροφοριών. Αυτού του είδους τα πρωτότυπα είναι ιδιαίτερα χρήσιμα για τον εντοπισμό των απαιτήσεων των χρηστών. Στον αντίποδα του προηγούμενου είδους βρίσκονται τα λειτουργικά πρωτότυπα. Τα λειτουργικά πρωτότυπα υλοποιούν πραγματικές λειτουργίες του συστήματος. Οι χρήστες πραγματοποιούν αναλύσεις με τη βοήθεια του συστήματος και διατυπώνουν τα σχόλια τους. Τα λειτουργικά πρωτότυπα μπορούν να εξελιχθούν σχετικά εύκολα σε τελικά συστήματα. Άλλα είδη πρωτοτύπων είναι το πρωτότυπο επίδειξης, το οποίο χρησιμοποιείται για να επιδειχθούν στους πελάτες οι δυνατότητες του συστήματος, και το διερευνητικό πρωτότυπο, το οποίο χρησιμοποιείται για να διερευνηθούν πιθανοί κίνδυνοι και να αποφασιστεί ή να απορριφθεί η υλοποίηση του συστήματος.

Άλλα θέματα που πρέπει να αντιμετωπιστούν κατά τη δημιουργία των πρωτοτύπων είναι:

- Οι άνθρωποι που θα συμμετέχουν στον σχεδιασμό, στη χρήση και στην επικύρωση του πρωτοτύπου.
- Τα δεδομένα που θα χρησιμοποιηθούν.
- Το υλικό και λογισμικό υπολογιστών που θα απαιτηθεί.

Η ανάλυση απαιτήσεων του συστήματος πρέπει να συμπεριλάβει και ζητήματα μεταδεδομένων. Κάθε μηχανογραφικό σύστημα παράγει και διαχειρίζεται δεδομένα. Ο όρος «μεταδεδομένα» σημαίνει άλλα, πρόσθετα δεδομένα, που περιγράφουν τα δεδομένα του μηχανογραφικού συστήματος. Τα μεταδεδομένα μπορούν να αφορούν τεχνικά, αλλά και επιχειρησιακά, μη τεχνικά θέματα. Τα επιχειρησιακά μεταδεδομένα διευκολύνουν τα διοικητικά στελέχη, τα οποία κατά κανόνα δεν έχουν μεγάλη τεχνική κατάρτιση, αλλά αντιλαμβάνονται πολύ καλά τα επιχειρησιακά θέματα, να κατανοήσουν το σύστημα, να πλοηγηθούν σε αυτό και να το χρησιμοποιήσουν αποτελεσματικά. Το υποσύστημα μεταδεδομένων μπορεί να θεωρηθεί ως ένα σημασιολογικό στρώμα του συστήματος ΕΕ, το οποίο αποδίδει νόημα στα δεδομένα.

Αναλυτικότερα, οι πληροφορίες που τηρούνται μέσω των μεταδεδομένων μπορούν να αναφέρονται σε θέματα όπως:

- Τα προβλήματα του οργανισμού που καλείται να επιλύσει το σύστημα ΕΕ.
- Το περιεχόμενο και τη σημασία της πληροφορίας που παράγεται.
- Τις εκθέσεις που θα παραχθούν μέσω του συστήματος και τα περιεχόμενα τους.
- Τη δομή των δεδομένων του συστήματος. Τα διαγράμματα τύπου οντότητας-σχέσης, που προέρχονται από τη θεωρία των σχεσιακών βάσεων δεδομένων, είναι πολύ αποτελεσματικά γι' αυτήν την εργασία.
- Τον καθορισμό του περιεχομένου των δεδομένων, όπως πχ τον ορισμό υπολογισμού αριθμοδεικτών.
- Λεπτομέρειες και οδηγίες χρήσης των αναλυτικών μεθόδων.
- Πληροφορίες σχετικά με τα πηγαία δεδομένα, όπως πχ η δομή τους, τα πηγαία συστήματα κλπ.
- Τις πολιτικές καθαρισμού, μετασχηματισμού και φόρτωσης των πηγαίων δεδομένων.

- Πληροφορίες σχετικά με τις εργασίες ETL που πραγματοποιήθηκαν, όπως πότε έγιναν, από ποιους, τι πηγαία δεδομένα χρησιμοποιήθηκαν, τι μετασχηματισμούς υπέστησαν κλπ.
- Πληροφορίες σχετικά με την πρόσβαση στο σύστημα, όπως ποιος χρησιμοποίησε το σύστημα, τι εργασίες εκτέλεσε κλπ.
- Ζητήματα δικαιωμάτων πρόσβασης στο σύστημα.

Η σημασία των μεταδεδομένων σε έναν οργανισμό είναι μεγάλη. Επιτρέπουν τη σωστή ερμηνεία των δεδομένων και την παραγωγή χρήσιμης πληροφορίας, επιβάλλουν την τυποποίηση δεδομένων και λειτουργιών, αποκλείοντας αυθαίρετες παρεμβάσεις και ερμηνείες από διαφορετικούς χρήστες. Επίσης, ορίζουν μέτρα ποιότητας των δεδομένων και τέλος καθιστούν ευκολότερη και αποτελεσματικότερη τη χρήση και τη συντήρηση του συστήματος. Ο οργανισμός μπορεί να προμηθευτεί έτοιμο λογισμικό για την τήρηση των μεταδεδομένων ή να κατασκευάσει λογισμικό σύμφωνα με τις δικές του απαιτήσεις. Το λογισμικό αυτό πρέπει να μπορεί να επικοινωνεί και να δέχεται είσοδο και από άλλα λογισμικά, όπως λογισμικό ETL, λογισμικό OLAP, λογισμικό εξόρυξης δεδομένων κλπ. Επίσης, είναι απαραίτητη η απασχόληση προσωπικού για την τήρηση των μεταδεδομένων.

12.2.4 Σχεδιασμός

Στο στάδιο αυτό πραγματοποιείται ο σχεδιασμός του νέου πληροφοριακού συστήματος. Όπως αναφέρθηκε και προηγουμένως, τα συστήματα Επιχειρηματικής Ευφυΐας είναι εντάσεως δεδομένων, δηλαδή σχετίζονται σε μεγάλο βαθμό με τη διαχείριση των δεδομένων. Για τον λόγο αυτό, στη καρδιά κάθε συστήματος ΕΕ βρίσκεται η Αποθήκη Δεδομένων (ΑΔ). Ο σχεδιασμός του συστήματος αφορά καταρχήν τον σχεδιασμό της Αποθήκης Δεδομένων.

Στις Αποθήκες Δεδομένων ισχύουν διαφορετικές σχεδιαστικές αρχές από τις κλασικές Σχεσιακές Βάσεις Δεδομένων (ΣΒΔ). Καταρχήν, στις ΣΒΔ βασικό μέλημα είναι η αποφυγή του πλεονασμού των δεδομένων (data redundancy). Σύμφωνα με αυτήν την αρχή, απαγορεύεται η αποθήκευση των ίδιων δεδομένων πολλές φορές, γιατί δημιουργούνται προβλήματα ασυνέπειας των δεδομένων. Η διαδικασία της κανονικοποίησης στοχεύει στην εξάλειψη του πλεονασμού. Αντιθέτως, στις ΑΔ η ανάγκη ταχείας πρόσβασης σε μεγάλους όγκους δεδομένων επιβάλλει την παραβίαση των κανονικών μορφών και την πολλαπλή αποθήκευση των ίδιων δεδομένων. Οι διαφορές μεταξύ των δύο συστημάτων δεν περιορίζονται στο ζήτημα του πλεονασμού. Οι ΑΔ τηρούν μεγάλους όγκους ιστορικής πληροφορίας, κάτι που δεν ισχύει στις ΣΒΔ. Επίσης, στις ΣΒΔ υπολογιζόμενα πεδία (πχ αθροίσματα, μέσοι όροι κλπ.) δεν αποθηκεύονται, αλλά υπολογίζονται κατά την εκτέλεση του ερωτήματος. Αντιθέτως, στις ΑΔ γίνεται μαζική αποθήκευση υπολογιζόμενων δεδομένων.

Οι ΑΔ δεν είναι συστήματα εντατικής και καθημερινής καταχώρησης δεδομένων. Τα δεδομένα εισέρχονται σε τακτά χρονικά διαστήματα με τις εργασίες ETL. Αντιθέτως, οι ΑΔ είναι συστήματα καθημερινής ανάκτησης δεδομένων. Εξαιτίας αυτού, η έμφαση δίνεται στις απαιτήσεις για ταχεία ανάκτηση. Επίσης, η πληροφορία πρέπει να είναι δομημένη με τρόπο κατανοητό και χρήσιμο σε στελέχη επιχειρήσεων και όχι σε ειδικούς πληροφορικής. Ένας επιπλέον λόγος γι αυτό, είναι ότι οι χρήστες σε πολλές περιπτώσεις (πχ πράξεις OLAP) χρησιμοποιούν απευθείας την ΑΔ, χωρίς τη μεσολάβηση κάποιας εφαρμογής, η οποία αποκρύπτει τη σχεδιαστική πολυπλοκότητα.

Οι ειδικές ανάγκες και συνθήκες που ισχύουν για τα συστήματα ΕΕ καθιστούν το κλασικό κανονικοποιημένο λογικό σχήμα των ΣΒΔ ακατάλληλο για τις ΑΔ. Στις ΑΔ οι λογικές πληροφορίες διαχωρίζονται σε γεγονότα (facts) και διαστάσεις (dimensions). Τα γεγονότα είναι συνήθως αριθμητικές ποσότητες συναλλαγών (πχ το ύψος των πωλήσεων). Οι διαστάσεις είναι πληροφορίες που συνδέονται με τα γεγονότα (πχ κατηγορία προϊόντος και γεωγραφική περιοχή). Αποθηκεύονται συνολικοποιημένα δεδομένα (πχ συνολικές πωλήσεις) σε σχέση με τις διαστάσεις (πχ πωλήσεις ανά κατηγορία προϊόντος και γεωγραφική περιοχή). Το λογικό σχήμα, το οποίο χρησιμοποιείται στις ΑΔ και που είναι κατάλληλο για τέτοιου τύπου μοντελοποίηση των δεδομένων, είναι το σχήμα του αστέρα (star schema) ή μια παραλλαγή του που ονομάζεται σχήμα χιονονιφάδας (snowflake schema). Ο αναγνώστης μπορεί να αναζητήσει περισσότερες λεπτομέρειες για τα ζητήματα αυτά στο Κεφάλαιο 4, το οποίο αναφέρεται στις Αποθήκες Δεδομένων.

Κατά τον λογικό σχεδιασμό της ΑΔ πρέπει να γίνει πολύ προσεκτικά ο καθορισμός των γεγονότων και των διαστάσεων. Τα γεγονότα και οι διαστάσεις συνδέονται άμεσα με τα επιχειρηματικά ζητήματα που καλείται να εξυπηρετήσει το σύστημα. Οι παράμετροι αυτών των ζητημάτων θα αποτελέσουν τις διαστάσεις. Η πληροφορία πρέπει να είναι δομημένη με τέτοιον τρόπο, ώστε τα επιχειρηματικά στελέχη να μπορούν να υποβάλλουν τα ερωτήματα που τους ενδιαφέρουν και να λαμβάνουν σαφείς και ορθές απαντήσεις. Σημαντικό είναι ότι το

σύστημα θα χρησιμοποιηθεί από διαφορετικά τμήματα του οργανισμού, τα οποία έχουν διαφορετικές ανάγκες πληροφόρησης, Για παράδειγμα, το τμήμα παραγωγής λογικά δεν ενδιαφέρεται για τη γεωγραφική διασπορά των πωλήσεων. Αντιθέτως, για το τμήμα μάρκετινγκ η πληροφορία αυτή μπορεί να είναι κρίσιμης σημασίας. Ο σχεδιασμός της ΑΔ πρέπει να είναι τέτοιος, ώστε να εξυπηρετούνται οι διαφορετικές ανάγκες των διαφορετικών τμημάτων του οργανισμού.

Ένα άλλο ζήτημα μεγάλης σημασίας είναι ο καθορισμός του βαθμού της λεπτομέρειας της πληροφορίας που θα τηρείται στο σύστημα. Υπερβολική λεπτομέρεια θα οδηγήσει στην υπερμεγέθυνση των δεδομένων και συνακόλουθα σε καθυστερήσεις στη λειτουργία του συστήματος, αυξημένες απαιτήσεις για αποθηκευτικό χώρο και άλλα προβλήματα. Από την άλλη πλευρά, η υπερβολική γενίκευση θέτει όρια στις δυνατότητες ανάλυσης. Είναι δυνατόν τα στελέχη που πραγματοποιούν τις αναλύσεις να διαπιστώσουν κάποια στιγμή ότι χρειάζονται αναλυτικότερα δεδομένα από αυτά που ορίστηκαν κατά τον σχεδιασμό του συστήματος. Για τους λόγους αυτούς, πρέπει να επιλεγεί προσεκτικά το ισοζύγιο μεταξύ λεπτομέρειας και γενίκευσης. Το Σχήμα 12.3 δείχνει δύο μη ικανοποιημένους χρήστες, με αντικρουόμενες ανάγκες πληροφόρησης.

Παρόλο που ο προτεινόμενος λογικός σχεδιασμός μιας ΑΔ ακολουθεί το σχήμα του αστέρα, πρέπει να σημειώσουμε ότι αυτό δεν έχει απόλυτη και καθολική ισχύ. Υπάρχουν περιπτώσεις όπου οι εργασίες ανάλυσης των δεδομένων που πραγματοποιούνται στον οργανισμό απαιτούν πρόσβαση σε αναλυτικά δεδομένα, ακόμα και αν αυτό συνεπάγεται μεγάλες καθυστερήσεις στην εκτέλεση των ερωτημάτων. Σε αυτές τις περιπτώσεις μπορεί να επιλεγεί το κανονικοποιημένο σχεσιακό σχήμα.

Πέραν των ζητημάτων που σχετίζονται με τον λογικό σχεδιασμό της Αποθήκης Δεδομένων, πρέπει να ληφθούν υπόψη και διάφορα τεχνικά και άλλα θέματα. Είναι σύνηθες φαινόμενο οι ΑΔ να περιέχουν πολύ μεγάλους όγκους δεδομένων, με τάξη μεγέθους terabyte. Αυτές οι βάσεις δεδομένων καλούνται Πολύ Μεγάλες Βάσεις Δεδομένων (Very Large Data Bases (VLDB)) και έχουν ειδικές τεχνικές απαιτήσεις. Τα ειδικά τεχνικά θέματα των Πολύ Μεγάλων Βάσεων Δεδομένων βρίσκονται έξω από τα όρια του παρόντος συγγράμματος. Ο αναγνώστης μπορεί να αναζητήσει λεπτομέρειες στη σχετική βιβλιογραφία. Κατά τον σχεδιασμό της Αποθήκης Δεδομένων πρέπει να αποφασιστούν και πολιτικές ασφαλείας, όπως ποιος θα έχει πρόσβαση στον σχεδιασμό της ΑΔ, ποιος στη χρήση της, τι πληροφορία θα είναι διαθέσιμη στον εκάστοτε χρήστη κλπ.

Οι Αποθήκες Δεδομένων δεν παράγουν πρωτογενώς δεδομένα μέσω της καθημερινής χρήσης τους, αλλά χρησιμοποιούν δεδομένα άλλων πληροφοριακών συστημάτων, τα οποία μεταφέρονται στην ΑΔ. Τα πηγαία συστήματα μπορεί να είναι διαφορετικά, διασκορπισμένα ή ασύμβατα. Επίσης, τα δεδομένα τους μπορεί να είναι αντικρουόμενα, ελλιπή ή εσφαλμένα. Για τους λόγους αυτούς, η διαδικασία Εξαγωγής, Μετασχηματισμού, και Φόρτωσης (ΕΜΦ) (Extract, Transform, Load (ETL)) των δεδομένων δεν είναι καθόλου απλή, καθώς στα πλαίσια της πρέπει να επιλυθούν μια σειρά από προβλήματα. Ο σχεδιασμός του συστήματος περιλαμβάνει ως απαραίτητο τμήμα του τον σχεδιασμό των εργασιών ΕΜΦ.

Σχήμα 12.3 *Αντικρουόμενες ανάγκες πληροφόρησης*

Εργασίες ΕΜΦ εκτελούνται αρχικά κατά τη δημιουργία της ΑΔ. Σε αυτό το στάδιο μεταφέρονται τα τρέχοντα λειτουργικά δεδομένα. Επίσης μεταφέρονται τα ιστορικά, παλαιότερα δηλαδή, δεδομένα. Δεν είναι

σπάνιο να υπάρχουν διαφορές ανάμεσα στα ιστορικά και τα τρέχοντα λειτουργικά δεδομένα, σε ζητήματα όπως ονοματοδοσίες, μορφές δεδομένων κλπ. Οι διαφορές αυτές πρέπει να διευθετηθούν. Ακολούθως, κατά τη διάρκεια ζωής της ΑΔ μεταφέρονται σε τακτά χρονικά διαστήματα τα καινούρια δεδομένα. Ο όρος καινούρια δεδομένα δεν περιγράφει μόνον πρόσθετα δεδομένα που παρήχθησαν κατά το χρονικό διάστημα από την προηγούμενη μεταφορά, αλλά και τροποποιήσεις παλαιότερων δεδομένων. Κατά συνέπεια, τίθεται ζήτημα κατάλληλων ενημερώσεων των δεδομένων.

Κατά την εκτέλεση των εργασιών ΕΜΦ πρέπει αρχικά να επιλεγούν τα δεδομένα που θα μεταφερθούν. Ακολούθως πρέπει να επιλυθούν μια σειρά από προβλήματα. Ένα συνηθισμένο πρόβλημα είναι η ύπαρξη πολλαπλών πρωτευόντων κλειδιών. Οι πολλοί διαφορετικοί κωδικοί πελάτη είναι ένα τυπικό παράδειγμα. Πρωτεύοντα κλειδιά, αλλά και άλλα πεδία, μπορεί να περιέχουν την ίδια πληροφορία αλλά να έχουν διαφορετικό τίτλο. Άλλο πρόβλημα είναι η ύπαρξη ασυνεπειών, η ύπαρξη δηλαδή διαφορετικών αντικρουόμενων τιμών δεδομένων για την ίδια πληροφορία. Διαφορές μπορεί να υπάρχουν και στη μορφή των δεδομένων ή στις μονάδες μέτρησης. Επίσης, μπορεί να υπάρχουν λανθασμένες τιμές. Το λογισμικό ΕΜΦ πρέπει να είναι ικανό να επιλύει αυτά τα προβλήματα.

Κατά κανόνα, τα δεδομένα δεν μεταφέρονται αυτούσια στην ΑΔ, αλλά μετασχηματίζονται. Ο βασικότερος μετασχηματισμός είναι η συνολικοποίηση τους, σύμφωνα με τα γεγονότα και τις διαστάσεις που ορίζονται στο σχήμα της ΑΔ. Άλλοι μετασχηματισμοί μπορεί να επιβάλλονται από τις ανάγκες των αναλυτικών μεθόδων. Για παράδειγμα, μπορεί να γίνει μετατροπή αριθμητικών τιμών σε ονομαστικές ή σε άλλες αριθμητικές τιμές.

Η διαδικασία εξαγωγής, μετασχηματισμού και φόρτωσης πρέπει να σχεδιαστεί προσεκτικά, ώστε να δίνει απαντήσεις σε όλα τα προηγούμενα ζητήματα. Η σχετική διαχείριση των δεδομένων μπορεί να καταγραφεί σε έναν πίνακα, που να ορίζει τα πηγαία δεδομένα, τους μετασχηματισμούς και τους τελικούς προορισμούς αποθήκευσης. Επίσης, η όλη διαδικασία μπορεί να αποτυπωθεί σε ένα διάγραμμα ροής εργασιών, που να ορίζει τα βήματα, την αλληλουχία τους και τις σχέσεις αλληλεξάρτησης τους. Στην αγορά διατίθενται έτοιμα εργαλεία ΕΜΦ. Τα εργαλεία αυτά αξιολογούνται, ώστε να επιλεγεί το πλέον κατάλληλο. Εναλλακτικά μπορεί να γραφεί κώδικας, που να εκτελεί τις εργασίες αυτές.

Ο σχεδιασμός του συστήματος περιλαμβάνει επίσης και τον σχεδιασμό του απαραίτητου λογισμικού, που θα χρησιμοποιηθεί για την ανάλυση των δεδομένων. Η σύγχρονη επιχειρηματική ευφυΐα δεν διεξάγεται μόνον με την οπτική διερεύνηση των δεδομένων και την εκτέλεση πράξεων OLAP. Η Εξόρυξη Δεδομένων ήρθε να εμπλουτίσει δραστικά το φάσμα των διαθέσιμων αναλυτικών τεχνικών και να προσφέρει νέες πρωτόγνωρες δυνατότητες. Εξελιγμένα λογισμικά, που κάνουν χρήση προχωρημένων στατιστικών μεθόδων, καθώς και μεθόδων τεχνητής νοημοσύνης, είναι διαθέσιμα και μπορούν να παράξουν πληροφόρηση, πολύτιμη για τη λήψη επιχειρηματικών αποφάσεων.

Η Εξόρυξη Δεδομένων προσφέρει μεγάλη ποικιλία μεθόδων, που καλύπτουν διαφορετικά αντικείμενα και έχουν διαφορετικές φιλοσοφίες, λειτουργίες, αποτελέσματα και απαιτήσεις ως προς τα δεδομένα. Οι μέθοδοι αυτές είναι κατάλληλες για κατηγοριοποίηση και πρόβλεψη αριθμητικών τιμών, ανάλυση συστάδων, εξόρυξη κανόνων συσχέτισης, ανάλυση χρονοσειρών, εντοπισμό εξαιρέσεων κλπ. Για κάθε μια από αυτές τις κατηγορίες εργασιών διατίθεται πληθώρα εναλλακτικών μεθόδων, με ειδικά πλεονεκτήματα και μειονεκτήματα. Αν για παράδειγμα οι χρήστες πρέπει να εκτελέσουν εργασίες κατηγοριοποίησης, μπορούν να χρησιμοποιήσουν μεθόδους δένδρων αποφάσεων, νευρωνικών δικτύων, Μπαϋεσιανών δικτύων, μηχανών διανυσμάτων υποστήριξης, k-πλησιέστερων γειτόνων και υβριδικών κατηγοριοποιητών. Κάθε μια από αυτές τις μεθόδους έχει πλεονεκτήματα και μειονεκτήματα ως προς τις επιδόσεις κατηγοριοποίησης, την ερμηνευσιμότητα των μοντέλων, τη δυνατότητα επεξεργασίας μεγάλου όγκου δεδομένων κλπ.

Οι σχεδιαστές του συστήματος πρέπει να είναι καλοί γνώστες των δυνατοτήτων και απαιτήσεων αυτών των μεθόδων. Επίσης, πρέπει να μπορούν να κατανοούν σε βάθος τις ανάγκες των χρηστών και να διαβλέπουν πόσες και ποιες από αυτές τις μεθόδους αποτελούν την ιδανική επιλογή για την επίλυση των προβλημάτων. Δεδομένου ότι οι μέθοδοι αυτές είναι εξαιρετικά σύγχρονες και ότι η εφαρμογή τους στα συστήματα Επιχειρηματικής Ευφυΐας βρίσκεται στα αρχικά στάδια, οι σχεδιαστές των συστημάτων μπορούν να πρωτοπορήσουν, προσφέροντας εργαλεία που επεκτείνουν τις πάγιες πρακτικές των χρηστών και ανοίγουν νέους δρόμους στην ανάκτηση επιχειρηματικής γνώσης από τα δεδομένα. Οι πάροχοι λογισμικού ΕΕ φροντίζουν να παρέχουν δυνατότητες επέκτασης στα προϊόντα τους. Είναι χαρακτηριστικό ότι τόσο η Oracle όσο και η Microsoft ενσωματώνουν τη γλώσσα R στα συστήματα τους.

Ο σχεδιασμός του συστήματος ΕΕ ολοκληρώνεται με τον σχεδιασμό του συστήματος μεταδεδομένων. Αρχική πηγή μεταδεδομένων είναι οι βάσεις δεδομένων, στις οποίες ο καθορισμός των μεταδεδομένων αποτελεί αναπόσπαστο τμήμα της λειτουργίας τους. Στις βάσεις δεδομένων πρέπει να οριστούν οι πίνακες, τα πρωτεύοντα κλειδιά, τα ξένα κλειδιά, κανόνες ακεραιότητας αναφοράς κλπ. Όμως, τεχνικά μεταδεδομένα πα-

ράγονται και από άλλα λογισμικά, όπως εργαλεία ΕΜΦ, εργαλεία καθαρισμού δεδομένων, λογισμικά OLAP κλπ. Επίσης, χρειάζονται λογικά μεταδεδομένα σχετικά με ερμηνείες επιχειρηματικών πληροφοριών που παράγονται από το σύστημα, τις εκθέσεις που συντάσσονται κλπ. Γενικώς τα συστήματα ΕΕ χρειάζονται πολλά μεταδεδομένα. Για τον λόγο αυτό, η καθιέρωση και χρήση των συστημάτων ΕΕ συνέβαλε καθοριστικά στην αναγνώριση της σημασίας των μεταδεδομένων. Η πληθώρα των μεταδεδομένων επιβάλλει την ύπαρξη κατάλληλου λογισμικού για τη διαχείριση τους.

Στην αγορά λογισμικού διατίθενται έτοιμα προϊόντα για την τήρηση των μεταδεδομένων. Τα εργαλεία αυτά πρέπει να αξιολογηθούν ως προς την ικανότητα τους να καλύπτουν τις ανάγκες του οργανισμού. Εάν δεν βρεθεί κατάλληλο έτοιμο λογισμικό, τότε το λογισμικό αυτό πρέπει να κατασκευαστεί. Κάθε μια από αυτές τις λύσεις έχει πλεονεκτήματα και μειονεκτήματα. Το έτοιμο λογισμικό είναι φθηνότερο και άμεσα εφαρμόσιμο, δύσκολα όμως θα ικανοποιήσει πλήρως όλες της ανάγκες. Με την κατασκευή λογισμικού καλύπτονται όλες οι απαιτήσεις, η λύση όμως αυτή συνεπάγεται επιβαρύνσεις σε κόστος και χρόνο. Σημαντικά ζητήματα, που σχετίζονται με το λογισμικό τήρησης μεταδεδομένων, είναι η δυνατότητα του να ενημερώνεται αυτόματα από άλλα λογισμικά, όπως βάσεις δεδομένων και εργαλεία ΕΜΦ, οι τρόποι και η διαβάθμιση της πρόσβασης στα μεταδεδομένα, καθώς και το εάν το σύστημα θα είναι κεντρικό ή κατανεμημένο.

12.2.5 Υλοποίηση

Κατά το στάδιο αυτό πραγματοποιείται η υλοποίηση του συστήματος. Ένα σημαντικό μερίδιο των εργασιών που εκτελεί το σύστημα είναι η τροφοδότηση της ΑΔ με δεδομένα. Η υλοποίηση των εργασιών ΕΜΦ είναι ένα επίπονο καθήκον. Εάν έχει απορριφθεί η λύση της αγοράς εργαλείων ΕΜΦ, τότε κατασκευάζεται το αναγκαίο λογισμικό. Ακολούθως, εκτελούνται οι αρχικές εργασίες πλήρωσης της ΑΔ με δεδομένα. Τα δεδομένα που παραβιάζουν καθορισμένους επιχειρηματικούς κανόνες καθαρίζονται και αποκαθίστανται οι ορθές τιμές. Υπολογίζονται τα αθροίσματα που προβλέπονται από το σχήμα της ΑΔ. Είναι πιθανόν να κατασκευαστούν νέα δεδομένα με συνδυασμό άλλων δεδομένων που προϋπήρχαν. Αντιμετωπίζονται τα προβλήματα των διαφορετικών ονομάτων και τιμών. Το ζητούμενο είναι να υπάρχει για ένα δεδομένο μια μοναδική τιμή, που να χαρακτηρίζεται από ένα μοναδικό όνομα.

Η μεταφορά των δεδομένων ακολουθείται από ελέγχους ορθότητας. Ελέγχονται οι διαδικασίες που υλοποιούνται στα λογισμικά, η ροή εκτέλεσης των εργασιών, οι αλληλεπιδράσεις μεταξύ διαφορετικών λογισμικών και τα τελικά δεδομένα. Στη διαδικασία συμμετέχουν όχι μόνο τεχνικά στελέχη, αλλά και στελέχη εξειδικευμένα στα επιχειρηματικά θέματα. Επειδή οι εργασίες ΕΜΦ είναι πολύ χρονοβόρες, είναι σημαντικό να ελεγχθούν οι επιδόσεις του συστήματος. Η όλη διαδικασία ελέγχων είναι αρκετά περίπλοκη, και για τον λόγο αυτό είναι χρήσιμη η κατάρτιση ενός σχεδίου που να οργανώνει αυτές τις ενέργειες.

Το λογισμικό που χρησιμοποιείται για την ανάλυση των δεδομένων μπορεί να παρουσιάζει μεγάλη ποικιλία ως προς τη φιλοσοφία του και τις δυνατότητες του. Στην απλούστερη εκδοχή του περιλαμβάνει τη χρήση μοντέλων για τη διεξαγωγή αναλύσεων what-if, αναλύσεων ευαισθησίας και αναλύσεων αναζήτησης στόχων. Επίσης, τα συστήματα ΕΕ περιέχουν κατά κανόνα λογισμικό για την εκτέλεση εργασιών OLAP. Τέλος, για τη διεξαγωγή πιο περίτεχνων αναλύσεων απαιτείται λογισμικό Εξόρυξης Δεδομένων. Πέραν του λογισμικού ανάλυσης, το σύστημα θα περιλαμβάνει και λογισμικό για την οπτικοποίηση των αποτελεσμάτων και τη σύνταξη αναφορών.

Οι πάροχοι συστημάτων ΕΕ προσφέρουν μια μεγάλη ποικιλία προϊόντων, που καλύπτουν σχεδόν κάθε είδος αναλυτικών καθηκόντων. Πολλά από αυτά τα λογισμικά είναι ειδικά σχεδιασμένα για την εκτέλεση συγκεκριμένων εργασιών, όπως η διαχείριση επίδοσης της επιχείρησης, η διαχείριση του ρίσκου, η αντιμετώπιση της απάτης, η διαχείριση της εφοδιαστικής αλυσίδας κλπ. Τα προϊόντα αυτά πρέπει να αξιολογηθούν προσεκτικά, ώστε ο οργανισμός να προμηθευτεί το κατάλληλο λογισμικό. Σε αρκετές περιπτώσεις είναι δυνατός ο συνδυασμός διαφορετικών λύσεων που προέρχονται από διαφορετικούς κατασκευαστές. Για παράδειγμα, είναι συνηθισμένη η χρήση προϊόντων της Microsoft, όπως το EXCEL και το Share Point, ως πλατφορμών εξαγωγής και οπτικοποίησης των αποτελεσμάτων.

Σε περίπτωση που έχει επιλεγεί η λύση της κατασκευής λογισμικού, στο στάδιο αυτό πραγματοποιείται η υλοποίηση του κώδικα. Εάν κατά το στάδιο της ανάλυσης έχει κατασκευαστεί κάποιο πρωτότυπο, τότε στο στάδιο αυτό το πρωτότυπο θα εξελιχθεί σε ολοκληρωμένο σύστημα. Ειδικά, τα λειτουργικά πρωτότυπα, τα οποία ενσωματώνουν ήδη πραγματικές αναλυτικές και άλλες λειτουργίες, μπορούν να εξελιχθούν σχετικά εύκολα σε τελικά συστήματα. Τόσο το αγορασμένο όσο και το κατασκευασμένο λογισμικό υποβάλλονται σε διαδικασίες σχολαστικών και επαναλαμβανόμενων ελέγχων.

Ένα άλλο ζήτημα, που πρέπει να αντιμετωπιστεί στα πλαίσια υλοποίησης, είναι αυτό της πρόσβασης στο

σύστημα. Ένας δημοφιλής τρόπος πρόσβασης είναι μέσω του διαδικτύου. Η διαδικτυακή διεπαφή με το σύστημα, πέραν της ευχρηστίας που προσφέρει, εξασφαλίζει τη δυνατότητα της εύκολης απομακρυσμένης πρόσβασης, και μάλιστα με χρήση κινητών συσκευών. Σε περίπτωση που έχει επιλεγεί μια τέτοια λύση, στο στάδιο αυτό γίνεται και η κατασκευή της πύλης ΕΕ. Ιδιαίτερη μέριμνα πρέπει να ληφθεί για την ελεγχόμενη και διαβαθμισμένη πρόσβαση στο σύστημα.

Τέλος, γίνεται η υλοποίηση του συστήματος μεταδεδομένων. Ο οργανισμός μπορεί να αγοράσει κάποιο έτοιμο εργαλείο τήρησης μεταδεδομένων ή να κατασκευάσει το δικό του λογισμικό. Τα μεταδεδομένα που θα εισαχθούν στο σύστημα μεταδεδομένων προέρχονται κυρίως από άλλα λογισμικά. Τέτοια λογισμικά είναι επεξεργαστές κειμένου, όπου έχουν καταγραφεί οδηγίες για τον σκοπό και τη λειτουργία του συστήματος ΕΕ, βιβλία εργασίας, όπου οι αναλυτές του οργανισμού έχουν εκτελέσει υπολογισμούς, συστήματα βάσεων δεδομένων που τηρούν τα δικά τους μεταδεδομένα σχετικά με πίνακες, κλειδιά, πεδία τιμών κλπ. εργαλεία ΕΜΦ που καταγράφουν πληροφορίες, όπως αλγορίθμους μετασχηματισμού δεδομένων, ιστορικά μεταφοράς δεδομένων κλπ. εργαλεία OLAP, όπου τηρούνται πληροφορίες όπως υπολογισμοί και αθροίσματα δεδομένων, και εργαλεία εξόρυξης δεδομένων με πληροφορίες σχετικά με τις μεθόδους, τις απαιτήσεις και τις δυνατότητες τους. Το σύστημα μεταδεδομένων πρέπει να μπορεί να επικοινωνεί με όλα αυτά τα συστήματα που το τροφοδοτούν με μεταδεδομένα. Τα έτοιμα συστήματα μεταδεδομένων διαθέτουν τέτοιες δυνατότητες, οι οποίες όμως μπορεί να μην είναι αρκετές.

Το σύστημα μεταδεδομένων πρέπει να τηρείται ενημερωμένο με σχολαστικότητα και να περιγράφει με ακρίβεια τα δεδομένα του συστήματος ΕΕ. Κάθε αλλαγή στα δεδομένα πρέπει να καταγράφεται στο σύστημα μεταδεδομένων. Χρήσιμο είναι να οριστεί κάποιος υπεύθυνος για την τήρηση του συστήματος μεταδεδομένων. Εκτός του διαχειριστή του συστήματος, το σύστημα μεταδεδομένων χρησιμοποιούν ειδικοί πληροφορικής, όπως διαχειριστές βάσεων δεδομένων, χειριστές συστημάτων ΕΜΦ, αλλά και τα στελέχη της επιχείρησης για να αντλήσουν πληροφόρηση. Αυτές οι κατηγορίες χρηστών έχουν διαφορετικές ανάγκες, και καλό είναι να προσφέρονται διαφορετικές διεπαφές πρόσβασης, που να τους διευκολύνουν. Επίσης, τα περιεχόμενα του συστήματος μεταδεδομένων μπορεί να οργανωθούν σε θεματικές ενότητες ανάλογα με τα ενδιαφέροντα αυτών των κατηγοριών χρηστών. Τέλος, πρέπει να προβλεφθούν μέτρα ασφάλειας για την πρόσβαση στα μεταδεδομένα.

12.2.6 Εφαρμογή

Κατά το στάδιο της εφαρμογής το σύστημα τίθεται σε λειτουργία. Καταρχάς γίνεται η εγκατάσταση του συστήματος. Η εργασία αυτή συνοδεύεται από μια σειρά τεχνικές ρυθμίσεις. Το σύστημα μπορεί να παραδοθεί ολόκληρο, αλλά προτιμότερο είναι να παραδίδεται τμηματικά, σε επαναλαμβανόμενα στάδια, κατά τη διάρκεια της εξέλιξης του. Η σταδιακή παράδοση περιορίζει τους κινδύνους, προσφέρει νωρίτερα χειροπιαστά αποτελέσματα στους χρήστες και διευκολύνει τη μεταφορά γνώσης, καθώς και τη διαχείριση της αλλαγής που επιφέρει η εγκατάσταση του συστήματος στον οργανισμό.

Την εγκατάσταση του συστήματος ακολουθεί η εκπαίδευση των χρηστών. Η εργασία αυτή περιλαμβάνει κάποια σημεία που χρήζουν προσοχής. Συνηθισμένη πρακτική είναι η παρουσίαση του συστήματος, δηλαδή η παράθεση των λειτουργιών του. Με τον τρόπο αυτό, οι χρήστες μαθαίνουν πως λειτουργεί το σύστημα. Αυτό δεν σημαίνει και ότι έχουν εμπεδώσει το πώς μπορούν να το χρησιμοποιήσουν για να εκτελέσουν τις εργασίες τους αποτελεσματικότερα. Το φαινόμενο αυτό είναι εντονότερο σε συστήματα ΕΕ, τα οποία δεν μηχανοργανώνουν τυποποιημένες εργασίες, αλλά επιτρέπουν στους χρήστες μεγαλύτερη ευελιξία, δημιουργικότητα και φαντασία. Ο χρήστης μπορεί να εκτελέσει διάφορες αναλύσεις. Το ποιες από αυτές είναι οι πλέον κατάλληλες επαφίεται στη διακριτική του ευχέρεια. Μια πραγματική εκπαίδευση σε ένα σύστημα ΕΕ συνίσταται στη βαθύτερη κατανόηση των δυνατοτήτων του συστήματος, και κυρίως σύνδεση αυτών των δυνατοτήτων με τις αναλυτικές ανάγκες του χρήστη. Αυτός ο εκπαιδευτικός στόχος μπορεί να επιτευχθεί με τη χρήση του συστήματος σε πραγματικές συνθήκες, δηλαδή με πραγματικά δεδομένα και για την αντιμετώπιση πραγματικών προβλημάτων. Επιδείξεις του συστήματος με απλοποιημένα δεδομένα και υποθετικά σενάρια δεν είναι ιδιαίτερα αποτελεσματικές. Επιπλέον, η εκπαίδευση πρέπει να επικεντρώνεται κυρίως στα επιχειρησιακά ζητήματα και δευτερευόντως στα τεχνολογικά – χειριστικά. Για την εκπαίδευση των χρηστών πολύ χρήσιμη μπορεί να είναι και η συνδρομή των στελεχών που συμμετείχαν στον σχεδιασμό του συστήματος, γιατί τα στελέχη αυτά έχουν ήδη κατανοήσει ποια επιχειρηματικά ζητήματα εξυπηρετεί το σύστημα, καθώς και τον τρόπο που τα εξυπηρετεί.

Η εφαρμογή και χρήση του συστήματος απαιτεί και εργασίες για τη συντήρηση του, έτσι ώστε να διατηρείται διαρκώς ενημερωμένο και λειτουργικό. Σε συστήματα ΕΕ, όπου τα δεδομένα προέρχονται από άλλα πλη-

ροφοριακά συστήματα, η σημαντικότερη εργασία συντήρησης είναι η μεταφορά των δεδομένων. Οι εργασίες ΕΜΦ, που εκτελούνται σε τακτά χρονικά διαστήματα, τροφοδοτούν το σύστημα με δεδομένα. Κάθε φορά, μετά τις εργασίες ΕΜΦ πρέπει να γίνεται έλεγχος ορθότητας των δεδομένων και να ενημερώνεται προσεκτικά το σύστημα μεταδεδομένων.

Ένα άλλο ζήτημα που αφορά τη λειτουργικότητα του συστήματος είναι η διαχείριση της μεγέθυνσης του. Τα συστήματα ΕΕ έχουν την τάση να μεγεθύνονται, και μάλιστα με γρήγορο ρυθμό. Όπως προαναφέρθηκε, τα συστήματα ΕΕ εξελίσσονται διαρκώς μέσα από μια κυκλική διαδικασία. Οι νέες λειτουργικότητες που προστίθενται απαιτούν την ύπαρξη νέων δεδομένων ή την επέκταση των παλαιότερων. Το αποτέλεσμα είναι η συνεχής μεγέθυνση των δεδομένων. Η μεγέθυνση αφορά και τη χρήση του συστήματος, καθώς όλο και περισσότεροι χρήστες επιδιώκουν να αξιοποιούν τις δυνατότητες του. Η μεγέθυνση του συστήματος και τα συνακόλουθα τεχνικά προβλήματα που αυτή προκαλεί πρέπει να έχουν προβλεφθεί. Σημαντικότατο κριτήριο για την επιλογή του υλικού και του λογισμικού πρέπει να είναι η δυνατότητα κλιμάκωσης του και ανταπόκρισης του σε μελλοντικές αυξημένες απαιτήσεις. Κατά τη διάρκεια της λειτουργίας του συστήματος θα απαιτηθούν αναβαθμίσεις στο υλικό και το λογισμικό, κυρίως σε σέρβερς, εύρος ζώνης και συστήματα βάσεων δεδομένων. Τέλος, η συντήρηση του συστήματος περιλαμβάνει την τήρηση των αντιγράφων ασφαλείας και την παρακολούθηση της χρήση του.

12.2.7 Αξιολόγηση

Στο τελευταίο στάδιο του κύκλου ζωής ενός συστήματος Επιχειρηματικής Ευφυΐας γίνεται η αξιολόγηση του. Στην απλούστερη εκδοχή της, η αξιολόγηση γίνεται έναντι του προϋπολογισμού του έργου και του χρονοδιαγράμματος εκτέλεσης του. Ελέγχονται ζητήματα όπως εάν τηρήθηκαν οι προκαθορισμένοι χρόνοι για κάθε υποέργο, που προέκυψαν καθυστερήσεις, πώς επηρέασαν την εκτέλεση των άλλων υποέργων κλπ. Επίσης, σχετικά με τον προϋπολογισμό του έργου, απογράφεται το κόστος εκτέλεσης των υποέργων και προμήθειας εξοπλισμού, ελέγχονται και αιτιολογούνται πιθανές υπερβάσεις.

Συνθετότερη είναι η αποτίμηση της ανάλυσης κόστους – οφέλους. Το κόστος του έργου είναι σχετικά εύκολο να υπολογιστεί. Το όφελος όμως του οργανισμού δεν είναι πάντα τόσο προφανές. Εάν το έργο έχει σχεδιαστεί στη βάση αντιμετώπισης συγκεκριμένων επιχειρηματικών ζητημάτων με μετρήσιμα αποτελέσματα, πχ συμπίεση κόστους συγκεκριμένων δραστηριοτήτων, τότε το όποιο όφελος είναι καθορισμένο και χειροπιαστό. Η αποτίμηση αναφέρεται στο κατά πόσο εκπληρώθηκαν οι στόχοι και στο τι θετικές μεταβολές επήλθαν στα μετρήσιμα αποτελέσματα. Όμως τα έργα Επιχειρηματικής Ευφυΐας προσφέρουν και άλλα λιγότερο προφανή, μη χειροπιαστά, δύσκολα μετρήσιμα και μακροχρόνια οφέλη. Σε αυτά περιλαμβάνονται η απόκτηση νέας επιχειρηματικής γνώσης, η βελτίωση των διαδικασιών, κυρίως στη λήψη αποφάσεων, και οι πιο αποτελεσματικές σχέσεις. Η ποσοτικοποίηση του οφέλους σε τέτοια ζητήματα είναι δυσκολότερη ή ίσως και αδύνατη. Ο Strassmann (1990) τονίζει ότι άμεσα οφέλη, όπως η αύξηση των εσόδων, εμφανίζονται γρήγορα στις χρηματοοικονομικές καταστάσεις, ενώ άλλα έμμεσα οφέλη, όπως η μείωση του ρίσκου και η αύξηση της ανταγωνιστικότητας, είναι δυσκολότερο να διαπιστωθούν. Μελλοντικές ερευνητικές εργασίες πιθανόν να δώσουν απαντήσεις στο ζήτημα της ποσοτικοποίησης τέτοιων οφελών.

Αποτίμηση του έργου γίνεται με κάθε ολοκλήρωση του κύκλου ζωής του συστήματος, ανεξαρτήτως του εάν θεωρείται ότι λειτουργεί «τέλεια» ή παρουσιάζει προβλήματα. Η διαδικασία αυτή προσφέρει γνώση χρήσιμη για τη διαρκή βελτίωση του συστήματος. Πέρα από θέματα προϋπολογισμού και χρονοπρογραμματισμού, ελέγχονται ο βαθμός κάλυψης των απαιτήσεων του συστήματος, η καταλληλότητα της παραγόμενης πληροφορίας, η ικανοποίηση των χρηστών, η ποιότητα των δεδομένων, η αποτελεσματικότητα των αναλυτικών μεθόδων, ο βαθμός χρήσης του συστήματος, η ποιότητα των εκθέσεων και αναφορών και η ταχύτητα λειτουργίας του συστήματος.

Η αξιολόγηση του έργου γίνεται μέσα από μια προγραμματισμένη και καλά καθορισμένη διαδικασία συνεντεύξεων και ερωτηματολογίων. Η έρευνα αυτή γίνεται μετά την παρέλευση αρκετού χρόνου από την εγκατάσταση και ενεργοποίηση του συστήματος, ώστε ο χρήστες να έχουν αποκτήσει πραγματική και αρκετή εμπειρία από τη χρήση του. Στην έρευνα συμμετέχουν εκτός από τους τελικούς χρήστες και όλα τα στελέχη που συμμετείχαν στην ομάδα έργου. Οι ερωτήσεις πρέπει να είναι γνωστές εκ των προτέρων και οι συμμετέχοντες πρέπει να έχουν αρκετό χρόνο για να προετοιμαστούν. Χρήσιμο είναι να ανατεθεί σε κάποιο στέλεχος ο συντονισμός και η διεξαγωγή αυτής της έρευνας.

Οι όροι «αξιολόγηση» ή «αποτίμηση» είναι μάλλον ανεπαρκείς για να περιγράψουν πλήρως όλες τις εργασίες που εκτελούνται σε αυτό το στάδιο του κύκλου ζωής του συστήματος. Τα συστήματα Επιχειρηματικής Ευφυΐας είναι συστήματα σε διαρκή εξέλιξη. Οι ανάγκες για ανάκτηση επιχειρηματικής γνώσης δεν είναι

στατικές. Η πραγματική ζωή αναδεικνύει διαρκώς νέα ζητήματα, για τα οποία πρέπει να ληφθούν αποφάσεις. Οι Yeoh and Koronios (2010) τονίζουν ότι το σύστημα θα επιτύχει μόνον εάν οι χρήστες συνεχώς ανακαλύπτουν και μοντελοποιούν νέα γνώση και τροποποιούν τα δεδομένα. Το γεγονός ότι τα συστήματα ΕΕ έλκουν διαρκώς νέους χρήστες, καθώς και το ότι η χρήση του συστήματος διαπερνά οριζοντίως όλα τα τμήματα του οργανισμού σε όλη την έκταση του, συνιστούν δύο παράγοντες που αναδεικνύουν νέες ανάγκες πληροφόρησης ή τροποποιούν τις απαιτήσεις του συστήματος.

Πρέπει να γίνει κατανοητό ότι το σύστημα θα τροποποιείται διαρκώς και ότι πρέπει να υιοθετηθεί μια λογική διαδοχικών εκδόσεων. Είναι πρακτικά αδύνατο το σύστημα να λειτουργεί χωρίς προβλήματα και να καλύπτει όλες τις ανάγκες με την πρώτη του έκδοση. Η παραγωγή και διάθεση διαδοχικών εκδόσεων είναι μια πάγια, και εν πολλοίς ευπρόσδεκτη τακτική, όταν προέρχεται από τους παρόχους λογισμικού. Η αναβάθμιση όμως συστημάτων που έχουν αναπτυχθεί στα πλαίσια κάποιου οργανισμού συχνά γίνεται αντιληπτή ως ένα ενοχλητικό πρόβλημα, το οποίο επιβαρύνει τον οργανισμό με έξοδα και προκαλεί καθυστερήσεις στη λειτουργία του. Για την περίπτωση συστημάτων ΕΕ, πρέπει να γίνει κοινή πεποίθηση στη διοίκηση, και μάλιστα στο ανώτατο επίπεδο της, ότι το σύστημα θα εξελίσσεται διαρκώς και ότι είναι απαραίτητη η διάθεση οικονομικών, ανθρώπινων και άλλων πόρων για την παραγωγή διαδοχικών εκδόσεων.

Οι χρήστες του συστήματος, οι οποίοι σημειωτέον είναι υψηλόβαθμα στελέχη του οργανισμού, με βαθιά γνώση των αναγκών του και του στρατηγικού του προσανατολισμού, κατά τη διάρκεια της εργασίας τους αντιλαμβάνονται τα όρια του συστήματος, και ανακαλύπτουν νέες ανάγκες και δυνατότητες που θα μπορούσαν να καλυφθούν σε επόμενες εκδόσεις. Οι διαπιστώσεις διαφορετικών χρηστών μπορεί να είναι ταυτόσημες, επικαλυπτόμενες ή αντικρουόμενες. Επίσης, σε μεγάλους οργανισμούς, που είναι γεωγραφικά διασκορπισμένοι, οι νέες ιδέες που παράγονται παραμένουν εγκλωβισμένες σε τοπικά όρια, και δεν κυκλοφορούν σε όλη την έκταση του οργανισμού. Είναι λοιπόν σημαντικό να οργανωθεί η παραγωγή απόψεων και ιδεών σχετικά με την επέκταση του συστήματος. Για την εξυπηρέτηση αυτού του στόχου, ιδιαίτερα χρήσιμη είναι η δημιουργία μέσων για την καταγραφή των απόψεων, τη διευκόλυνση της επικοινωνίας και τη διεξαγωγή του διαλόγου. Εργαλεία που προέρχονται από το Web 2.0, όπως εταιρικά blogs και wikis, μπορούν να αποτελέσουν πλατφόρμα καταγραφής και επικοινωνίας απόψεων, που θα προετοιμάσουν τον καθορισμό των απαιτήσεων των νέων εκδόσεων του συστήματος και θα αποτελέσουν την απαρχή του νέου κύκλου ζωής του.

Ο κύκλος ζωής ανάπτυξης συστημάτων Επιχειρηματικής Ευφυΐας είναι ένα ευρύ αντικείμενο, το οποίο περιλαμβάνει πολλά εξειδικευμένα λογικά και τεχνικά προβλήματα. Η πλήρης κάλυψη του στα πλαίσια του παρόντος συγγράμματος δεν είναι εφικτή, καθώς η έκταση του θέματος είναι τέτοια που δικαιολογεί την έκδοση ενός πλήρους βιβλίου αφιερωμένου σε αυτό. Στόχος του συγγραφέα είναι να σκιαγραφήσει τα βασικότερα σχετικά ζητήματα. Ο αναγνώστης μπορεί να αναζητήσει περισσότερες λεπτομέρειες, ειδικότερες οδηγίες για την εκτέλεση του έργου, ρόλους εμπλεκόμενων μερών, ειδικά παραδοτέα κάθε σταδίου, καθώς και πολλά άλλα, στο βιβλίο των Moss and Atre (2003).

12.2.8 Η Επιχειρηματική Ευφυΐα Ως Υπηρεσία

Σύμφωνα με όσα έχουν αναφερθεί μέχρι τώρα, η ανάπτυξη συστημάτων ΕΕ είναι μια δραστηριότητα που απαιτεί σημαντικούς πόρους. Μεγάλες επιχειρήσεις με πιεστικές ανάγκες για πληροφόρηση θα αποφασίσουν ευκολότερα να διαθέσουν τους αναγκαίους πόρους. Για μικρότερες όμως επιχειρήσεις το κόστος μπορεί να αποδειχθεί αποτρεπτικός ή και απαγορευτικός παράγοντας. Άλλοι ανασχετικοί παράγοντες είναι η πολυπλοκότητα του εγχειρήματος και η έλλειψη εμπειρίας. Μια λύση, που περιορίζει δραστικά το κόστος και διευκολύνει στην ανάπτυξη του συστήματος, είναι η Επιχειρηματική Ευφυΐα ως Υπηρεσία (ΕΕΩΥ) (Business Intelligence As A Service), κατά τα πρότυπα του Λογισμικού ως Υπηρεσία (Software As A Service)

Η υπολογιστική νέφους (cloud computing) αλλάζει τη θεώρηση των υπολογιστικών συστημάτων. Τα συστήματα δεν είναι πλέον ιδιόκτητα, εγκατεστημένα σε χώρους του ιδιοκτήτη, και δεν λειτουργούν με δική του μέριμνα. Αντιθέτως, υπηρεσίες υπολογιστικών συστημάτων προσφέρονται από παρόχους, μέσω του Διαδικτύου. Οι πάροχοι έχουν την ευθύνη της λειτουργίας του συστήματος και διαθέτουν την αναγκαία υποδομή. Στα πλαίσια της υπολογιστικής νέφους, μπορεί να προσφέρονται εφαρμογές (Software As A Service), εργαλεία ανάπτυξης (Platform As A Service), καθώς και υπολογιστικοί πόροι υποδομής (Infrastructure As A Service). Η προσέγγιση αυτή εξασφαλίζει οικονομία κλίμακας, καθώς οι χρήστες μοιράζονται υπολογιστικούς πόρους, τεχνογνωσία, υπηρεσίες υποστήριξης κλπ.

Δεδομένου ότι ένα σύστημα ΕΕ απαρτίζεται από πολλά υποσυστήματα, όπως η Αποθήκη Δεδομένων, το λογισμικό για τις εργασίες ΕΜΦ, το λογισμικό για τη διεξαγωγή αναλύσεων, το λογισμικό οπτικοποίησης των αποτελεσμάτων και το σύστημα τήρησης των μεταδεδομένων, τίθεται το ερώτημα ποια υποσυστήματα θα με-

ταφερθούν στο νέφος και ποια θα παραμείνουν ως ιδιόκτητα συστήματα. Επίσης, η επιλογή του κατάλληλου παρόχου, καθώς και η αξιολόγηση των υπηρεσιών που προσφέρει και της αξιοπιστίας του, είναι ένα περίπλοκο πρόβλημα.

Ένα ζήτημα που χρήζει ιδιαίτερης προσοχής είναι αυτό της ασφάλειας. Τα συστήματα ΕΕ περιέχουν απόρρητα δεδομένα, και το γεγονός ότι τα δεδομένα αυτά θα αποθηκεύονται σε εξωτερικούς υπολογιστές ή ότι θα «ταξιδεύουν» στο διαδίκτυο προκαλεί σκεπτικισμό. Πάροχοι λύσεων ΕΕΩΥ, αναγνωρίζοντας το δίκαιο του αιτήματος για ασφάλεια των δεδομένων, εξασφαλίζουν πιστοποίηση μέσω τρίτων φορέων. Το Πιστοποιητικό του Ελεγκτικού Πρότυπου 70 (Statement of Auditing Standards 70 (SAS 70)) του American Institute of Certified Public Accountants (AICPA) διασφαλίζει ότι πάροχοι υπηρεσιών, οι οποίοι φιλοξενούν ή επεξεργάζονται δεδομένα που ανήκουν σε πελάτες τους, διαθέτουν επαρκείς ελέγχους και μέσα προστασίας για τα δεδομένα. Ένα άλλο βασικό ζήτημα είναι οι επιδόσεις του συστήματος. Η παροχή της υπηρεσίας μέσω δικτύου μπορεί να επιφέρει καθυστερήσεις στην εκτέλεση των εργασιών. Το πρόβλημα μπορεί να οφείλεται στον πάροχο, και σε αυτήν την περίπτωση ο χρήστης δεν έχει δυνατότητα παρέμβασης. Μια δυνατή λύση είναι η υπογραφή σύμβασης με τον πάροχο, η οποία θα προβλέπει ότι είναι υποχρεωμένος να ρυθμίσει το σύστημα του έτσι ώστε να εξασφαλίζει ικανοποιητικές επιδόσεις. Τέλος, το λογισμικό που προσφέρεται ως υπηρεσία διαθέτει τις λειτουργικότητες που έχει ορίσει ο κατασκευαστής. Εξειδικευμένες ανάγκες διαχείρισης των δεδομένων ή ανάλυσης τους, ανάγκες πιθανότατα σημαντικές για έναν οργανισμό, μπορεί να μην καλύπτονται από το έτοιμο αυτό λογισμικό.

Η Επιχειρηματική Ευφυΐα Ως Υπηρεσία βρίσκεται ακόμα στα αρχικά της στάδια. Μοντέλα, που αποτυπώνουν τα δομικά της στοιχεία και περιγράφουν ένα σκεπτικό και κριτήρια για την επιλογή του κατάλληλου παρόχου, έχουν προταθεί από την ακαδημαϊκή κοινότητα. Ενδεικτικά υποδεικνύονται στον αναγνώστη οι εργασίες των Baars and Kemper (2010) καθώς και των Muriithi and Kotze (2013), στις οποίες παρουσιάζονται τέτοια μοντέλα.

Μεγάλοι κατασκευαστές λογισμικού Επιχειρηματικής Ευφυΐας έχουν αξιολογήσει τη σημασία της ΕΕΩΥ και παρέχουν σχετικά προϊόντα. Η Oracle με το λογισμικό Business Intelligence Cloud Service προσφέρει συλλογή και μεταφόρτωση δεδομένων, εργασίες ΕΜΦ, προχωρημένες δυνατότητες ανάλυσης, εργαλεία οπτικοποίησης και δημιουργίας dashboards, διαδραστική πρόσβαση μέσω κινητών συσκευών και ασφάλεια δεδομένων. Η Microsoft αξιοποιεί την τεχνογνωσία της στον αυτοματισμό γραφείου και προτείνει τη λύση Power Business Intelligence for Office 365. Το λογισμικό εξασφαλίζει ευελιξία και απομακρυσμένη πρόσβαση, αυξημένες δυνατότητες συνεργασίας, μοντελοποίηση στο EXCEL, διαχείριση δεδομένων, καθώς και δημιουργία dashboards και τρισδιάστατων απεικονίσεων. Η IBM προσφέρει μια σειρά λύσεων βασισμένων στην υπολογιστική νέφους, που καλύπτουν τη διαχείριση πελατών με δυνατότητες διαδικτυακού, κινητού και κοινωνικού μάρκετινγκ, την ανάλυση λειτουργιών, την ανάλυση κινδύνου για τον υπολογισμό και τη διαχείριση του ρίσκου, τη διεξαγωγή προγνωστικών αναλύσεων και τη λήψη αποφάσεων από άτομα και ομάδες ατόμων και, τέλος, την ανάλυση των χρηματοοικονομικών στοιχείων και τη σύνταξη των χρηματοοικονομικών καταστάσεων. Η γκάμα των διαθέσιμων προϊόντων είναι μεγάλη και περιλαμβάνει εκδόσεις υπολογιστικής νέφους για τα λογισμικά Cognos, SPSS και Watson Analytics.

12.3 Παράγοντες Επιτυχίας σε Έργα Επιχειρηματικής Ευφυΐας

Τα συστήματα Επιχειρηματικής Ευφυΐας εμφανίστηκαν στο προσκήνιο ως τεχνολογικές λύσεις, οι οποίες προσφέρουν ολοκλήρωση δεδομένων και αυξημένες δυνατότητες ανάλυσης τους. Τελικός στόχος τους είναι να παρέχουν την απαραίτητη πληροφόρηση στα στελέχη που εμπλέκονται στη διαδικασία λήψης αποφάσεων. Επιτυχημένα συστήματα υψηλής ποιότητας, τα οποία χρησιμοποιούνται κατάλληλα από τους χρήστες τους, έχουν σημαντικές επιπτώσεις στην ποιότητα των αποφάσεων που λαμβάνονται, και τελικά στην άσκηση διοίκησης σε έναν οργανισμό. Με αυτόν τον τρόπο, τα επιτυχημένα συστήματα ΕΕ είναι ικανά να παράξουν πολύτιμη αξία. Για τον λόγο αυτό, έχει βαρύνουσα σημασία η μελέτη ζητημάτων που σχετίζονται με την ποιότητα τους, καθώς και με τις αναγκαίες συνθήκες και προϋποθέσεις για τον σχεδιασμό, υλοποίηση, χρήση και εφαρμογή τους. Τα ζητήματα αυτά δεν σχετίζονται αποκλειστικά με τεχνικές παραμέτρους, όπως πχ την ταχύτητα ή τους τρόπους μετάδοσης της πληροφορίας, αλλά αναφέρονται κυρίαρχα στους τρόπους με τους οποίους οι οργανισμοί μπορούν να συλλέξουν την αξία της πληροφορίας.

Για την επιτυχή διαχείριση και ολοκλήρωση ενός έργου Επιχειρηματικής Ευφυΐας είναι πολύ σημαντικό να εντοπιστούν και να ληφθούν υπόψη οι λεγόμενοι Κρίσιμοι Παράγοντες Επιτυχίας (ΚΠΕ). Οι παράγοντες αυτοί παίζουν καθοριστικό ρόλο στην επίτευξη των στόχων του έργου, ενώ η ανεπαρκής αντιμετώπιση και διαχείριση τους μπορεί να οδηγήσει το έργο σε αποτυχία. Σύμφωνα με τον ορισμό του Rockart (1979), κρίσι-

μοι παράγοντες επιτυχίας είναι ένας περιορισμένος αριθμός περιοχών, στις οποίες τα αποτελέσματα, εάν είναι ικανοποιητικά, θα διασφαλίσουν την επιτυχή και ανταγωνιστική επίδοση ενός οργανισμού.

Ο εντοπισμός και ανάδειξη παραγόντων επιτυχίας σε διάφορες κατηγορίες έργων πληροφορικής τεχνολογίας είναι συνηθισμένη πρακτική, και σε πολλές περιπτώσεις έχει αποδώσει αναγνωρισμένα και ευρέως αποδεκτά αποτελέσματα. Το πολυδιάστατο μοντέλο επιτυχίας πληροφοριακών συστημάτων των DeLone and McLean (1992) είναι ένα από τα πλέον αναφερόμενα. Στην περίπτωση όμως έργων Επιχειρηματικής Ευφυΐας δεν υπάρχει ένα καθορισμένο και ευρύτερα αποδεκτό σύνολο παραγόντων επιτυχίας, και μάλιστα όπως αυτοί γίνονται αντιληπτοί από την πλευρά του μάνατζμεντ. Ένας από τους λόγους για αυτό είναι ότι μέχρι τώρα η αγορά της Επιχειρηματικής Ευφυΐας καθοδηγείται από τη βιομηχανία της πληροφορικής και τους παρόχους λογισμικού. Ένας άλλος λόγος είναι ότι η ανάπτυξη συστημάτων ΕΕ είναι ένα σχετικά καινούργιο αντικείμενο, τουλάχιστον σε σύγκριση με την ανάπτυξη συστημάτων παρακολούθησης συναλλαγών. Τα συστήματα παρακολούθησης συναλλαγών είναι εγκατεστημένα εδώ και πολλές δεκαετίες στις επιχειρήσεις, και για τον λόγο αυτό υπάρχει πολύ τεχνογνωσία σχετικά με την ανάπτυξη και την εφαρμογή τους. Αντιθέτως, η ζήτηση για συστήματα Επιχειρηματικής Ευφυΐας παρουσιάζει έξαρση τα τελευταία χρόνια. Για τον λόγο αυτό, η εμπειρία που σχετίζεται με ζητήματα ανάπτυξης τους δεν είναι ανάλογη με την εμπειρία σχετικά με την ανάπτυξη άλλων συστημάτων.

Τα συστήματα ΕΕ παρουσιάζουν από τη φύση τους πολλές ιδιαιτερότητες. Σε μεγάλο βαθμό αυτό οφείλεται στο γεγονός ότι τα συστήματα αυτά δεν σχετίζονται με τις καλά καθορισμένες, και εν πολλοίς τυποποιημένες επιχειρηματικές διεργασίες καθημερινής λειτουργίας, οι οποίες είναι ευκολότερο να μηχανογραφηθούν, αλλά αφορούν την υποστήριξη της λήψης αποφάσεων από τη διοίκηση, διαδικασία λιγότερο τυποποιημένη, συχνά μεταβαλλόμενη, και που περιλαμβάνει σημαντικό βαθμό αβεβαιότητας. Τα διοικητικά καθήκοντα είναι λιγότερα συχνά οργανωμένα με χρήση αυστηρά καθορισμένων διαδικασιών (Popovic, Hackney, Coelho & Jaklic, 2012). Ο μικρότερος βαθμός δόμησης προκαλεί δυσκολίες στον εντοπισμό των απαιτήσεων του χρήστη. Επιπλέον, σε αντίθεση με τα συστήματα παρακολούθησης συναλλαγών, τα συστήματα ΕΕ, τα οποία βελτιώνουν τη διαδικασία λήψης αποφάσεων, προσφέρουν οφέλη μακροπρόθεσμα, έμμεσα και δύσκολα μετρήσιμα. Μια άλλη ιδιαιτερότητα των συστημάτων ΕΕ είναι ότι χρησιμοποιούν δεδομένα άλλων συστημάτων, τα οποία σε πολλές περιπτώσεις μπορεί να είναι διάσπαρτα, μη ομογενή ή και αδόμητα. Καθορισμένοι παράγοντες επιτυχίας άλλων πληροφοριακών συστημάτων δεν μπορούν να εφαρμοστούν αυτόματα σε συστήματα ΕΕ, ακριβώς λόγω της ιδιαιτερότητας της φύσης τους.

Η ανάγκη συστηματικής μελέτης των παραγόντων, οι οποίοι επηρεάζουν την επιτυχία έργων Επιχειρηματικής Ευφυΐας, προκύπτει και από το γεγονός ότι έχουν παρατηρηθεί πολύ μεγάλες αποκλίσεις στις αποδόσεις των συστημάτων ΕΕ. Αναφέρονται περιπτώσεις εξαιρετικά επιτυχημένων έργων ΕΕ. Για παράδειγμα, η Continental Airlines επέτυχε απόδοση (Return On Investment (ROI)) 1000%, αυξάνοντας τα έσοδα της και συμπιέζοντας τα κόστη. Αντίθετα, άλλες επιχειρήσεις δεν μπόρεσαν να επιτύχουν τόσο υψηλά ποσοστά απόδοσης (Sangar & Iahad, 2013). Επιπλέον, καταγράφεται ένα μεγάλο ποσοστό έργων ΕΕ τα οποία έχουν οδηγηθεί στην αποτυχία, και είτε εγκαταλείφθηκαν είτε δεν επέτυχαν τους στόχους τους (Frolick & Lindsey, 2003). Οι Williams and Williams (2007) επισημαίνουν ότι η απόδοση της επένδυσης (ROI) σε συστήματα ΕΕ είναι αμφισβητήσιμη, εξαιτίας των προκλήσεων στη διαδικασία ανάπτυξης τους.

Η ακαδημαϊκή βιβλιογραφία που ασχολείται με το θέμα των παραγόντων επιτυχίας σε έργα ΕΕ είναι μάλλον περιορισμένη (Yeoh & Koronios, 2010; Dinter, 2013). Ορισμένες μελέτες συναρτούν το αντικείμενο με συγκεκριμένα πεδία εφαρμογής, όπως πχ οργανισμούς παροχής υπηρεσιών υγείας, ενώ άλλες αναφέρονται σε εμπειρίες που προέρχονται από συγκεκριμένες χώρες. Κάποιες εργασίες αντιμετωπίζουν το ζήτημα ευρύτερα, και ασχολούνται με παράγοντες επιτυχίας, οι οποίοι ισχύουν γενικότερα σε έργα επιχειρηματικής ευφυΐας.

Μια σημαντική εργασία, η οποία ασχολείται με τους παράγοντες επιτυχίας έργων ΕΕ, είναι αυτή των Yeoh and Koronios (2010). Οι συγγραφείς εφάρμοσαν μια ερευνητική μεθοδολογία, η οποία συνδυάζει συνεντεύξεις ειδικών με τη μελέτη περιπτώσεων (case studies). Εξετάστηκαν διάφοροι παράγοντες επιτυχίας, συναρτημένοι με τρεις διαστάσεις ενδιαφέροντος, δηλαδή τον οργανισμό, τη διαδικασία και την τεχνολογία. Τα αποτελέσματα της έρευνας συνοψίζονται στα παρακάτω:

Παράγοντας 1. Σταθερή υποστήριξη από τη Διοίκηση και ανάληψη του αναγκαίου κόστους. Οι περισσότεροι συμμετέχοντες ανέδειξαν τον παράγοντα αυτόν ως τον σημαντικότερο. Όπως αναφέρθηκε και παραπάνω, η ανάπτυξη συστημάτων ΕΕ είναι μια κυκλική διαδικασία σχεδιασμού, υλοποίησης και αποτίμησης, η οποία επαναλαμβάνεται διαρκώς. Η συνεχής εξέλιξη του έργου, μέσα από επαναλαμβανόμενους κύκλους, απαιτεί τη διάθεση των κατάλληλων οικονομικών και ανθρώπινων πόρων. Η υποστήριξη από τη διοίκηση, και μάλιστα από το ανώτατο επίπεδο, εξασφαλίζει τους αναγκαίους πόρους. Επιπλέον, τα έργα ΕΕ σχετίζονται με τον στρατηγικό προσανατολισμό των επιχειρήσεων, και διαπερνούν οριζοντίως τα επιμέρους τμήματα

των οργανισμών. Το γεγονός αυτό μπορεί να προκαλέσει προστριβές που σχετίζονται με τις επιχειρησιακές διαδικασίες, την ποιότητα των δεδομένων κλπ. Με τη συνδρομή της διοίκησης οι διαφορές επιλύονται αποτελεσματικά.

Παράγοντας 2. Καθαρό στρατηγικό όραμα και σαφές επιχειρηματικό ζητούμενο. Για την επιτυχία του έργου, πρέπει ο οργανισμός να έχει ξεκάθαρο μακροπρόθεσμο στρατηγικό όραμα. Επίσης, το επιχειρηματικό ζητούμενο που θα εξυπηρετηθεί από το σύστημα ΕΕ πρέπει να είναι σαφές και να εντάσσεται οργανικά στον στρατηγικό προσανατολισμό. Έργα εναρμονισμένα με τη στρατηγική του οργανισμού κερδίζουν ευκολότερα την υποστήριξη της διοίκησης,, και άρα τους αναγκαίους πόρους για την υλοποίηση τους. Αντιθέτως, μη εναρμονισμένα έργα δεν εξυπηρετούν τους στόχους του οργανισμού και οδηγούνται στην αποτυχία. Επίσης, πολύ σημαντικό είναι το σύστημα να απαντά στις ανάγκες ενός συγκεκριμένου επιχειρηματικού προβλήματος. Τα έργα ΕΕ πρέπει να αποτελούν απάντηση σε ένα καθορισμένο επιχειρηματικό ζητούμενο - περίπτωση. Συγκεκριμένες επιχειρηματικές ανάγκες θέτουν ένα πρόβλημα το οποίο θα απαντηθεί μέσω του έργου ΕΕ.

Παράγοντας 3. Σταθμισμένη σύνθεση ομάδας έργου με τη συμμετοχή ειδικών άριστων σε επιχειρηματικά θέματα. Η σύνθεση της ομάδας που θα ασχοληθεί με το έργο ΕΕ είναι κομβικής σημασίας. Σε συνήθη έργα ανάπτυξης πληροφοριακών συστημάτων, η διεύθυνση και η εκτέλεση του έργου ανατίθεται σε ειδικούς πληροφορικής. Στα έργα ΕΕ οι ανάγκες είναι διαφορετικές. Ιδιαίτερης σημασίας είναι η συμμετοχή στελεχών εξειδικευμένων στα επιχειρησιακά ζητήματα. Οι ειδικοί αυτοί μπορούν να κατανοήσουν ή και να προβλέψουν τις επιχειρηματικές ανάγκες και τις σχετικές απαιτήσεις για πληροφόρηση. Επίσης, αντιλαμβάνονται τον στρατηγικό προσανατολισμό της επιχείρησης και μπορούν να εντάξουν σε αυτόν το έργο ΕΕ. Για την επιτυχία του έργου απαιτείται κατάλληλη μείξη άριστων στελεχών εξειδικευμένων στα επιχειρηματικά και στα τεχνικά θέματα

Παράγοντας 4. Οριοθέτηση του πεδίου και σταδιακή ανάπτυξη. Ο πλατειασμός του αντικειμένου είναι ένας ενδεχόμενος κίνδυνος για τα έργα ΕΕ. Για τον λόγο αυτό, απαιτείται η οριοθέτηση του πεδίου του έργου, δηλαδή ο ακριβής καθορισμός του τι θα περιλαμβάνεται στο έργο και τι θα αποκλείεται. Η οριοθέτηση του πεδίου επιτρέπει την καλύτερη κατανόηση του έργου, τον καλύτερο προγραμματισμό του και την επικέντρωση σε κρίσιμα ζητήματα.

Η οριοθέτηση πεδίου διευκολύνει και τη σταδιακή ανάπτυξη του έργου. Ειδικοί σε έργα ΕΕ συνιστούν ότι είναι προτιμότερο να ολοκληρώνονται και να παραδίδονται αυτόνομα τμήματα του έργου, που αφορούν πχ ένα τμήμα της επιχείρησης, και όχι να παραδοθεί εξαρχής ολόκληρο το έργο. Με την προσέγγιση αυτή περιορίζεται ο κίνδυνος αποτυχίας, το έργο οργανώνεται σε μικρότερα και καλύτερα κατανοητά βήματα, οι χρήστες έχουν ταχύτερα απτά αποτελέσματα με συνέπεια να αυξάνεται η υποστήριξη προς το έργο, και διευκολύνεται η μεταφορά γνώσης ως προς τον χειρισμό και τις δυνατότητες του συστήματος.

Παράγοντας 5. Συμμετοχή χρηστών στην ανάπτυξη του συστήματος. Σημαντικός παράγοντας επιτυχίας είναι η ενεργή συμμετοχή των χρηστών στη διαδικασία ανάπτυξης του συστήματος. Καταρχήν, η συμμετοχή των χρηστών επιτρέπει την ανάδειξη των αναγκών τους. Επιπλέον, οι χρήστες των συστημάτων ΕΕ, και ειδικά σε εργασίες OLAP, χρησιμοποιούν απευθείας τα μοντέλα δεδομένων, χωρίς να μεσολαβεί κάποια εφαρμογή. Η συμμετοχή των χρηστών θα διευκολύνει τον ορισμό των σωστών μοντέλων δεδομένων, των διαστάσεων τους, των περιεχομένων, των επιχειρηματικών κανόνων και των μεταδεδομένων.

Παράγοντας 6. Προσαρμοστικότητα και επεκτασιμότητα του συστήματος. Τα συστήματα ΕΕ είναι σε διαρκή εξέλιξη, ώστε να ανταποκρίνονται σε μεταβαλλόμενες ή και σε νέες ανάγκες πληροφόρησης. Για τον λόγο αυτό, η τεχνική υποδομή του συστήματος σε ότι αφορά το λογισμικό και το υλικό των υπολογιστών πρέπει να είναι ευπροσάρμοστη, και να επιτρέπει την επέκταση και την κλιμάκωση του. Πρέπει να είναι δυνατή η προσθήκη νέων πηγών δεδομένων, νέων στηλών και νέων διαστάσεων των κύβων. Επίσης, η ενσωμάτωση νέων δεδομένων από εξωτερικές πηγές είναι ένα συνηθισμένο παράδειγμα επέκτασης συστημάτων ΕΕ.

Παράγοντας 7. Ποιότητα των δεδομένων. Ένας πολύ σημαντικός παράγοντας επιτυχίας σε έργα ΕΕ είναι η ποιότητα των δεδομένων, και ειδικότερα των πηγαίων δεδομένων. Η ποιότητα των αρχικών δεδομένων επηρεάζει την ποιότητα της παραγόμενης πληροφορίας. Τα συστήματα ΕΕ συγκεντρώνουν και ομογενοποιούν δεδομένα από πολλές πηγές. Σε αρκετές περιπτώσεις τα προβλήματα των πηγαίων δεδομένων αποκαλύπτονται κατά τη μεταφορά τους στην Αποθήκη Δεδομένων ή ακόμα και κατά τη χρήση του συστήματος. Ένα πολύ συνηθισμένο πρόβλημα, που σχετίζεται με την ποιότητα των δεδομένων, είναι αυτό της διαφορετικής ονοματοδοσίας ταυτόσημων δεδομένων. Διαφορετικά τμήματα μιας επιχείρησης επιλέγουν ονόματα δεδομένων σύμφωνα με τη δική τους οπτική και ανάλογα με τις δικές τους ανάγκες. Τα διαφορετικά ονόματα δημιουργούν συγχύσεις και προβλήματα συνεννόησης. Ένα άλλο πρόβλημα είναι η χρήση διαφορετικών μονάδων μέτρησης. Για τη δημιουργία ενός συστήματος που διαπερνά οριζοντίως την επιχείρηση, όπως τα συστήματα ΕΕ, απαιτείται η τυποποίηση των δεδομένων.

Το ζήτημα του εντοπισμού σημαντικών παραγόντων, που επηρεάζουν την επιτυχία ή αποτυχία συστημάτων Επιχειρηματικής Ευφυΐας, εξετάζεται και σε δημοσιευμένη εργασία της Dinter (2013). Η συγγραφέας επεκτείνει τις τρέχουσες προσεγγίσεις ανάπτυξης συστημάτων ΕΕ, εισάγοντας τον όρο «εφοδιαστική της πληροφορίας» (information logistics) και τονίζει την ανάγκη οργάνωσης μιας στρατηγικής, που συστηματικά θα επιδιώκει σε επίπεδο ολόκληρης της επιχείρησης την άντληση πληροφόρησης. Η στρατηγική αυτή θα βρίσκεται σε συνδυασμό τόσο με τον ευρύτερο στρατηγικό προσανατολισμό της επιχείρησης όσο και με τη στρατηγική της για την ανάπτυξη πληροφοριακών συστημάτων. Έμφαση δίνεται στη διάχυση της πληροφορίας σε όλο το εύρος του οργανισμού, καθώς και στην αξιοποίηση συνεργειών που ξεπερνούν τοπικές και εξειδικευμένες διαδικασίες υποστήριξης αποφάσεων συγκεκριμένων χρηστών. Πολλά έργα ΕΕ καθοδηγούνται από ειδικές ανάγκες τμημάτων ενός οργανισμού. Στα πλαίσια μια στρατηγικής εφοδιαστικής της πληροφορίας, οι «τοπικές» αυτές λύσεις θα πρέπει να εντάσσονται και να ολοκληρώνονται σε ένα ευρύτερο σύστημα, που θα διασφαλίζει την ομαλή ροή της πληροφορίας σε όλον τον οργανισμό. Για την υποστήριξη μιας τέτοιας στρατηγικής είναι αναγκαία και η συγκρότηση διοικητικών δομών, που θα υλοποιούν τη στρατηγική, θα οργανώνουν συνέργειες και θα μεριμνούν για την ένταξη του συνολικού έργου στη γενική στρατηγική της επιχείρησης.

Η συγγραφέας πραγματοποίησε έρευνα, στην οποία συμμετείχαν με τη διαδικασία της ημιδομημένης συνέντευξης 226 στελέχη μεγάλων επιχειρήσεων, τα οποία ήταν εξειδικευμένα στην Επιχειρηματική Ευφυΐα. Η έρευνα πραγματοποιήθηκε κατά τη διάρκεια συνεδρίου αφιερωμένου στην Επιχειρηματική Ευφυΐα. Συγκεντρώθηκαν συνολικά 160 ερωτηματολόγια. Επιπροσθέτως, η συγγραφέας προσέφυγε στη σχετική βιβλιογραφία για άντληση πληροφοριών. Στα πλαίσια της έρευνας της, η συγγραφέας εντοπίζει και μελετά μια σειρά από παράγοντες, που επηρεάζουν την πορεία έργων επιχειρηματικής ευφυΐας. Αξιοσημείωτο είναι ότι οι παράγοντες δεν διαφέρουν ριζικά από αυτούς που εξετάζονται στην εργασία των Yeoh and Koronios (2010). Αντιθέτως, υπάρχουν αρκετές επικαλύψεις. Βεβαίως, η αντίληψη της συγγραφέως περί μιας ολιστικής στρατηγικής εφοδιαστικής της πληροφορίας αναπροσαρμόζει την οπτική γωνία θέασης του αντικειμένου. Συνοπτικά οι παράγοντες που μελετήθηκαν στα πλαίσια της εν λόγω έρευνας είναι οι παρακάτω:

Περιεκτικότητα. Υπό τον όρο «περιεκτικότητα» (comprehensiveness) εξετάζεται ο βαθμός διαφοροποίησης μιας τακτικής ανάπτυξης βραχυπρόθεσμων, εξειδικευμένων και αυτόνομων έργων ΕΕ από μια ολιστική τακτική, σύμφωνα με την οποία τα επιμέρους έργα εντάσσονται σε ευρύτερες στρατηγικές διαχείρισης της πληροφορίας.

Προσαρμοστικότητα. Η προσαρμοστικότητα αναφέρεται στη δυνατότητα προσαρμογής σε μεταβαλλόμενες καταστάσεις και επιχειρηματικές ανάγκες, αλλά και στη συνεχή αποτίμηση των στόχων και στη συνακόλουθη αναπροσαρμογή της στρατηγικής.

Πραγματογνωμοσύνη και υψηλή εξειδίκευση, κυρίως σε σχέση με την ικανότητα στρατηγικής αντίληψης.

Υποστήριξη από τη διοίκηση και μάλιστα από το ανώτατο επίπεδο της.

Επικοινωνία σχετικά με τους σκοπούς του έργου μεταξύ της διοίκησης και του προσωπικού

Προσανατολισμός πληροφοριακής στρατηγικής. Ευθυγράμμιση της στρατηγικής ανάπτυξης της εφοδιαστικής της πληροφορίας με τη γενικότερη στρατηγική ανάπτυξης πληροφοριακών συστημάτων.

Μεταφορά επιχειρηματική γνώσης. Αναφέρεται στην ολοκλήρωση των σχεδιασμών σε πληροφορική τεχνολογία με τα επιχειρηματικά σχέδια. Οι ειδικοί πληροφορικής τεχνολογίας πρέπει να γνωρίζουν τα επιχειρηματικά σχέδια και τη στρατηγική του οργανισμού, αλλά και να συμμετέχουν σε αυτούς τους σχεδιασμούς.

Συνεργασία ειδικών σε επιχειρηματικά ζητήματα με τους ειδικούς πληροφορικής τεχνολογίας.

Χρησιμοποιώντας τη στατιστική μέθοδο Partial Least Square, μελετήθηκε ο βαθμός επίδρασης των παραπάνω παραγόντων σε δύο ζητήματα, στην ποιότητα του συστήματος και στην επάρκεια παροχής πληροφορίας. Τα αποτελέσματα που προέκυψαν από την έρευνα είναι μικτά, επαληθεύουν αποτελέσματα άλλων προηγούμενων ερευνών, ενώ αντικρούουν άλλα. Διαπιστώνεται μια ισχυρή επίδραση της προσαρμοστικότητας, της υποστήριξης από τη διοίκηση και του προσανατολισμού πληροφοριακής στρατηγικής στην ποιότητα του συστήματος. Επίσης, η επάρκεια παροχής πληροφορίας φαίνεται να συναρτάται ισχυρά με την περιεκτικότητα, την επικοινωνία και τη συνεργασία. Η περιεκτικότητα επηρεάζει οριακά την ποιότητα του συστήματος. Σύμφωνα με τα αποτελέσματα, δεν υποστηρίζεται η συσχέτιση μεταξύ μεταφοράς επιχειρηματικής γνώσης και ποιότητας συστήματος – επάρκειας παροχής πληροφορίας, ούτε η συσχέτιση μεταξύ πραγματογνωμοσύνης και συνεργασίας με την ποιότητα του συστήματος. Σημειωτέον ότι μερικά από αυτά τα συμπεράσματα δεν επαληθεύονται από τα αποτελέσματα άλλων προηγούμενων ερευνών.

Άλλες ερευνητικές εργασίες συνεισφέρουν και αυτές με τα αποτελέσματα τους χρήσιμη γνώση σχετικά με τον εντοπισμό παραγόντων, οι οποίοι επηρεάζουν την επιτυχία έργων Επιχειρηματικής Ευφυΐας. Οι Wixom and Watson (2001) πραγματοποίησαν έρευνα με χρήση ερωτηματολογίων, τα οποία συμπλήρωσαν στελέχη εξειδικευμένα σε αποθήκες δεδομένων και πάροχοι δεδομένων από 111 οργανισμούς. Τα αποτελέσματα, τα

οποία προέκυψαν μετά από στατιστική ανάλυση των απαντήσεων, καταδεικνύουν συσχετίσεις ανάμεσα στην ποιότητα του συστήματος και των δεδομένων και στα αντιληπτά οφέλη. Και σε αυτήν την έρευνα η υποστήριξη από τη διοίκηση αναδεικνύεται ως ένας σημαντικός παράγοντας επιτυχίας, γιατί επιτρέπει την αποτελεσματική αντιμετώπιση ζητημάτων που προκύπτουν κατά τη διάρκεια του έργου. Σημαντική είναι επίσης η διάθεση επαρκών πόρων. Η συμμετοχή των χρηστών στη διαδικασία ανάπτυξης του συστήματος, η ποιότητα των δεδομένων των πηγαίων συστημάτων και η σύνθεση της ομάδας έργου, η οποία πρέπει να αποτελείται από στελέχη υψηλής ειδίκευσης και αυξημένων ικανοτήτων, είναι επιπλέον παράγοντες που επηρεάζουν την πορεία εξέλιξης του έργου.

Οι Seah, Hsieh and Weng (2010), διεξάγοντας μια μελέτη περίπτωσης σε μια κινεζική εταιρεία τηλεπικοινωνιών, η οποία ολοκλήρωσε με επιτυχία την ανάπτυξη συστήματος Επιχειρηματικής Ευφυΐας, μελετούν παράγοντες, που επηρέασαν την υλοποίηση του έργου. Οι συγγραφείς διαπιστώνουν ότι το έργο συνάντησε αντίσταση από τους εργαζόμενους και ότι αναδείχθηκαν ζητήματα διαχείρισης της μεταβολής (change management). Η στάση των ανώτατων διοικητικών στελεχών, οι οποίοι ακολούθησαν μια ισχυρή, αποφασιστική αλλά και προσαρμοστική πολιτική, η οποία λαμβάνει υπόψη τις ιδιαίτερες συνθήκες του εργασιακού περιβάλλοντος, αντιμετώπισε τα προβλήματα και αποτέλεσε σημαντικό παράγοντα επιτυχίας του έργου.

Οι Popovic et al. (2012) μελετούν ζητήματα ποιότητας συστημάτων Επιχειρηματικής Ευφυΐας, χρησιμοποιώντας έννοιες και τροποποιώντας μοντέλα που προέρχονται από το πεδίο της ποιότητας πληροφοριακών συστημάτων. Ως διαστάσεις ποιότητας ενός συστήματος ΕΕ ορίζονται η ωριμότητα του συστήματος, η ποιότητα των περιεχομένων της πληροφορίας, η ποιότητα της πρόσβασης στην πληροφορία και η αναλυτική κουλτούρα στη λήψη αποφάσεων, ενώ ως μέτρο της ποιότητας του συστήματος ορίζεται η χρήση της πληροφορίας στις επιχειρηματικές διεργασίες. Σημειωτέον ότι η έννοια της ποιότητας της πληροφορίας διασπάται, και μελετάται ξεχωριστά το ζήτημα του περιεχόμενου της πληροφορίας και το ζήτημα της πρόσβασης στην πληροφορία. Θεωρείται ότι η χρήση της πληροφορίας στις επιχειρηματικές διεργασίες παράγει επιχειρηματική αξία. Επίσης, μελετάται η σχέση αλληλεπίδρασης μεταξύ των διαστάσεων ποιότητας. Αναλυτικότερα, οι διαστάσεις ποιότητας του συστήματος καθορίζονται ως εξής:

Ωριμότητα. Η ωριμότητα του συστήματος σχετίζεται με δύο παράγοντες, α)την ολοκλήρωση των δεδομένων και β)τις αναλυτικές δυνατότητες. Η ολοκλήρωση των δεδομένων μετριέται με τον τρόπο που τα δεδομένα συγκεντροποιούνται και ολοκληρώνονται, και με το εάν δεδομένα που προέρχονται από διαφορετικές πηγές είναι συνεπή και δεν παρουσιάζουν αντιφάσεις. Οι αναλυτικές δυνατότητες αναφέρονται σε τρόπους ανάλυσης, όπως σύνταξη αναφορών, OLAP, εξόρυξη δεδομένων, Κρίσιμους Δείκτες Επίδοσης κλπ.

Ποιότητα περιεχομένων πληροφορίας. Η ποιότητα των περιεχομένων της πληροφορίας αναφέρεται σε ζητήματα, όπως το εάν η πληροφορία είναι επαρκής, ακριβής, κατανοητή, σχετική με το προς διερεύνηση ζητούμενο, επίκαιρη και απαλλαγμένη από λάθη και αποκλίσεις.

Πρόσβαση στην πληροφορία. Η ποιότητα της πρόσβασης στην πληροφορία καθορίζεται από την ταχύτητα πρόσβασης, τις δυνατότητες διαδραστικής πρόσβασης, τον βαθμό κάλυψης των αναγκών των χρηστών από την παρεχόμενη πληροφορία, καθώς και από τη γνώση σχετικά με πηγές της πληροφορίας.

Αναλυτική κουλτούρα στη λήψη αποφάσεων. Η πληροφορία, όσο ποιοτική και αν είναι, είναι εν δυνάμει χρήσιμη. Για να αποδώσει καρπούς η πληροφορία πρέπει να χρησιμοποιηθεί μέσα στις επιχειρηματικές διεργασίες και να βελτιώσει τις αποφάσεις. Ο βαθμός χρήσης αναλυτικών πρακτικών στις διαδικασίες λήψης αποφάσεων μετριέται με την ύπαρξη καλά καθορισμένων διαδικασιών για τη λήψη αποφάσεων, με την παγίωση πολιτικών ενσωμάτωσης πληροφοριών σε κάθε διαδικασία λήψης αποφάσεων και με τη χρήση των πληροφοριών σε κάθε απόφαση που λαμβάνεται.

Οι συγγραφείς διεξήγαγαν έρευνα με χρήση ερωτηματολογίων. Στην έρευνα συμμετείχαν ανώτατα στελέχη μεσαίου και μεγάλου μεγέθους επιχειρήσεων από τη Σλοβενία. Οι συμμετέχοντες είχαν επαρκή γνώση σε ζητήματα συστημάτων Επιχειρηματικής Ευφυΐας. Συνολικά συγκεντρώθηκαν 181 ερωτηματολόγια, τα στοιχεία των οποίων υποβλήθηκαν σε στατιστική ανάλυση. Στόχος της έρευνας ήταν η κατανόηση των παραγόντων που επηρεάζουν την ποιότητα συστημάτων Επιχειρηματικής Ευφυΐας, αλλά και των σχέσεων αλληλεπίδρασης μεταξύ αυτών των παραγόντων.

Τα αποτελέσματα της έρευνας συνοψίζονται στα παρακάτω. Και η ολοκλήρωση των δεδομένων και οι αναλυτικές δυνατότητες είναι σημαντικές για την ωριμότητα του συστήματος. Όμως η εισαγωγή αναλυτικών τεχνολογιών, όπως OLAP και η εξόρυξη δεδομένων, είναι αυτή που προκαλεί την επίτευξη υψηλότερου επιπέδου ωριμότητας. Η ωριμότητα του συστήματος έχει θετική επίδραση και στην ποιότητα των περιεχομένων της πληροφορίας και στην ποιότητα πρόσβασης στην πληροφορία. Οι επιχειρήσεις που υλοποιούν έργα ΕΕ επικεντρώνουν περισσότερο στην ποιοτική πρόσβαση στην πληροφορία και λιγότερο στην ποιότητα περιεχομένου της πληροφορίας. Ωστόσο, αναφορικά με τη χρήση της πληροφορίας στις επιχειρηματικές διεργασίες,

διαπιστώνεται ότι μόνο η ποιότητα του περιεχόμενου της πληροφορίας είναι σημαντική και όχι η ποιότητα της πρόσβασης. Αυτό σημαίνει ότι εάν ο χρήστης διαπιστώσει την αξία της πληροφορίας, τότε θα την αναζητήσει, ανεξάρτητα από το πόσο δύσκολο είναι να μπορέσει να τη βρει. Αντιστρόφως, η ακατάλληλη πληροφορία θα αποτρέψει τον χρήστη από το να τη χρησιμοποιήσει ή θα τον οδηγήσει σε μέτριες ή και εσφαλμένες αποφάσεις. Σημαντικός είναι και ο ρόλος της ύπαρξης μιας κουλτούρας που αξιοποιεί αναλυτικές τεχνικές για τη λήψη αποφάσεων. Η ύπαρξη μιας τέτοιας κουλτούρας πριμοδοτεί τη χρήση του συστήματος, ακόμα και όταν η ποιότητα της πληροφορίας δεν είναι ιδιαίτερα υψηλή. Η αξιοποίηση της πληροφορίας μεγεθύνεται όταν συνδυάζεται η ποιοτική πληροφορία με μια κουλτούρα χρήσης αναλυτικών τεχνικών στη λήψη αποφάσεων. Τα παραπάνω οδηγούν στο συμπέρασμα ότι κατά την ανάπτυξη συστημάτων ΕΕ, η βασική μέριμνα πρέπει να είναι η ποιοτική κάλυψη των πραγματικών αναγκών των χρηστών και όχι η γρήγορη επεξεργασία και μεταφορά της πληροφορίας. Αυτές οι πληροφοριακές ανάγκες σχετίζονται με διοικητικές διαδικασίες που συνδέουν τις επιχειρησιακές στρατηγικές με τη διαχείριση των επιχειρηματικών διαδικασιών, και περιλαμβάνουν τη στοχοθεσία, τον απολογισμό και τη συνακόλουθη δράση.

Οι εργασίες που επιλέχθηκαν για να συμπεριληφθούν στο παρόν Κεφάλαιο δίνουν μια αποκαλυπτική εικόνα της τρέχουσας κατάστασης πραγμάτων στον χώρο της εκτίμησης της ποιότητας των συστημάτων Επιχειρηματικής Ευφυΐας, καθώς και του καθορισμού των παραγόντων που παίζουν κρίσιμο ρόλο στην επιτυχία τους. Καταρχάς, είναι κοινή διαπίστωση ότι δεν υπάρχει ένα καλά καθορισμένο και ευρέως αποδεκτό σύνολο κρίσιμων παραγόντων επιτυχίας, ούτε υπάρχουν παγιωμένα και καθιερωμένα μοντέλα επιτυχίας. Τέτοια μοντέλα που αναφέρονται στα πληροφοριακά συστήματα υπάρχουν, είναι όμως αδύνατη η αυτόματη μεταφορά τους στα συστήματα ΕΕ, λόγω της ιδιαιτερότητας της φύσης τους. Όλοι οι συγγραφείς επισημαίνουν την ανάγκη διεξαγωγής πρόσθετης σχετικής έρευνας. Επιπλέον, στις παραπάνω εργασίες είναι εμφανής η ποικιλομορφία των προσεγγίσεων και η διαφορετικότητα στις οπτικές γωνίες θέασης του αντικειμένου. Άλλη εργασία αντιμετωπίζει το αντικείμενο χρησιμοποιώντας έννοιες και μοντέλα της θεωρίας ποιότητας πληροφοριακών συστημάτων, άλλη το συναρτά προνομιακά με μία στρατηγική εφοδιαστικής της πληροφορίας, και άλλη επικεντρώνει στην εμπειρική έρευνα, αποφεύγοντας τις θεωρητικές προεκτάσεις. Η ποικιλομορφία των προσεγγίσεων είναι ενδεικτική του ενδιαφέροντος που παρουσιάζει το αντικείμενο και του βαθμού προσέλκυσης ερευνητών με διαφορετικό υπόβαθρο. Ταυτόχρονα όμως επαληθεύει την ανυπαρξία καθιερωμένης θεωρίας και μοντέλων για την επιτυχία των συστημάτων ΕΕ.

Αναφορικά με τους παράγοντες που επηρεάζουν την επιτυχία των συστημάτων Επιχειρηματικής Ευφυΐας, μπορούμε καταρχάς να κάνουμε έναν βασικό διαχωρισμό ανάμεσα σε αυτούς που σχετίζονται με τεχνικά ζητήματα και σε αυτούς που αφορούν τον οργανισμό και τις διαδικασίες του. Κοινή συνισταμένη όλων των εργασιών είναι ότι οι δεύτεροι είναι περισσότερο σημαντικοί και καθορίζουν σε μεγαλύτερο βαθμό την επιτυχία του έργου. Αναλυτικότερα, η στάση της ανώτατης διοίκησης απέναντι στο έργο, η αποφασιστική συμμετοχή και υποστήριξη της, η διάθεση των απαραίτητων πόρων, η συνδρομή της για την αντιμετώπιση προβλημάτων και αντιθέσεων αναγνωρίζεται από πολλούς ερευνητές ως ένας πολύ σημαντικός, ίσως ο σημαντικότερος παράγοντας επιτυχίας. Ένα άλλο ζήτημα που αναδύεται ως καθοριστικός παράγοντας είναι αυτό της στρατηγικής διάστασης. Οργανισμοί με ξεκάθαρο στρατηγικό προσανατολισμό και με ικανότητα σύνδεσης της επιχειρησιακής στρατηγικής με τη διαχείριση των επιχειρηματικών διαδικασιών μπορούν να εντάξουν οργανικά ένα σύστημα Επιχειρηματικής Ευφυΐας στις πρακτικές παραγωγής πληροφορίας, χρήσιμης για τη λήψη αποφάσεων και την εξυπηρέτηση των στρατηγικών στόχων. Η ύπαρξη επιχειρησιακής στρατηγικής επιτρέπει και τη συνάρθρωση της με μια πληροφοριακή στρατηγική, που συστηματικά θα επιδιώκει σε επίπεδο ολόκληρης της επιχείρησης την άντληση σχετικής πληροφόρησης.

Η στελέχωση της ομάδας έργου μπορεί να επηρεάσει τον βαθμό επιτυχίας του. Απαιτείται σταθμισμένη σύνθεση, η οποία, εκτός από ειδικούς πληροφορικής, θα περιλαμβάνει με βαρύνουσα σημασία στελέχη εξειδικευμένα στα επιχειρησιακά ζητήματα, ικανά να συνδέσουν τα ζητήματα αυτά με τις ανάγκες για σχετική πληροφόρηση, καθώς και να αντιληφθούν τον τρόπο ένταξης του έργου στη στρατηγική του οργανισμού. Η προσαρμοστικότητα του συστήματος είναι ένας ακόμα σημαντικός παράγοντας. Οι καταστάσεις και τα επιχειρηματικά ζητούμενα δεν είναι σταθερά αλλά μεταβάλλονται. Μαζί τους μεταβάλλονται και οι ανάγκες για πληροφόρηση, χρήσιμη για τη λήψη αποφάσεων. Το σύστημα θα πρέπει να είναι ικανό να ανταποκρίνεται σε αυτές τις ανάγκες, να επεκτείνει τις λειτουργίες του και να ενσωματώνει νέα δεδομένα, πιθανότατα εξωτερικά. Το ζήτημα της ποιότητας των δεδομένων παίζει και αυτό ουσιαστικό ρόλο, καθώς προβληματικά πηγαία δεδομένα ενδέχεται να προκαλέσουν την παραγωγή μη ορθής πληροφορίας.

Ως προς τα ποιοτικά χαρακτηριστικά του συστήματος, η ωριμότητα του επηρεάζεται από την ολοκλήρωση και ποιότητα των δεδομένων, κυρίως όμως από τις αναλυτικές του δυνατότητες. Η ωριμότητα του συστήματος θα συμβάλει στην παραγωγή πληροφορίας υψηλής ποιότητας ως προς το περιεχόμενο της, η οποία με τη σειρά

της θα ενισχύσει τον βαθμό χρήσης του συστήματος. Τέλος, η ύπαρξη μιας επιχειρησιακής κουλτούρας για την αξιοποίηση αναλυτικών τεχνικών στη λήψη αποφάσεων πριμοδοτεί τη χρήση του συστήματος και μεγεθύνει την αξιοποίηση της πληροφορίας.

Ο Πίνακας 12.1 συνοψίζει βασικούς παράγοντες επιτυχίας για την ευτυχή ολοκλήρωση έργου ΕΕ

Βασικοί παράγοντες επιτυχίας έργων Επιχειρηματικής Ευφυΐας
Αποφασιστική υποστήριξη από τη διοίκηση και μάλιστα από το ανώτατο επίπεδο της
Ξεκάθαρος στρατηγικός προσανατολισμός
Στελέχωση της ομάδας έργου. Βαρύνουσα σημασία έχει η συμμετοχή ειδικών στα επιχειρηματικά ζητήματα
Προσαρμοστικότητα του συστήματος
Ποιότητα δεδομένων
Αυξημένες αναλυτικές δυνατότητες
Ύπαρξη κουλτούρας για χρήση αναλυτικών τεχνικών στη λήψη αποφάσεων

Πίνακας 12.1 *Παράγοντες επιτυχίας έργων Επιχειρηματικής Ευφυΐας*

Βιβλιογραφία/Αναφορές

Alexander, I. F., & Maiden, N. (2004). *Scenarios, Stories, Use Cases: Through the Systems Development Life-cycle*. Chinchester, UK: John Wiley & Sons Ltd.

Baars, H., & Kemper, H. G. (2010). Business Intelligence in the Cloud?. *Proceedings of the 2010 Pacific Asia Conference on Information Systems*, 1528-1539.

Bara, A., Botha, I., Diakonita, V., Lungu, I., Velicanu, A., & Velicanu, M. (2009). A Model for Business Intelligence Systems Development. *Informatica Economica, 13*(4), 99-108.

Boehm, B. W. (1988). A Spiral Model of Software Development and Enhancement. *Computer, 21*(5), 61-72. doi: 10.1109/2.59

Davis, A. M., Bersoff, E. H., & Comer, E. R. (1988). A Strategy for Comparing Alternative Software Development Life Cycle Models. *IEEE Transactions on Software Engineering, 14*(10), 1453-1461. doi: 10.1109/32.6190

DeLone, W. H., & McLean, E. R. (1992). Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research, 3*(1), 60–95. doi: 10.1287/isre.3.1.60

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean Model of Information Systems Success: A Ten-year Update. *Journal of Management Information Systems, 19*(4), 9–30.

Dinter, B. (2013). Success Factors for Information Logistics Strategy – An Empirical Investigation. *Decision Support Systems, 54*(3), 1207-1218. doi: 10.1016/j.dss.2012.09.001

Frolick, M. N., & Lindsey, K. (2003). Critical Factors for Data Warehouse Failure. *Journal of Data Warehousing, 8*(1), 48-54.

Fuchs, G. (2006). The Vital BI Maintenance Process. In B. Sujatha (Ed.), *Business Intelligence Implementation: Issues and Perspectives* (pp. 116-123). Hyderabad, India: ICFAI University Press.

Gangadharan, G. R., & Swami, S. N. (2004). Business Intelligence Systems: Design and Implementation Strategies. *Proceedings of the 26th International Conference on Information Technology Interfaces,* 139-144. Cavtat: IEEE.

Henderson-Sellers, B., & Edwards, J. M. (1990). The Object-Oriented Systems Life Cycle. *Communications of the ACM, 33*(9), 142-159. doi: 10.1145/83880.84529

Larman, C., & Basili, V. R. (2003). Iterative and Incremental Development: A Brief History. *Computer*, *36*(6), 47-56. doi: 10.1109/mc.2003.1204375

Li, E. Y. (1997). Perceived Importance of Information Systems Success Factors: A Meta Analysis of Group Differences. *Information and Management*, *32*(1), 15-28. doi: 10.1016/s0378-7206(97)00005-0

Moss, L., & Atre, S. (2003). *Business Intelligence Roadmap: The Complete Lifecycle for Decision-Support Applications*. Boston, MA: Pearson Education Inc.

Muriithi, G. M., & Kotze, J. E. (2013). A Conceptual Framework for Delivering Cost Effective Business Intelligence Solutions as a Service. *Proceedings of the South African Institute for Computer Scientists and Information Technologies Conference,* 96-100. New York, NY: ACM. doi:10.1145/2513456.2513502

Popovic, A., Hackney, R., Coelho, P., & Jaklic, J. (2012). Towards Business Intelligence System Success: Effects of Maturity and Culture on Analytical Decision Making. *Decision Support Systems, 54*(1), 728-739. doi: 10.1016/j.dss.2012.08.017

Poon, P., & Wagner, C. (2001). Critical Success Factors Revisited: Success and Failure Cases of Information Systems for Senior Executives. *Decision Support Systems, 30*(4), 393-418. doi: 10.1016/s0167-9236(00)00069-5

Rockart, J. G. (1979). Chief Executives Define their own Data Needs. *Harvard Business Review, 57*(2), 81-93.

Sangar, A. B., & Iahad, N. B. (2013). Critical Factors That Affect the Success of Business Intelligence Systems (BIS) Implementation in an Organization. *International Journal of Scientific and Technology Research, 2*(2), 176- 180.

Seah, M., Hsieh, M. H., & Weng, P. D. (2010). A case analysis of Savecom: The Role of Indigenous Leadership in Implementing a Business Intelligence System. *International Journal of Information Management, 30*(4), 368-373. doi: 10.1016/j.ijinfomgt.2010.04.002

Strassmann, P. (1990). *The Business Value of Computers*. New Canaan, CT: The Information Economics Press.

Sumner, M. (1999). Critical success factors in enterprise wide information management system projects. *Proceedings of the 1999 ACM SIGCPR conference on Computer personnel research,* 297-303. New York, NY: ACM. doi: 10.1145/299513.299722

Watson, H. J., Abraham, D., Chen, D., Preston, D. S., & Thomas, D. (2004). Data Warehousing ROI: Justifying and Assessing a Data Warehouse. *Business Intelligence Journal, 9*(2), 6-17.

Williams, S., & Williams, N. (2007). *The Profit Impact of Business Intelligence*. San Francisco, CA: Morgan Kaufmann Publishers.

Wixom, B. H., & Watson, H. J. (2001). An Empirical Investigation of the Factors Affecting Data Warehouses Success. *MIS Quarterly, 25*(1), 17-41. doi: 10.2307/3250957

Yeoh, W., & Koronios, A. (2010, Spring). Critical Success Factors for Business Intelligence Systems. *Journal of Computer Information Systems*, 23-32.

13 Οδηγός WEKA

Σύνοψη

Το WEKA είναι μια σουίτα λογισμικού για μηχανική μάθηση και εξόρυξη δεδομένων. Αναπτύχθηκε στο πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και διατίθεται ως ελεύθερο λογισμικό. Η μεγάλη ποικιλία μεθόδων εξόρυξης δεδομένων που περιλαμβάνει, η συνεχής υποστήριξη και εξέλιξη του από μια διεθνή ομάδα προγραμματιστών, η ελεύθερη διανομή του πηγαίου κώδικα και η δυνατότητα εγκατάστασης του σε διαφορετικές πλατφόρμες υλικού και λογισμικού είναι ορισμένοι από τους παράγοντες που συμβάλλουν στην ευρύτερη αποδοχή και στη μεγάλη διάδοση του. Επίσης, η γραφική διεπαφή που διαθέτει επιτρέπει τη χρήση του από χρήστες, οι οποίοι δεν έχουν ικανότητες προγραμματισμού. Το παρόν κεφάλαιο αποτελεί μια σύντομη παρουσίαση του WEKA. Ειδικότερα, το κεφάλαιο αναφέρεται στο WEKA Explorer, το οποίο αποτελεί και τη δημοφιλέστερη διεπαφή. Με το WEKA Explorer ο χρήστης μπορεί να εκτελέσει εργασίες προεπεξεργασίας δεδομένων, κατηγοριοποίησης, ανάλυσης συστάδων, ανάλυσης κανόνων συσχέτισης, επιλογής χαρακτηριστικών και οπτικοποίησης των δεδομένων. Σε ότι αφορά την προεπεξεργασία των δεδομένων, γίνεται αναφορά στις διάφορες πηγές δεδομένων και παρουσιάζονται τα αρχεία τύπου ARFF. Το γραφικό περιβάλλον του tab "Preprocess" επιτρέπει την εύκολη διερεύνηση της κατανομής τιμών στα διάφορα πεδία, τη διαγραφή πεδίων και την εκτέλεση διαφόρων αλγορίθμων προεπεξεργασίας, οι οποίοι εμφανίζονται υπό την ονομασία "filters". Παρέχονται εργαλεία για προσθήκη νέων υπολογιζόμενων πεδίων, για κανονικοποίηση και διακριτοποίηση αριθμητικών τιμών, για συγχώνευση ονομαστικών πεδίων, για δειγματοληψία, για μείωση διαστάσεων με Ανάλυση Κυρίων Συνιστωσών, για επιλογή χαρακτηριστικών κλπ.
Οι αλγόριθμοι και τα εργαλεία κατηγοριοποίησης που διαθέτει το WEKA είναι αξιοσημείωτοι. Παρέχονται υλοποιήσεις όλων των κύριων μεθόδων κατηγοριοποίησης, όπως Δένδρα Αποφάσεων, Νευρωνικά Δίκτυα, Μηχανές Διανυσμάτων Υποστήριξης, Μπαϋεσιανοί κατηγοριοποιητές, Λογιστική Παλινδρόμηση, k-Πλησιέστεροι Γείτονες κλπ. Για κάθε μέθοδο υπάρχουν πολλές δυνατότητες παραμετροποίησης. Επίσης, διατίθενται πολλές παραλλαγές των βασικών μεθόδων, αλλά και εργαλεία για τη δημιουργία σύνθετων κατηγοριοποιητών bagging και boosting, κατηγοριοποιητών ευαίσθητων στο κόστος, κατηγοριοποιητών που χρησιμοποιούν ανάλυση συστάδων κλπ. Ο χρήστης μπορεί να επικυρώσει τα μοντέλα του εφαρμόζοντας τη μέθοδο cross validation, τη μέθοδο holdout ή χρησιμοποιώντας ένα ανεξάρτητο σύνολο δεδομένων. Για κάθε μοντέλο παρουσιάζονται αναλυτικά στοιχεία για τις επιδόσεις και τη δομή του (πχ τα βάρη των συνδέσεων ενός δικτύου Multilayer Perceptron). Το WEKA περιλαμβάνει αρκετούς αλγορίθμους Ανάλυσης Συστάδων, όπως τον k-Means, τη Συσσωρευτική Ιεραρχική ΑΣ και το DBSCAN. Κάθε αλγόριθμος μπορεί να παραμετροποιηθεί. Επίσης, υπάρχει δυνατότητα οπτικής αναπαράστασης της κατανομής των παρατηρήσεων στις συστάδες. Το tab "Associate" περιλαμβάνει αλγορίθμους για ανάλυση Κανόνων Συσχέτισης, μεταξύ των οποίων και τον βασικό αλγόριθμο Apriori. Υπάρχει η δυνατότητα εξόρυξης κανόνων συσχέτισης σε δεδομένα με πεδίο κλάσης. Οι κανόνες αυτοί θα έχουν στο δεξιό τμήμα τους μια τιμή κλάσης. Στο tab "Select attributes" ο χρήστης μπορεί να πειραματιστεί με διάφορες μεθόδους επιλογής χαρακτηριστικών και να συνδυάσει μεθόδους αναζήτησης με μεθόδους αξιολόγησης χαρακτηριστικών. Τέλος, στο tab "Visualize" υπάρχει ένας πίνακας διαγραμμάτων διασποράς. Ο χρήστης, κάνοντας κλικ σε ένα διάγραμμα, μπορεί να το προβάλει σε ξεχωριστό παράθυρο.

Προηγούμενη Γνώση

Στο παρόν κεφάλαιο γίνεται παρουσίαση του λογισμικού μηχανικής μάθησης και εξόρυξης δεδομένων WEKA. Περιλαμβάνονται αλγόριθμοι για προεπεξεργασία των δεδομένων, κατηγοριοποίηση, ανάλυση συστάδων, κανόνες συσχέτισης, επιλογή χαρακτηριστικών και οπτικοποίηση. Οι αλγόριθμοι αυτοί έχουν παρουσιαστεί αναλυτικά σε προηγούμενα κεφάλαια του παρόντος συγγράμματος. Το παρόν κεφάλαιο επικεντρώνει στο συγκεκριμένο λογισμικό, στις χειριστικές ιδιαιτερότητες του και στις ειδικές δυνατότητες του. Για την κατανόηση του κεφαλαίου είναι απαραίτητη η προηγούμενη κατανόηση της σχετικής θεωρίας και των αλγορίθμων. Ειδικότερα, είναι απαραίτητη η εμπέδωση των περιεχομένων του Κεφαλαίου 6, το οποίο εισάγει τον αναγνώστη στην εξόρυξη δεδομένων, του Κεφαλαίου 7, το οποίο καλύπτει θέματα προεπεξεργασίας των δεδομένων, του Κεφαλαίου 8, το οποίο αναφέρεται στους κανόνες συσχέτισης, των Κεφαλαίων 9 και 10, τα οποία παρουσιάζουν αναλυτικά τους αλγόριθμους κατηγοριοποίησης και του Κεφαλαίου 11, το οποίο ασχολείται με μεθόδους ανάλυσης συστάδων. Παρουσίαση των διαγραμμάτων διασποράς και των πινάκων διαγραμμάτων διασποράς γίνεται στο Κεφάλαιο 5.

13.1 Εισαγωγή

Το WEKA (Waikato Environment for Knowledge Analysis) είναι μια σουίτα λογισμικού για μηχανική μάθηση και Εξόρυξη Δεδομένων. Αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και πήρε το όνομα του από το Weka, ένα μικρό και υπό εξαφάνιση πουλί της Ν. Ζηλανδίας. Το WEKA ανήκει στην κατηγορία του λεγόμενου "ελεύθερου λογισμικού" (freeware) και διατίθεται δημοσίως σύμφωνα με τους όρους της άδειας GNU General Public License, η οποία επιτρέπει στους χρήστες να χρησιμοποιούν, αλλά και να τροποποιούν ελεύθερα το λογισμικό.

Το WEKA είναι ένα από τα πιο διαδεδομένα λογισμικά Εξόρυξης Δεδομένων. Έχει χρησιμοποιηθεί σε μεγάλο αριθμό επιστημονικών εργασιών, και αρκετά βιβλία Εξόρυξης Δεδομένων αναφέρονται σε αυτό. Η μεγάλη δημοφιλία του οφείλεται στα ειδικά χαρακτηριστικά του και στις δυνατότητες που προσφέρει. Αναλυτικότερα το WEKA:

- Περιέχει αρκετά μεγάλη ποικιλία μεθόδων για κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, και κανόνες συσχέτισης. Επίσης, παρέχει δυνατότητες για προεπεξεργασία των δεδομένων, καθώς και εργαλεία οπτικοποίησης.
- Είναι λογισμικό ανοικτού κώδικα. Αυτό σημαίνει ότι ο πηγαίος κώδικας είναι δημοσίως διαθέσιμος. Χρήστες με γνώσεις προγραμματισμού μπορούν να τροποποιούν και να εξελίσσουν τους αλγορίθμους.
- Είναι γραμμένο σε γλώσσα Java, γεγονός που το καθιστά ικανό να εγκαθίσταται σε διαφορετικές πλατφόρμες υλικού και λογισμικού.
- Διαθέτει γραφικό περιβάλλον εργασίας. Στο διαδίκτυο υπάρχει διαθέσιμη μεγάλη ποικιλία βιβλιοθηκών για μηχανική μάθηση και εξόρυξη δεδομένων. Ωστόσο, η χρήση τους απαιτεί τη συγγραφή κώδικα. Αντιθέτως, το γραφικό περιβάλλον του WEKA επιτρέπει τη χρήση του λογισμικού από τελικούς χρήστες, οι οποίοι δεν διαθέτουν γνώσεις προγραμματισμού.

Το WEKA διατίθεται σε δύο διαφορετικές εκδόσεις:

- Στη λεγόμενη "σταθερή" (stable) έκδοση, η οποία απευθύνεται σε τελικούς χρήστες και αντιστοιχεί στην τελευταία έκδοση του βιβλίου των Witten, Frank and Hall (2011).
- Στην έκδοση η οποία απευθύνεται σε προγραμματιστές. Η έκδοση αυτή χρησιμοποιείται από την κοινότητα των προγραμματιστών του WEKA για τη διόρθωση σφαλμάτων και την επέκταση των δυνατοτήτων του λογισμικού.

Το πανεπιστήμιο του Waikato διατηρεί μια ιδιαίτερα πλούσια ιστοθέση αφιερωμένη στο WEKA (http://www.cs.waikato.ac.nz/ml/weka/). Στην ιστοθέση αυτή οι χρήστες μπορούν:

- Να προμηθευτούν το WEKA (http://www.cs.waikato.ac.nz/ml/weka/downloading.html). Προσφέρονται διαφορετικές επιλογές για λειτουργικά συστήματα Windows, Mac OS X και Linux.
- Να αναζητήσουν τεκμηρίωση σχετικά με το λογισμικό. Η τεκμηρίωση περιλαμβάνει το manual του λογισμικού, οδηγίες για την αντιμετώπιση προβλημάτων, απαντήσεις σε συχνές ερωτήσεις, οδηγίες για σύνδεση με γλώσσες προγραμματισμού όπως το Matlab και η R, παρουσιάσεις και ηλεκτρονικά σεμινάρια, καθώς και πολλά άλλα.
- Να προμηθευτούν το Application Programming Interface (API) του λογισμικού, καθώς και μια μεγάλη λίστα πρόσθετων πακέτων για διάφορες εργασίες μηχανικής μάθησης και εξόρυξης δεδομένων.
- Να προμηθευτούν ένα σημαντικό αριθμό συνόλων δεδομένων, τα οποία μπορούν να χρησιμοποιήσουν για εξάσκηση.

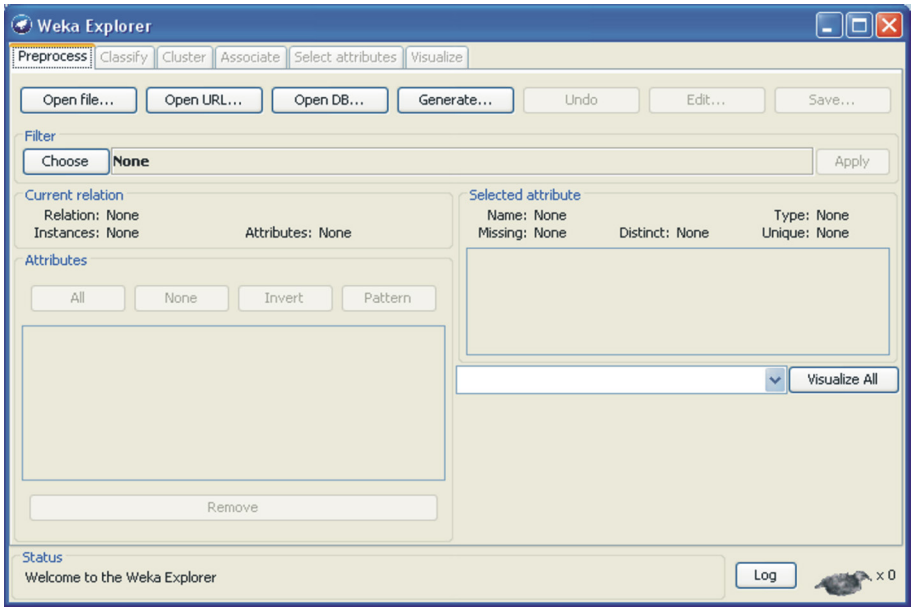
Η πιο πρόσφατη έκδοση είναι η 3.6.12 και θα παρουσιαστεί στα πλαίσια του παρόντος κεφαλαίου

Εικόνα 13.1 *Εκκίνηση του WEKA*

Με την εκκίνηση του WEKA εμφανίζεται το παράθυρο της Εικόνας 13.1. Από το σημείο αυτό ο χρήστης μπορεί να εκκινήσει τις κύριες εφαρμογές του WEKA:

- Ο **Explorer** είναι η πιο δημοφιλής διεπαφή. Ο χρήσης μπορεί να εκτελέσει όλες τις κύριες εργασίες Εξόρυξης Δεδομένων, όπως κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, ανακάλυψη κανόνων συσχέτισης, προεπεξεργασία των δεδομένων και οπτικοποίηση.
- Ο **Experimenter** είναι ένα περιβάλλον για διεξαγωγή πειραμάτων, όπου αξιολογούνται μέθοδοι κατηγοριοποίησης και παλινδρόμησης. Διευκολύνει τη σύγκριση της επίδοσης διαφορετικών μοντέλων και παρουσιάζει τα αποτελέσματα σε μορφή πίνακα.
- Το **Knowledge Flow** είναι ένα περιβάλλον που επιτρέπει τη διεξαγωγή των ιδίων εργασιών με τον Explorer, διαθέτει όμως διαφορετική διεπαφή (interface). Στο περιβάλλον αυτό χρησιμοποιούνται components, τα οποία συνδέονται μεταξύ τους με γραφικό τρόπο, ο οποίος ορίζει τη ροή εργασίας. Υπάρχουν components για τη φόρτωση των δεδομένων, την προεπεξεργασία τους, τη δημιουργία και εκπαίδευση μοντέλων, την οπτικοποίηση κλπ.

Όπως αναφέρθηκε και προηγουμένως, ο Explorer είναι το πιο δημοφιλές περιβάλλον. Για τον λόγο αυτό, στον παρόντα μικρό οδηγό θα γίνει παρουσίαση του Explorer. Το περιβάλλον εργασίας του Explorer παρουσιάζεται στην Εικόνα 13.2. Το παράθυρο της εφαρμογής περιλαμβάνει 6 tabs για προεπεξεργασία των δεδομένων, κατηγοριοποίηση, ανάλυση συστάδων, επιλογή γνωρισμάτων και οπτικοποίηση,

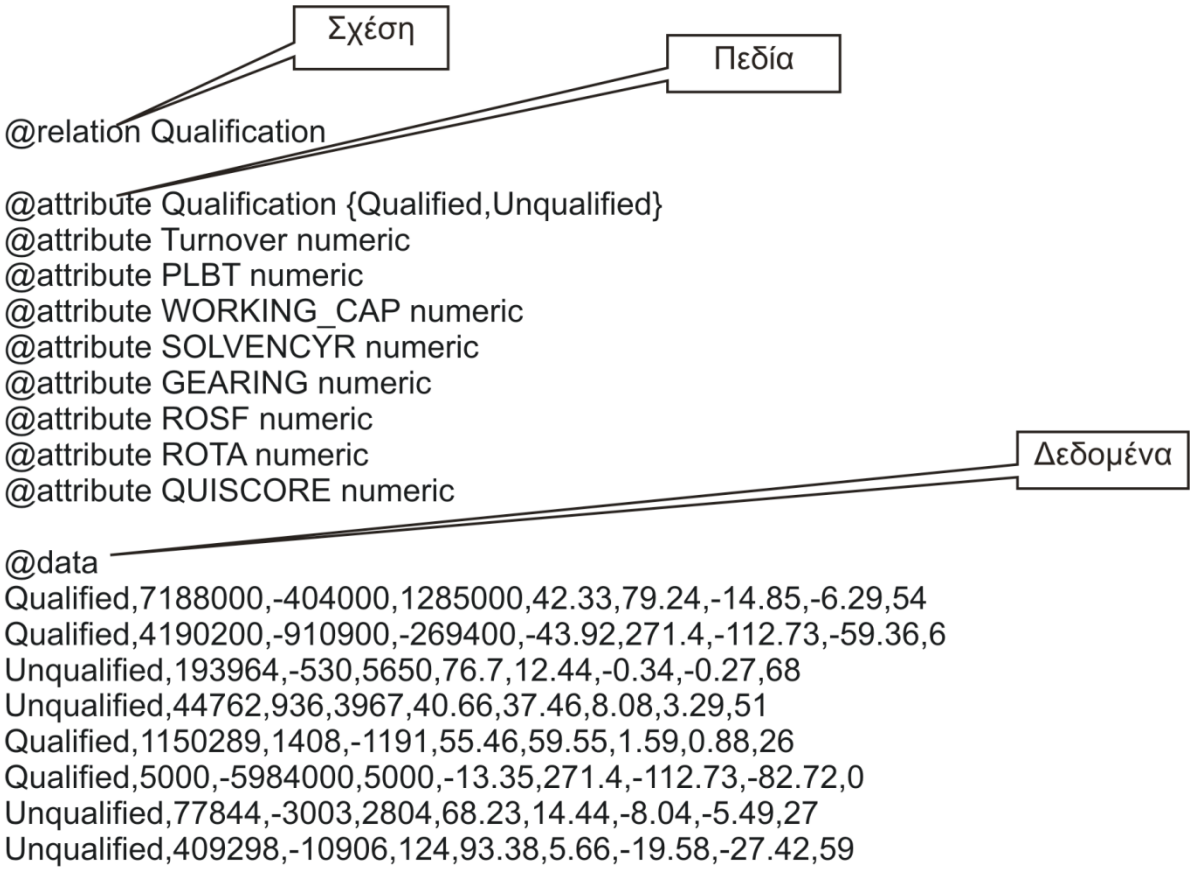


Εικόνα 13.2 *WEKA Explorer*

13.2 Προεπεξεργασία

Κατά τη χρήση του WEKA Explorer, το πρώτο βήμα είναι η εισαγωγή των δεδομένων. Δεδομένα μπορούν να εισαχθούν από μια SQL βάση δεδομένων (με χρήση του Java Data Base Connectivity (JDBC)) ή από μια διεύθυνση URL. Ο πιο συνηθισμένος όμως τρόπος φόρτωσης δεδομένων είναι μέσω ενός αρχείου ARFF. Τα αρχεία ARFF είναι απλά αρχεία κειμένου, όπου οι τιμές διαχωρίζονται με κόμμα (Coma Separated Values (CSV)). Επιπλέον, το αρχείο περιέχει μια επικεφαλίδα, στην οποία ορίζονται το όνομα της σχέσης (πίνακα δεδομένων) και τα πεδία. Παράδειγμα αρχείου ARFF παρουσιάζεται στην Εικόνα 13.3.

Στην αρχή του αρχείου αναφέρεται η λέξη ″@relation″ και ακολουθεί το όνομα του πίνακα δεδομένων (Qualification). Στη συνέχεια γίνεται η δήλωση των πεδίων. Για κάθε πεδίο χρειάζεται μια γραμμή, στην αρχή της οποίας υπάρχει η λέξη ″@attribute″, ακολουθεί το όνομα του πεδίου (πχ Turnover), και κατόπιν δηλώνεται ο τύπος του πεδίου. Αν το πεδίο είναι αριθμητικό, χρησιμοποιείται η λέξη ″numeric″. Αν το πεδίο είναι ονομαστικό, δηλώνονται οι δυνατές τιμές μέσα σε αγκύλες. Για παράδειγμα, το πρώτο πεδίο έχει όνομα ″Qualification″, είναι ονομαστικό και μπορεί να πάρει δύο τιμές, την τιμή ″Qualified″ και την τιμή ″Unqualified″. Μετά τη δήλωση των πεδίων ακολουθούν τα δεδομένα. Πριν από τα καθεαυτό δεδομένα υπάρχει μια γραμμή με τη λέξη ″@data″. Τα δεδομένα είναι τιμές, οι οποίες χωρίζονται με κόμμα. Σημειώνεται ότι τα δεδομένα του παραδείγματος αφορούν επιχειρήσεις και τον εξωτερικό τους έλεγχο. Κάθε γραμμή των δεδομένων αντιστοιχεί σε μια επιχείρηση. Στο πρώτο πεδίο καταγράφεται το αποτέλεσμα του εξωτερικού ελέγχου. Επιχειρήσεις οι οποίες πήραν δυσμενή σχόλια από τους εξωτερικούς ελεγκτές χαρακτηρίζονται ″Qualified″, ενώ επιχειρήσεις οι οποίες δεν πήραν δυσμενή σχόλια από τους εξωτερικούς ελεγκτές χαρακτηρίζονται ″Unqualified″. Τα υπόλοιπα πεδία είναι διάφοροι αριθμοδείκτες. Τα δεδομένα αυτά, με περισσότερους όμως αριθμοδείκτες, θα χρησιμοποιηθούν στα παραδείγματα που θα ακολουθήσουν.



Εικόνα 13.3 *Αρχείο ARFF*

Στο Διαδίκτυο διατίθεται εφαρμογή μετατροπής αρχείων Excel σε αρχεία ARFF. Μπορείτε να προμηθευτείτε την εφαρμογή από ιστοδελίδα της sourceforge.

Μετά την εισαγωγή των δεδομένων, στο παράθυρο της προεπεξεργασίας παρουσιάζονται διάφορες πληροφορίες για τα δεδομένα. Επίσης, ο χρήστης μπορεί να εκτελέσει εργασίες διερευνητικής ανάλυσης και προεπεξεργασίας:

- Στο αριστερό μέρος του παραθύρου εμφανίζονται τα πεδία. Ο χρήστης μπορεί να επιλέξει ορισμένα από αυτά και να τα διαγράψει πατώντας το κουμπί ″Remove″.
- Ο χρήστης μπορεί να δει και να τροποποιήσει τις τιμές των δεδομένων πατώντας το κουμπί ″Edit″. Επίσης, μπορεί να διαγράψει ολόκληρες γραμμές. Τα τροποιημένα δεδομένα μπορούν να αποθηκευτούν σε ένα νέο αρχείο ARFF με το πάτημα του κουμπιού ″Save″.
- Με την επιλογή ενός πεδίου, στο δεξιό μέρος του παράθυρου εμφανίζονται στοιχεία για το πεδίο. Εάν το πεδίο είναι αριθμητικό, παρουσιάζονται η μέγιστη και η ελάχιστη τιμή, η μέση τιμή και η τυπική απόκλιση. Εάν το πεδίο είναι ονομαστικό, εμφανίζονται οι δυνατές τιμές και το πλήθος των παρατηρήσεων οι οποίες έχουν την εκάστοτε τιμή.
- Εάν τα δεδομένα περιέχουν πεδίο κλάσης τότε ο χρήστης ορίζει το πεδίο αυτό. Εάν δεν υπάρχει πεδίο κλάσης τότε χρησιμοποιείται η τιμή ″no class″.
- Στο κάτω και δεξιό τμήμα του παράθυρου εμφανίζεται γραφικά η κατανομή των τιμών του επιλεγμένου πεδίου. Εάν έχει οριστεί πεδίο κλάσης, τότε σε κάθε ράβδο το πλήθος των παρατηρήσεων για την εκάστοτε τιμή κλάσης παρουσιάζεται χρησιμοποιώντας διαφορετικά χρώματα.

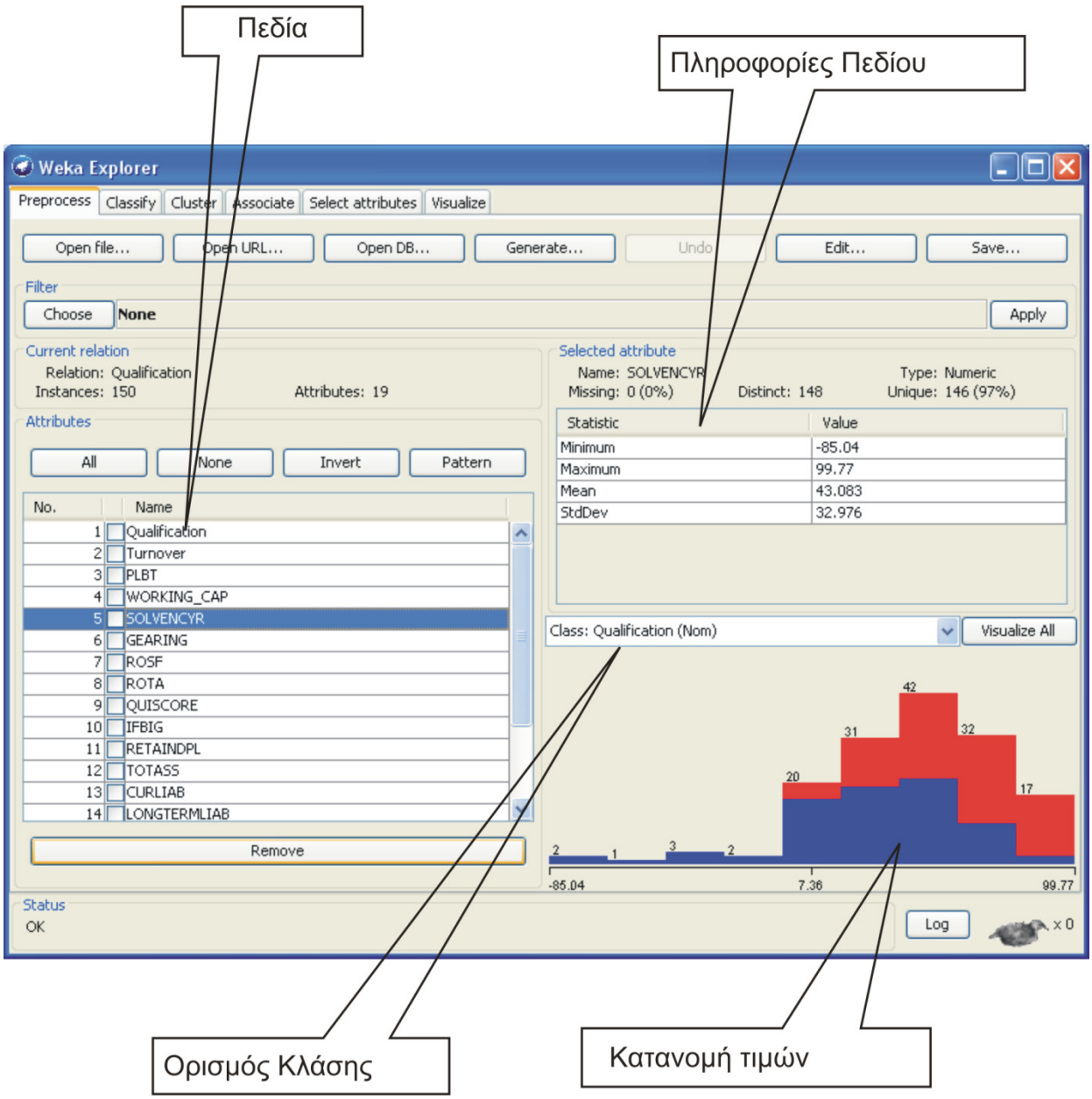
Στην Εικόνα 13.4 παρουσιάζεται παράδειγμα με το παράθυρο της προεπεξεργασίας, μετά την εισαγωγή των δεδομένων. Ως πεδίο κλάσης έχει οριστεί το ″Qualification″. Το πεδίο ″SolvencyR″ είναι επιλεγμένο. Αναγράφονται η μεγαλύτερη, η μικρότερη και η μέση τιμή, καθώς και η τυπική απόκλιση. Ιδιαίτερα χρήσιμη είναι η οπτική αναπαράσταση της κατανομής των τιμών. Στο παράδειγμα μας, οι εταιρείες οι οποίες πήραν σχόλια (Qualified) συμβολίζονται με μπλέ χρώμα, ενώ οι εταιρείες οι οποίες δεν πήραν σχόλια (Unqualified) συμβολίζονται με κόκκινο χρώμα. Από τη γραφική αναπαράσταση κατανομής των τιμών φαίνεται ότι υπάρχει διαφοροποίηση των τιμών της μεταβλητής SolvencyR ανάλογα με την κλάση, και ότι οι Qualified εταιρείες τείνουν να έχουν μικρότερες τιμές Solvency Ratio από τις Unqualified.

Στο παράθυρο της προεπεξεργασίας ο χρήστης μπορεί να διενεργήσει διερευνητική ανάλυση. Επιλέγοντας ένα – ένα τα πεδία παρατηρεί την κατανομή των τιμών. Με τον τρόπο αυτό, εντοπίζει μεταβλητές στις οποίες υπάρχει σημαντική διαφοροποίηση των τιμών ανάλογα με την κλάση. Επίσης, μπορεί να εντοπίσει μεταβλητές, στις οποίες παρουσιάζονται ακραίες τιμές. Κάνοντας edit τα δεδομένα, ο χρήστης εντοπίζει και μελετά τις αντίστοιχες παρατηρήσεις, και εάν τις θεωρήσει θόρυβο, μπορεί να τις διαγράψει. Το κουμπί ″Visualize All″ εμφανίζει την κατανομή των τιμών για όλες τις μεταβλητές.

Πέραν της διερευνητικής ανάλυσης, ο χρήστης μπορεί να εφαρμόσει μεθόδους αυτοματοποιημένης προεπεξεργασίας των δεδομένων. Ο καθορισμός της μεθόδου που θα εφαρμοστεί γίνεται στο πεδίο ″filter″. Ο χρήστης πατώντας το κουμπί ″Choose″ επιλέγει από λίστα τη μέθοδο προεπεξεργασίας την οποία θα εφαρμόσει. Υπάρχουν επιβλεπόμενες και μη επιβλεπόμενες μέθοδοι για την προεπεξεργασία πεδίων (στηλών) και παρατηρήσεων (γραμμών). Αφού επιλεγεί μια μέθοδος, μπορεί να γίνει ρύθμιση των παραμέτρων της. Ο χρήστης, κάνοντας κλικ στο όνομα της μεθόδου, ανοίγει το παράθυρο ρύθμισης των παραμέτρων της. Αφού καθοριστεί η επιθυμητή μέθοδος, γίνεται εφαρμογή της με το πάτημα του κουμπιού ″Apply″.

Το WEKA προσφέρει μεγάλη ποικιλία μεθόδων για προεπεξεργασία δεδομένων. Ορισμένες από τις συνηθέστερες εργασίες που μπορούν να εκτελεστούν είναι:

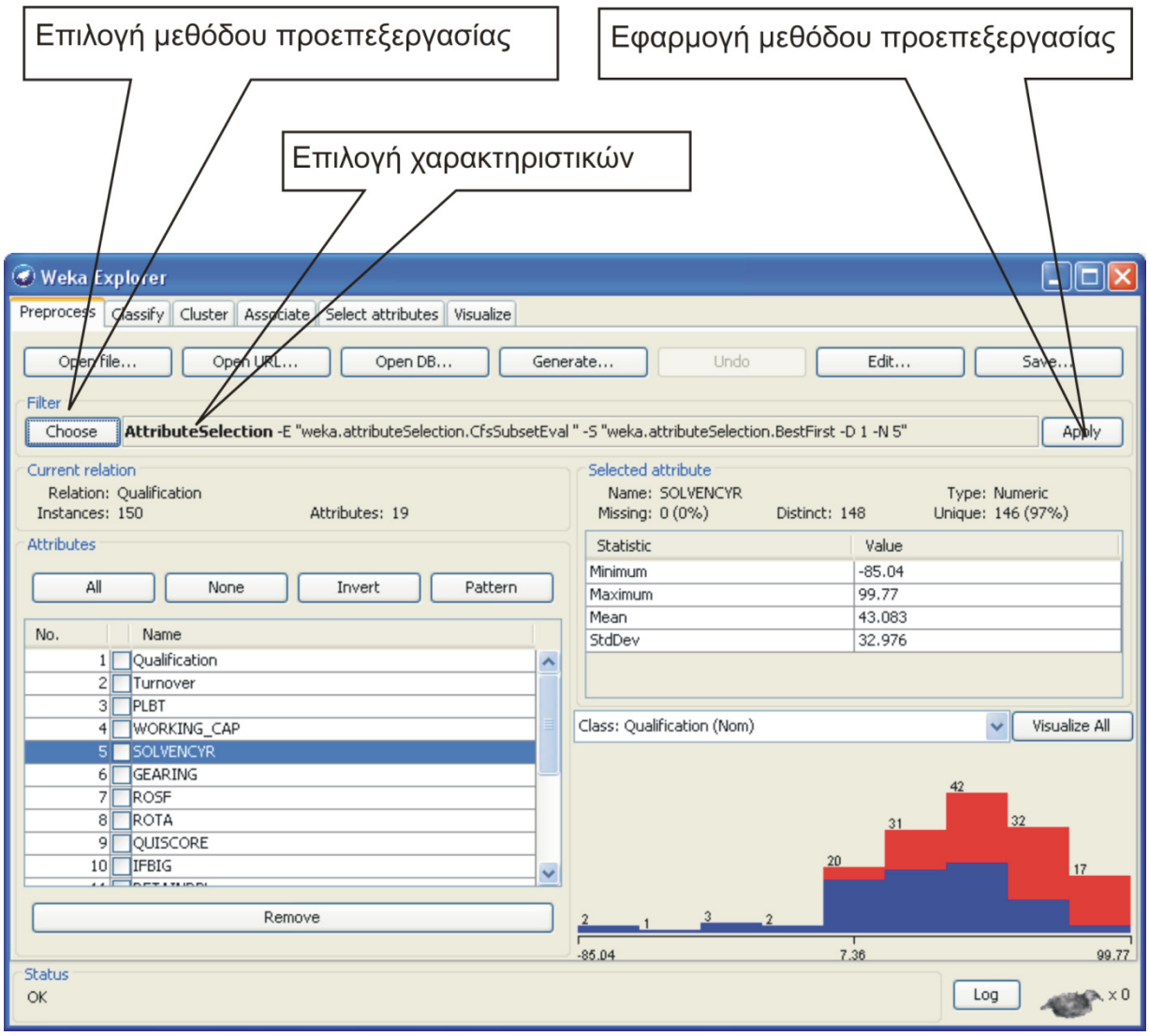
- Η προσθήκη νέων υπολογιζόμενων πεδίων.
- Η κανονικοποίηση αριθμητικών τιμών.
- Η διακριτοποίηση αριθμητικών τιμών.
- Η μετατροπή αριθμητικών και ονομαστικών πεδίων σε δυαδικά.
- Η συγχώνευση δύο ονομαστικών πεδίων.
- Η μείωση των διαστάσεων με Ανάλυση Κυρίων Συνιστωσών.
- Η δημιουργία νέων συνόλων δεδομένων με εφαρμογή δειγματολειψίας.



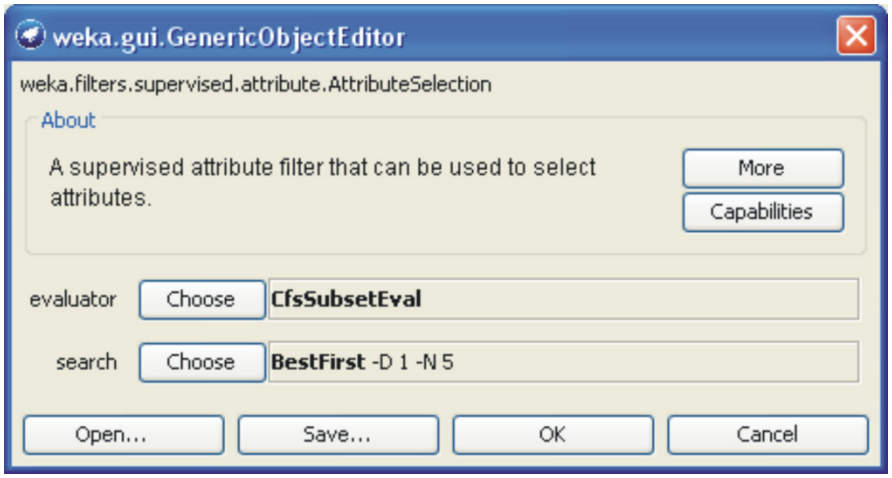
Εικόνα 13.4 *Προεπεξεργασία Δεδομένων*

Η σημαντικότερη ίσως εργασία προεπεξεργασίας των δεδομένων είναι η επιλογή σημαντικών στηλών. Από την αναδιπλούμενη λίστα που εμφανίζεται με το πάτημα του κουμπιού "Choose" στο πεδίο "Filter", ο χρήστης επιλέγει την εργασία weka/filters/supervised/attribute/AttributeSelection. Στην Εικόνα 13.5 ο χρήστης έχει δηλώσει ότι θα εκτελέσει επιλογή χαρακτηριστικών.

Το WEKA προσφέρει πολλές μεθόδους επιλογής χαρακτηριστικών. Για να καθορίσει μια συγκεκριμένη μέθοδο, ο χρήστης κάνει κλικ στο όνομα της μεθόδου προεπεξεργασίας (AttributeSelection) και ανοίγει το αντίστοιχο παράθυρο (Εικόνα 13.6). Στο πεδίο "evaluator" γίνεται ο καθορισμός της συγκεκριμένης μεθόδου επιλογής χαρακτηριστικών η οποία θα εφαρμοστεί. Το WEKA προτείνει εξ ορισμού τη μέθοδο CFS (Hall, 1999).

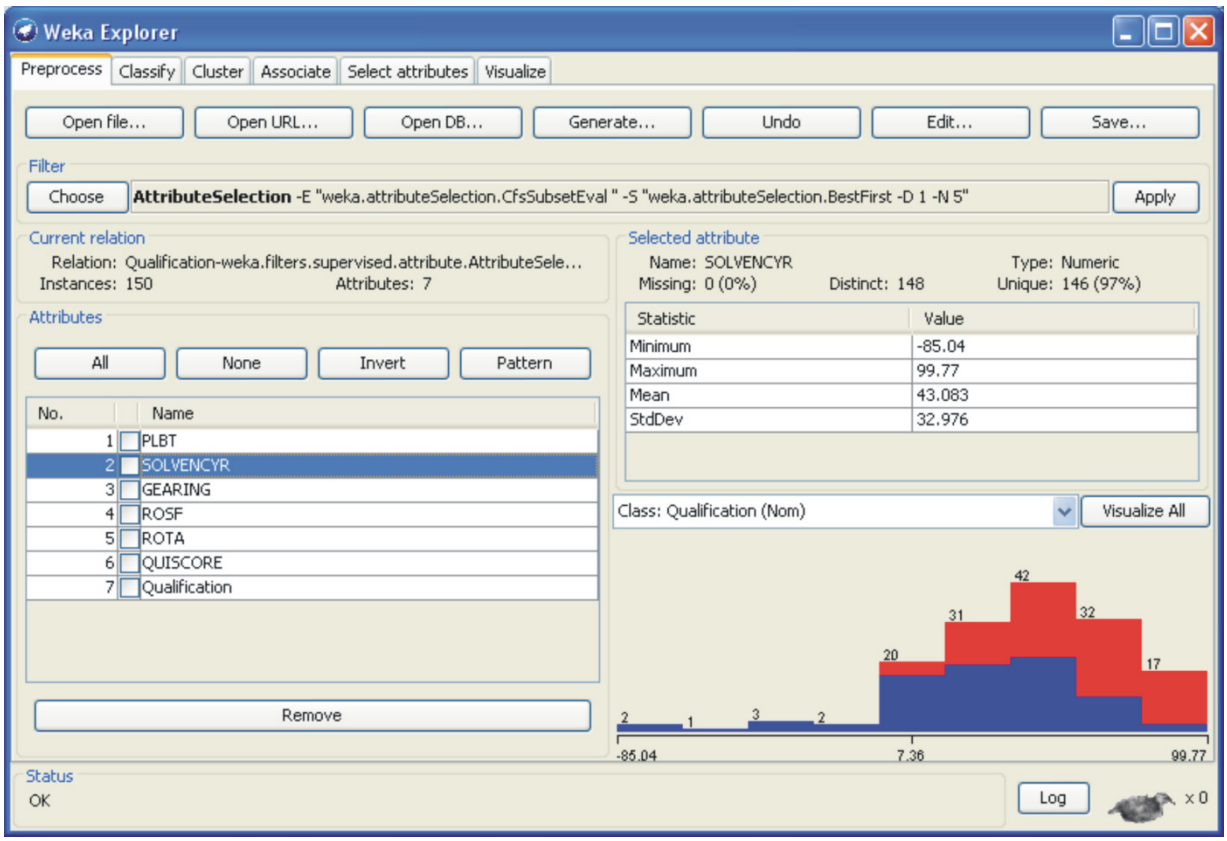


Εικόνα 13.5 *Επιλογή χαρακτηριστικών*



Εικόνα 13.6 *Καθορισμός μεθόδου επιλογής χαρακτηριστικών*

Ο χρήστης, πατώντας το κουμπί "Apply", εφαρμόζει τη μέθοδο. Το αποτέλεσμα παρουσιάζεται στην Εικόνα 13.7. Παρατηρούμε ότι πριν την εφαρμογή της μεθόδου υπήρχαν 19 πεδία, ενώ μετά την εφαρμογή της μεθόδου απέμειναν μόνο 7. Τα πεδία αυτά κρίθηκαν σημαντικά από τη μέθοδο CFS και θα χρησιμοποιηθούν στην περαιτέρω ανάλυση.



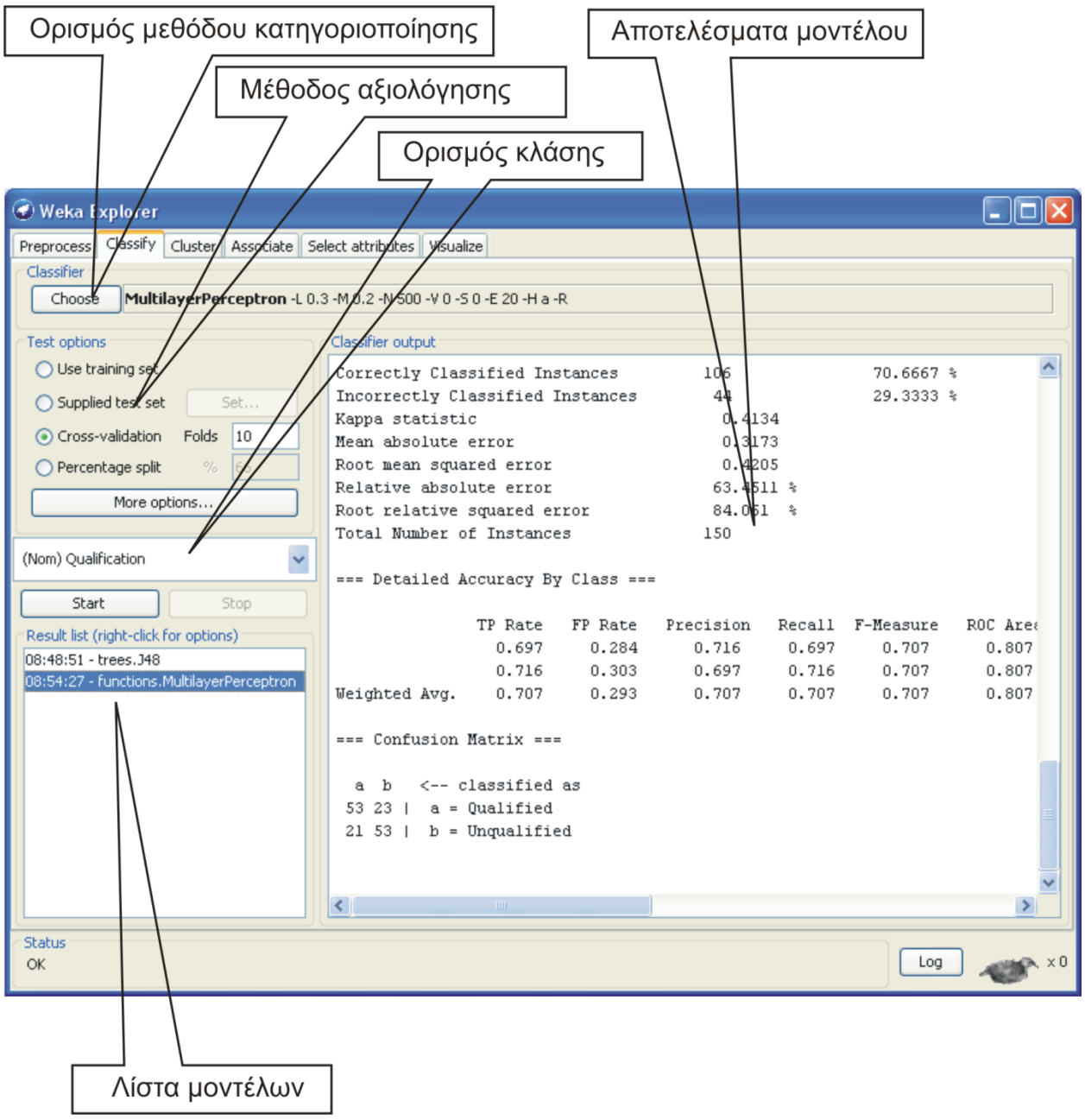
Εικόνα 13.7 *Επιλογή χαρακτηριστικών με τη μέθοδο CFS.*

13.3 Κατηγοριοποίηση

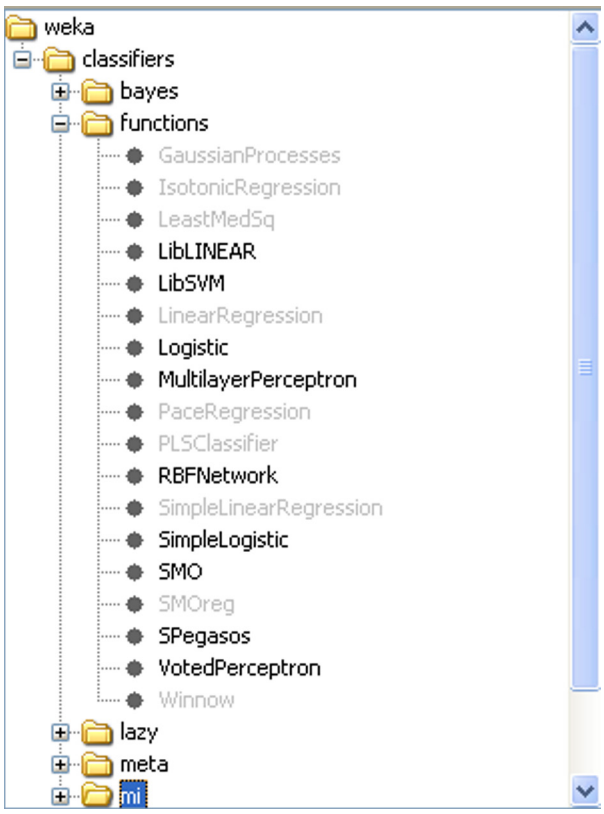
Το WEKA προσφέρει μια μεγάλη ποικιλία εργαλείων για κατηγοριοποίηση. Οι σχετικές εργασίες μπορούν να εκτελεστούν στο tab "Classify". Το tab Classify παρουσιάζεται στην Εικόνα 13.8.

Ο χρήστης αρχικά ορίζει τη μέθοδο κατηγοριοποίησης που θα εφαρμόσει. Η εργασία αυτή γίνεται πατώντας το κουμπί "Choose" στο πεδίο "Classifier". Το WEKA περιλαμβάνει μεγάλο αριθμό μεθόδων κατηγοριοποίησης. Οι μέθοδοι είναι ομαδοποιημένες σε κατηγορίες, οι οποίες παρουσιάζονται σε μορφή δένδρου. Το δένδρο των μεθόδων κατηγοριοποίησης παρουσιάζεται στην Εικόνα 13.9. Ορισμένες από τις κυριότερες μεθόδους κατηγοριοποίησης που περιλαμβάνονται είναι τα Μπαϋεσιανά Δίκτυα, οι Μηχανές Διανυσμάτων Υποστήριξης, η Λογιστική Παλινδρόμηση, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron και τα Δένδρα Αποφάσεων C4.5. Στο παράδειγμα της Εικόνας 13.8 έχει επιλεγεί ένα Νευρωνικό Δίκτυο τύπου MultiLayer Perceptron. Ιδιαίτερο ενδιαφέρον παρουσιάζει η κατηγορία "meta", στην οποία περιλαμβάνονται εργαλεία για τη δημιουργία σύνθετων κατηγοριοποιητών bagging και boosting, κατηγοριοποιητών ευαίσθητων στο κόστος, κατηγοριοποιητών που χρησιμοποιούν ανάλυση συστάδων κλπ.

Ο χρήστης, αφού επιλέξει μια μέθοδο κατηγοριοποίησης, μπορεί να ρυθμίσει τις παραμέτρους της. Η εργασία αυτή γίνεται κάνοντας κλικ στο όνομα της μεθόδου. Στην Εικόνα 13.10 παρουσιάζονται οι παράμετροι του Νευρωνικού Δικτύου. Στο πεδίο "About" παρέχονται σύντομες πληροφορίες για την εκάστοτε μέθοδο. Το κουμπί "More" ανοίγει ένα νέο παράθυρο, στο οποίο παρέχονται πρόσθετες πληροφορίες για τη μέθοδο, καθώς και διευκρινίσεις για την κάθε παράμετρο. Πιθανότατα ο χρήστης θα χρειαστεί αυτές της διευκρινίσεις για τη ρύθμιση των παραμέτρων. Για παράδειγμα, στο πεδίο "hiddenLayers" της Εικόνας 13.10 ορίζονται το πλήθος των κρυφών στρωμάτων και των κρυφών νευρώνων. Για κάθε κρυφό στρώμα αναγράφεται το πλήθος των νευρώνων του, και οι τιμές αυτές διαχωρίζονται με κόμμα για κάθε κρυφό στρώμα. Για παράδειγμα, η τιμή "10,4" δηλώνει την ύπαρξη δύο κρυφών στρωμάτων, όπου το πρώτο διαθέτει 10 νευρώνες και το δεύτερο 4. Στη θέση της κάθε τιμής μπορεί να μπεί ένα κωδικό σύμβολο. Για παράδειγμα, το σύμβολο "a", το οποίο υπάρχει στο παράδειγμα της Εικόνας 13.10, σημαίνει ότι το πλήθος των νευρώνων ισούται με το κλάσμα (πλήθος μεταβλητών εισόδου + πλήθος τιμών κλάσης) / 2. Τέτοιες χειριστικές λεπτομέρειες είναι ιδιαιτερότητες του WEKA, και ο χρήστης θα βρεί πολύτιμη σχετική καθοδήγηση μέσω του κουμπιού "More".



Εικόνα 13.8 *Κατηγοριοποίηση*

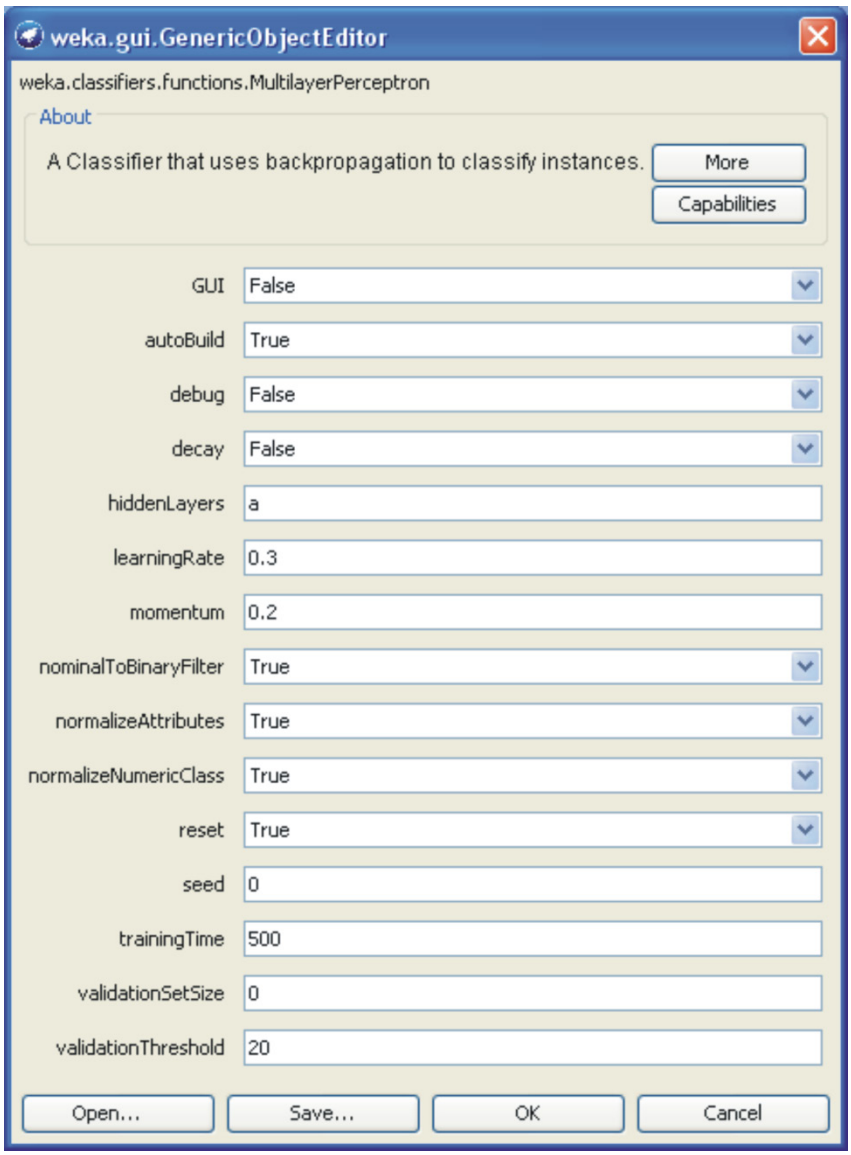


Εικόνα 13.9 *Μέθοδοι κατηγοριοποίησης*

Ο χρήστης, αφού ορίσει τη μέθοδο κατηγοριοποίησης που θα χρησιμοποιήσει, ορίζει το πεδίο της κλάσης και τη μέθοδο αξιολόγησης του κατηγοριοποιητή. Το WEKA προσφέρει τέσσερεις εναλλακτικές:

- Η επιλογή ″Use training set″ υπολογίζει τις επιδόσεις του μοντέλου, χρησιμοποιώντας το σύνολο εκπαίδευσης.
- Η επιλογή ″Supplied Test Set″ χρησιμοποιεί για επικύρωση ένα διαφορετικό σύνολο δεδομένων.
- Η επιλογή ″Cross-validation″ εφαρμόζει την ομόνυμη μέθοδο επικύρωσης. Ο χρήστης μπορεί να ορίσει το πλήθος των τμημάτων.
- Η επιλογή ″Percentage split″ εφαρμόζει τη μέθοδο holdout και διασπά το σύνολο των παρατηρήσεων σε υποσύνολο εκπαίδευσης και υποσύνολο επικύρωσης, σύμφωνα με τα ποσοστά που ορίζει ο χρήστης.

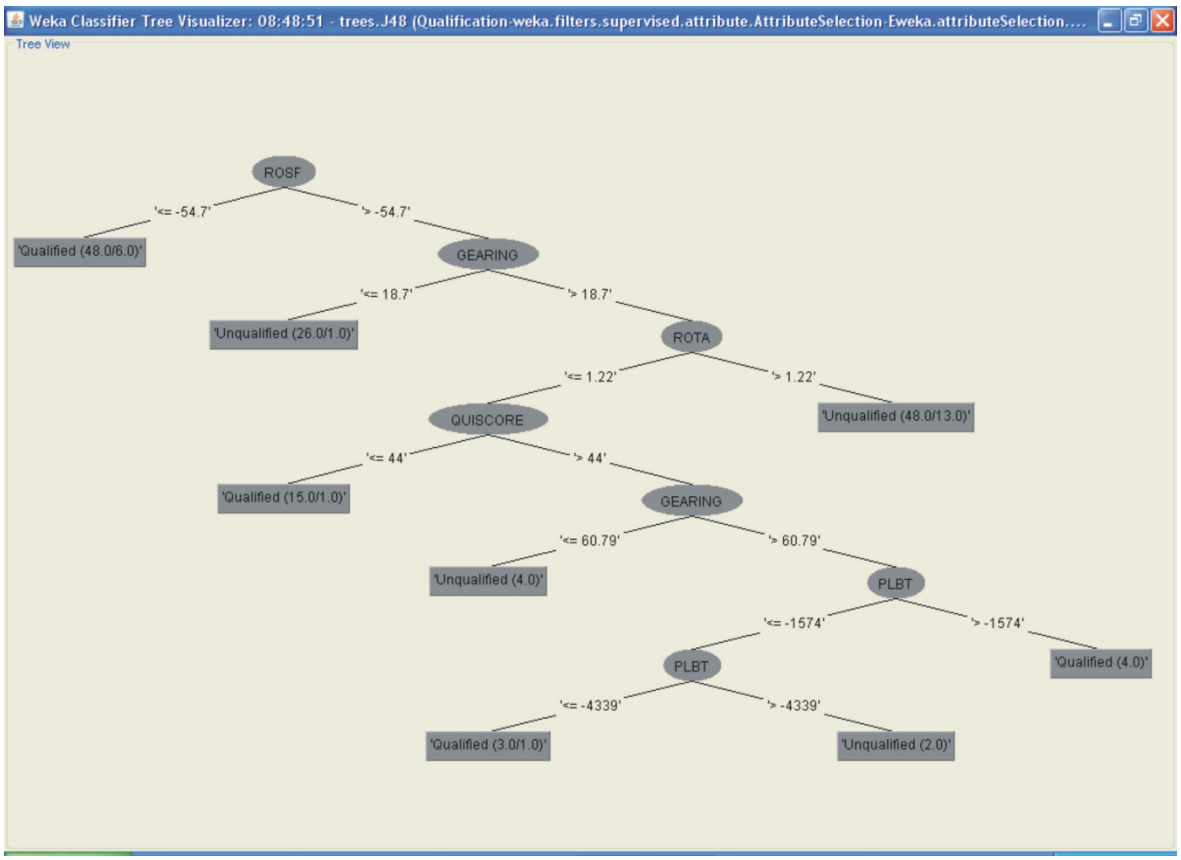
Αφού ολοκληρωθούν όλες οι προηγούμενες εργασίες, μπορεί να γίνει η εκπαίδευση και η αξιολόγηση του μοντέλου. Οι εργασίες αυτές εκτελούνται με το πάτημα του κουμπιού ″Start″. Μετά την ολοκλήρωση αυτών των εργασιών εμφανίζεται το μοντέλο στο πεδίο ″Results List″, στο κάτω και αριστερό μέρος του παραθύρου. Στο πεδίο ″Classifier output″ παρουσιάζονται τα αποτελέσματα του μοντέλου. Για κάθε κατηγοριοποιητή παρουσιάζονται το πλήθος των ορθών και εσφαλμένων προβλέψεων, πληροφορίες σχετικά με την αναλυτική ακρίβεια ανά κλάση, καθώς και ο πίνακας σύγχυσης (confusion matrix). Στο παράδειγμα της Εικόνας 13.8 βλέπουμε ότι το Νευρωνικό Δίκτυο προέβλεψε σωστά τις 106 παρατηρήσεις (ποσοστό 70,6667%). Αναλυτικότερα, προέβλεψε σωστά τις 53 από τις 76 ″Qualified″ εταιρείες (ποσοστό 69.7%) και τις 53 από τις 74 ″Unqualified″ εταιρείες (ποσοστό 71.6%). Τα στοιχεία αυτά παρουσιάζονται στο confusion matrix και στην αναλυτική ακρίβεια ανά κλάση, στη στήλη ″TP Rate″. Τα αποτελέσματα υπολογίστηκαν με τη μέθοδο επικύρωσης 10 fold cross validation.



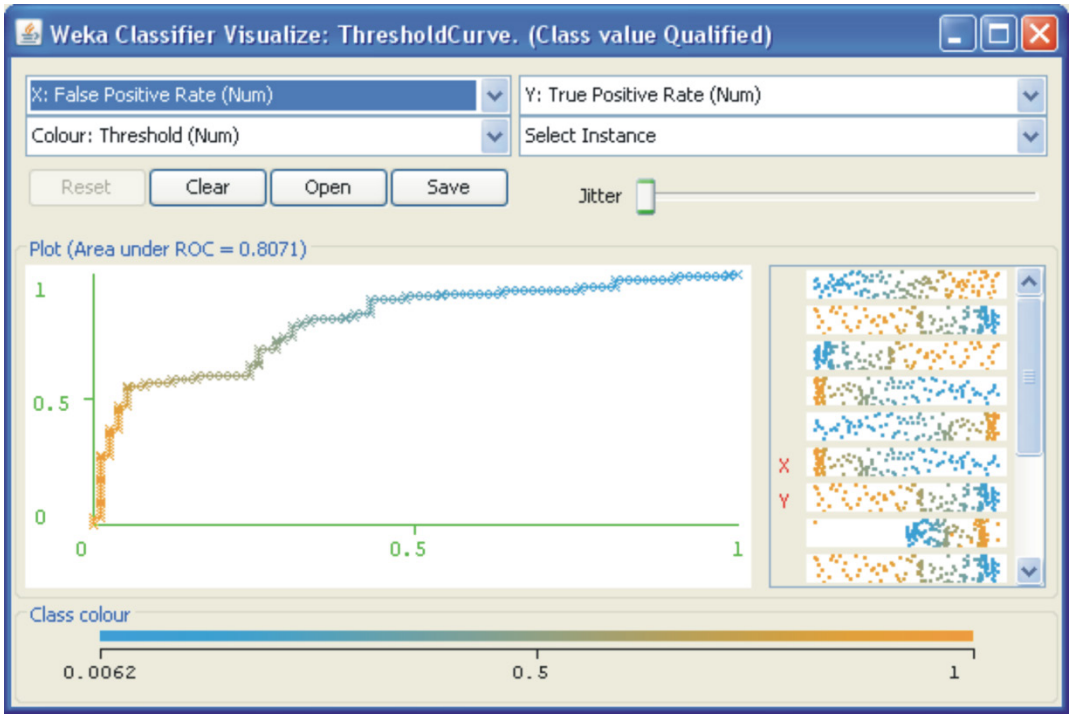
Εικόνα 13.10 *Παράμετροι Multilayer Perceptron.*

Εκτός από τα παραπάνω στοιχεία, τα οποία είναι κοινά για κάθε μέθοδο κατηγοριοποίησης, στο πεδίο ″Classifier output″ εμφανίζονται πρόσθετες πληροφορίες, οι οποίες διαφέρουν ανάλογα με τη μέθοδο κατηγοριοποίησης. Για παράδειγμα, εάν χρησιμοποιηθεί ένα Νευρωνικό Δίκτυο τότε εμφανίζονται τα βάρη των συνδέσεων, ενώ εάν χρησιμοποιηθεί ένα Δένδρο Αποφάσεων τότε εμφανίζεται η δομή του δένδρου. Ο χρήστης μπορεί να εφαρμόσει διαφορετικές μεθόδους και να πάρει αντίστοιχα μοντέλα ή να εφαρμόσει την ίδια μέθοδο με διαφορετική ρύθμιση παραμέτρων. Για κάθε εκτέλεση αλγορίθμου κατηγοριοποίησης ένα αντίστοιχο μοντέλο παρουσιάζεται στο πεδίο ″Results List″. Στο παράδειγμα της Εικόνας 13.8 έχουν εκπαιδευτεί ένα Δένδρο Αποφάσεων και ένα Νευρωνικό Δίκτυο. Κάνοντας κλικ σε ένα μοντέλο της λίστας παρουσιάζονται τα αποτελέσματα του στο πεδίο ″Classifier output″.

Κάνοντας δεξί κλικ σε ένα μοντέλο του πεδίου ″Results list″ ανοίγει το αντίστοιχο μενού συντομίας. Για ορισμένες μεθόδους, όπως τα Δένδρα Αποφάσεων ή τα Μπαϋεσιανά Δίκτυα, από το μενού συντομίας μπορεί να γίνει οπτική αναπαράσταση του μοντέλου. Στην Εικόνα 13.11 παρουσιάζεται το Δένδρο Αποφάσεων. Από το μενού συντομίας ο χρήστης μπορεί να προβάλει και τις καμπύλες ROC μοντέλου. Η σχετική επιλογή του μενού είναι ″Visualize threshold curve″. Η καμπύλη ROC του Νευρωνικού Δικτύου παρουσιάζεται στην Εικόνα 13.12.



Εικόνα 13.11 *Οπτική αναπαράσταση Δένδρου Αποφάσεων*



Εικόνα 13.12 *Καμπύλη ROC στο WEKA*

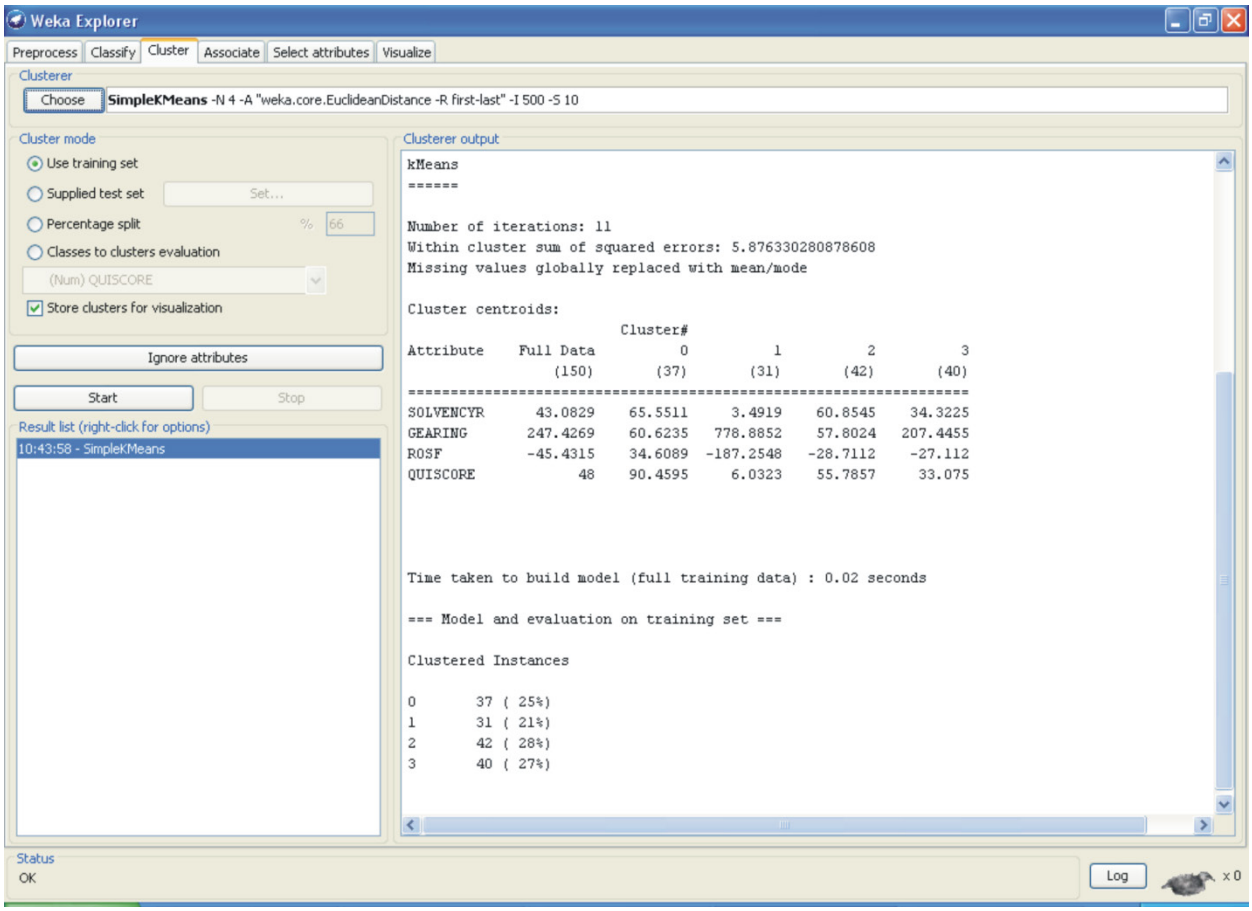
13.4 Ανάλυση Συστάδων

Το WEKA παρέχει εργαλεία και για ανάλυση συστάδων. Εργασίες ανάλυσης συστάδων εκτελούνται στο

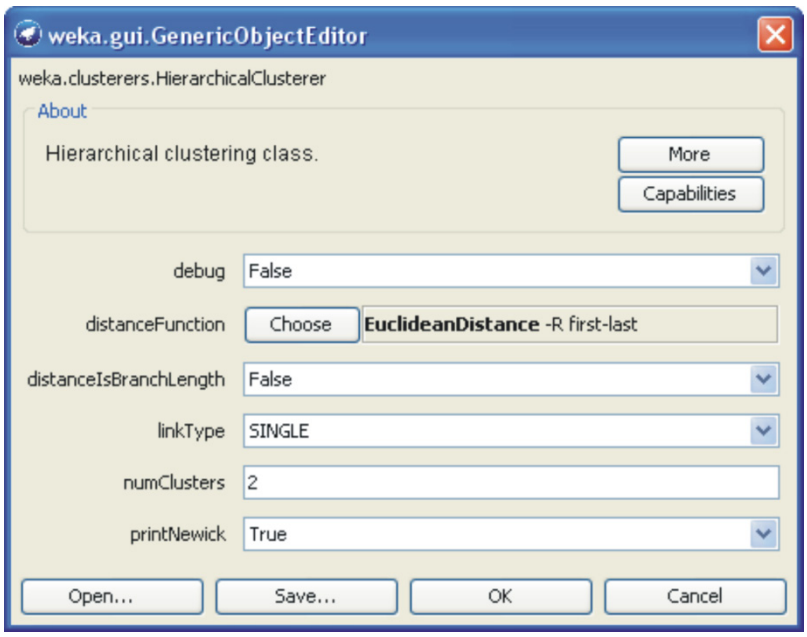
tab ″Cluster″. Το tab ″Cluster″ παρουσιάζεται στην Εικόνα 13.13. Αρχικά, ο χρήστης επιλέγει μέθοδο ΑΣ, κάνοντας κλικ στο κουμπί ″Choose″ του πεδίου ″Clusterer″. Το WEKA περιλαμβάνει αρκετές μεθόδους ΑΣ, αν και αισθητά λιγότερες από τις διαθέσιμες μεθόδους κατηγοριοποίησης. Ανάμεσα στους αλγορίθμους που διατίθονται περιλαμβάνονται ο k-Means, η Συσσωρευτική Ιεραρχική ΑΣ, η Expected Maximazation (EM), και ο DBSCAN.

Ο χρήστης, αφού επιλέξει μέθοδο ΑΣ, μπορεί να ρυθμίσει τις παραμέτρους της κάνοντας κλικ στο όνομα της μεθόδου. Στην Εικόνα 13.14 παρουσιάζεται το παράθυρο ρύθμισης παραμέτρων για τη μέθοδο της Ιεραρχικής ΑΣ. Στο συγκεκριμένο παράθυρο, ο χρήστης μπορεί να ορίσει τη συνάρτηση απόστασης, τον τύπο σύνδεσης (link Type) και το πλήθος των συστάδων. Μετα τον καθορισμό των παραμέτρων, ο χρήστης μπορεί να εκτελέσει τον αλγόριθμο κάνοντας κλικ στο κουμπί ″Start″. Για κάθε εκτέλεση αλγορίθμου προστίθεται μια εγγραφή στο πεδίο ″Result List″. Τα αντίστοιχα αποτελέσματα παρουσιάζονται στο πεδίο ″Clusterer Output″.

Στο παράδειγμα της Εικόνας 13.13 εφαρμόστηκε Ανάλυση Συστάδων στα δεδομένα των 150 επιχειρήσεων. Τα πεδία που συμπεριλήφθηκαν στην ανάλυση είναι ο αριθμοδείκτης αξιοπιστίας Solvency Ratio, ο αριθμοδείκτης δανειοληψίας Gearing, ο αριθμοδείκτης κερδοφορίας Return on Sharholders Funds – ROSF και ο δείκτης πιστοληπτικής ικανότητας Quiscore. Η μέθοδος ΑΣ η οποία εφαρμόστηκε είναι η k-Means. Προκαθορίστηκε να δημιουργηθούν τέσσερεις συστάδες. Σύμφωνα με τα αποτελέσματα, πραγματοποιήθηκαν 11 επαναλήψεις και δημιουργήθηκαν τέσσερεις συστάδες. Η πρώτη συστάδα περιλαμβάνει 37 παρατηρήσεις, η δεύτερη 31, η τρίτη 42 και η τέταρτη 40. Για κάθε συστάδα αναφέρονται οι συντεταγμένες του αντίστοιχου κεντρικού σημείου (centroid) για τους τέσσερεις αριθμοδείκτες.



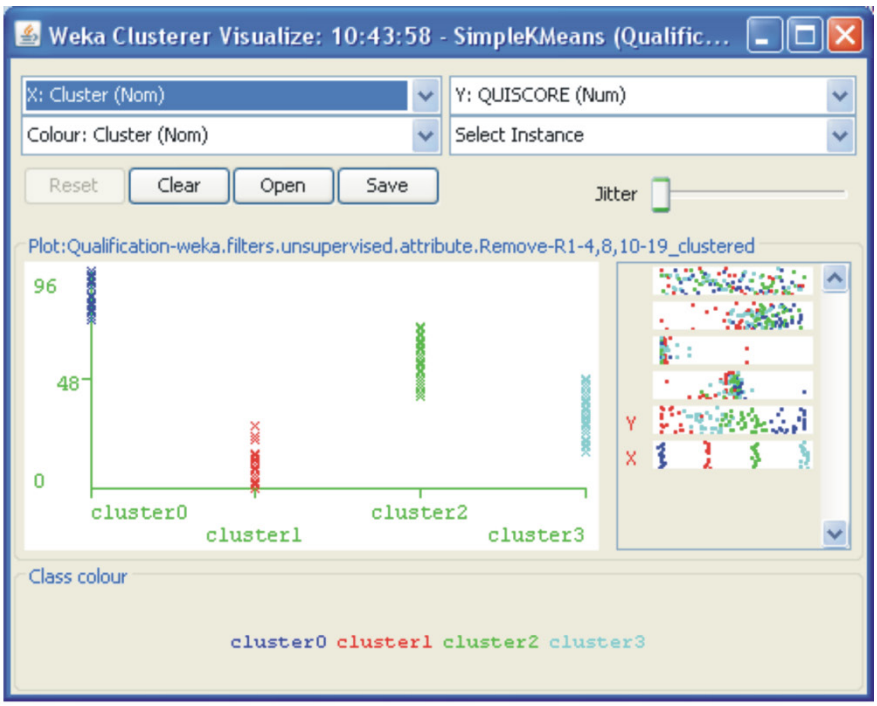
Εικόνα 13.13 *Ανάλυση Συστάδων στο WEKA*



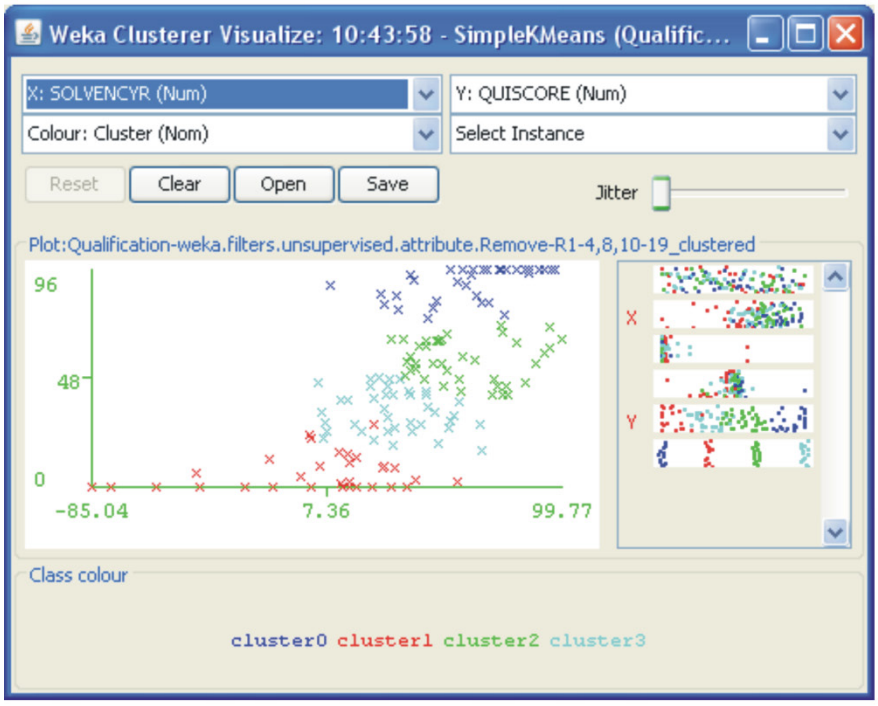
Εικόνα 13.14 *Ορισμός παραμέτρων Ιεραρχικής ΑΣ*

Κάνοντας δεξί κλικ σε μια εγγραφή του πεδίου ″Results List″, ανοίγει το αντίστοιχο μενού συντομίας. Εάν η μέθοδος ΑΣ που χρησιμοποιήθηκε είναι Ιεραρχική, ο χρήστης μπορεί να προβάλει το δενδρόγραμμα. Επίσης, από το μενού συντομίας ο χρήστης μπορεί να λάβει την οπτική αναπαράσταση της κατανομής των παρατηρήσεων σε συστάδες.

Στην Εικόνα 13.15 παρουσιάζεται οπτικά η κατανομή των επιχειρήσεων σε συστάδες. Στο τμήμα Α) στον άξονα *x* βρίσκονται οι τέσσερεις συστάδες και στον άξονα *y* η μεταβλητή Quiscore. Είναι εμφανές ότι στη συστάδα 1 βρίσκονται οι επιχειρήσεις με μικρές τιμές Quiscore, στη συστάδα 3 οι επιχειρήσεις με μεσαίες προς μικρές τιμές Quiscore, στη συστάδα 2 οι επιχειρήσεις με μεσαίες προς μεγάλες τιμές Quiscore και στη συστάδα 0 οι επιχειρήσεις με μεγάλες τιμές Quiscore. Στο τμήμα Β) στον άξονα *x* βρίσκεται η μεταβλητή SolvencyR(atio) και στον άξονα *y* η μεταβλητή Quiscore. Οι τέσσερεις συστάδες συμβολίζονται με διαφορετικά χρώματα. Είναι εμφανής ο σχηματισμός των συστάδων ανάλογα με τις τιμές τους στις δύο μεταβλητές.



A)

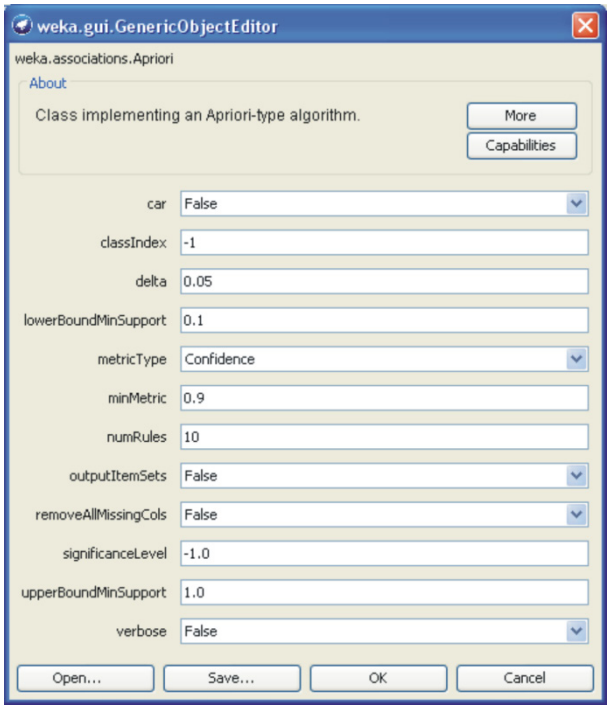


B)

Εικόνα 13.15 *Οπτική αναπαράσταση συστάδων*

13.5 Κανόνες Συσχέτισης

Το WEKA περιλαμβάνει και αλγορίθμους για την εξόρυξη Κανόνων Συσχέτισης. Ο χρήστης μπορεί να βρει τους σχετικούς αλγορίθμους στο tab "Associate". Περιλαμβάνονται ορισμένοι αλγόριθμοι, μεταξύ των οποίων και ο Apriori. Τα δεδομένα πρέπει να είναι διακριτά. Με την εφαρμογή του Apriori μπορούν να βρεθούν κανόνες, οι οποίοι υπερβαίνουν τις ελάχιστες τιμές υποστήριξης και εμπιστοσύνης.



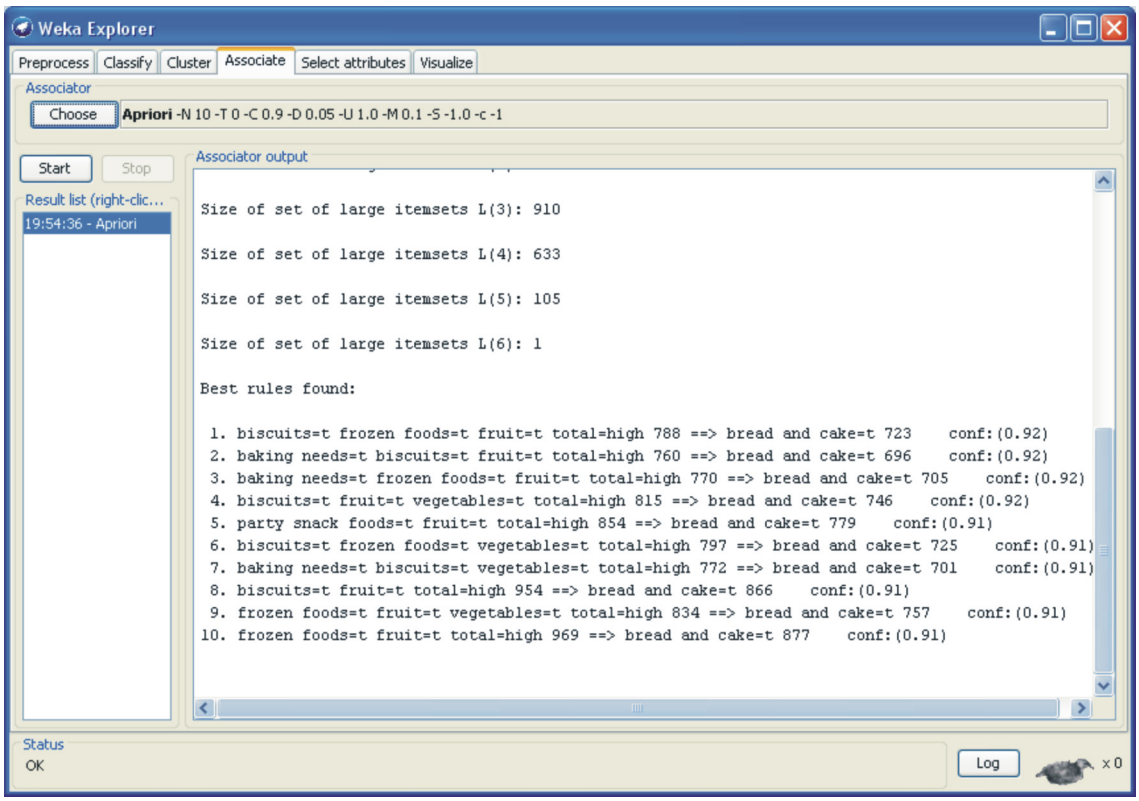
Εικόνα 13.16 *Παράμετροι Apriori.*

Στην Εικόνα 13.16 παρουσιάζεται το παράθυρο ρύθμισης παραμέτρων του Apriori. Στην υλοποίηση του Apriori, η οποία υπάρχει στο WEKA, εκτελείται μια επαναλαμβανόμενη διαδικασία, όπου ο αλγόριθμος μειώνει σταδιακά την τιμή της υποστήριξης, μέχρι να βρεί έναν προκαθορισμένο αριθμό κανόνων. Στο πεδίο "upperBoundMinSupport" ορίζεται η τιμή εκκίνησης για την ελάχιστη υποστήριξη, στο πεδίο "lowerBoundMinSupport" ορίζεται η μικρότερη δυνατή τιμή για την ελάχιστη υποστήριξη και στο πεδίο "delta" ορίζεται το βήμα μείωσης. Το πλήθος των κανόνων που θα εξαχθούν ορίζεται στο πεδίο "numRules". Προβλέπονται διάφορες μετρικές (πεδίο "metricType") για την εξαγωγή των κανόνων, με προτεινόμενη επιλογή την εμπιστοσύνη. Στο πεδίο "minMetric" ορίζεται η ελάχιστη τιμή της μετρικής.

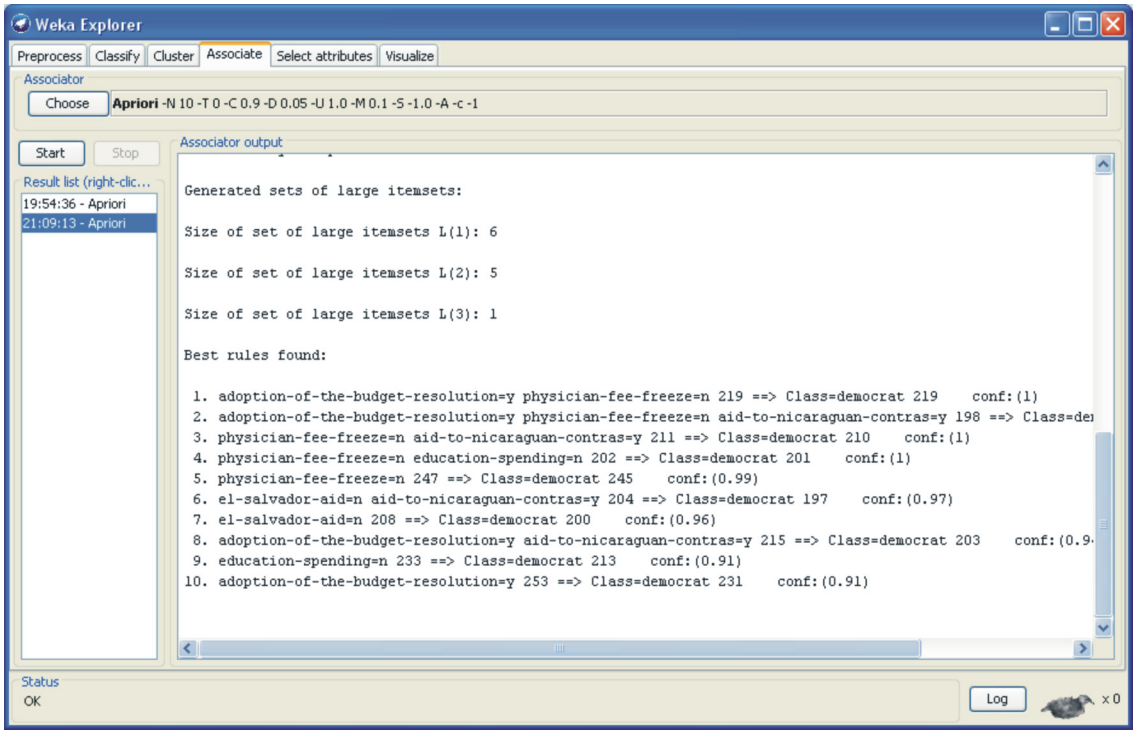
Στην Εικόνα 13.17 απεικονίζονται τα αποτελέσματα εξόρυξης Κανόνων Συσχέτισης με εφαρμογή του Apriori και χρήση του συνόλου δεδομένων "supermarket", το οποίο διατίθεται από την ιστοσελίδα του WEKA. Παρουσιάζονται οι δέκα κανόνες οι οποίοι έχουν εξαχθεί, καθώς και οι τιμές υποστήριξης και εμπιστοσύνης. Οι κανόνες συσχετίζουν την αγορά ορισμένων προϊόντων με την αγορά κάποιων άλλων.

Η συγκεκριμένη υλοποίηση του Apriori προβλέπει τη δυνατότητα εξόρυξης κανόνων προκαθορισμένης μορφής, και ειδικότερα κανόνων στο δεξιό τμήμα των οποίων θα βρίσκεται ένα προκαθορισμένο πεδίο. Κανόνες αυτής της μορφής είναι ιδιαίτερα χρήσιμοι για δεδομένα στα οποία υπάρχει γνώρισμα κλάσης. Οι κανόνες οι οποίοι θα προκύψουν θα έχουν στο αριστερό τους τμήμα έναν συνδυασμό τιμών κάποιων γνωρισμάτων και στο δεξιό τους τμήμα μια τιμή κλάσης. Ο χρήστης, για να εξορύξει κανόνες τέτοιας μορφής, πρέπει να θέσει στο παράθυρο ορισμού παραμέτρων (Εικόνα 13.16) και στο πεδίο "car" την τιμή "True". Επίσης, στο πεδίο "classIndex" ορίζει τον αριθμό του γνωρίσματος κλάσης. Η κωδική τιμή "-1" σημαίνει ότι το γνώρισμα κλάσης είναι το τελευταίο πεδίο.

Στην Εικόνα 13.18 εμφανίζονται κανόνες οι οποίοι στο δεξιό τους τμήμα έχουν ένα προκαθορισμένο πεδίο. Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι το "Vote", το οποίο διατίθεται από την ιστοθέση του WEKA. Το συγκεκριμένο σύνολο δεδομένων περιέχει αποτελέσματα ψηφοφοριών από τη βουλή των ΗΠΑ. Κάθε γραμμή αντιστοιχεί σε έναν βουλευτή (representative) και κάθε στήλη σε ένα νομοσχέδιο. Οι δυνατές τιμές δεδομένων είναι "y" και "n", και καταγράφουν εάν ο βουλευτής υπερψήφισε ή καταψήφισε το εκάστοτε νομοσχέδιο. Στην τελευταία στήλη καταγράφεται η κομματική τοποθέτηση του βουλευτή (democrat/republican). Σύμφωνα με τον πρώτο κανόνα, 219 βουλευτές υπερψήφισαν τον νόμο "adoption-of-the-budget-resolution" και καταψήφισαν τον νόμο "physician-fee-freeze". Όλοι αυτοί οι βουλευτές ήταν "δημοκρατικοί" (confidence = 1).



Εικόνα 13.17 *Κανόνες Συσχέτισης*



Εικόνα 13.18 *Κανόνες Συσχέτισης με γνώρισμα κλάσης.*

13.6 Επιλογή Χαρακτηριστικών

Στο tab ″Select Attributes″ του WEKA περιλαμβάνονται εργαλεία για την επιλογή σημαντικών χαρακτηριστικών. Οι μέθοδοι επιλογής χαρακτηριστικών του WEKA αποτελούνται από δύο τμήματα:

- Μια μέθοδο αναζήτησης. Περιλαμβάνονται μέθοδοι πρόσθιας αναζήτησης, οπίσθιας αναζήτησης, γενετικοί αλγόριθμοι κλπ.
- Μια μέθοδο αξιολόγησης. Διατίθεται μια σημαντική ποικιλία μεθόδων, η οποία περιλαβάνει την CFS, στατιστική Chi-Square, ευαίσθητες στο κόστος μεθόδους, wrappers, κριτήριο Gain Ratio, χρήση κατηγοριοποιητή SVM, ανάλυση κυρίων συνιστωσών κλπ.

Ο χρήστης μπορεί να συνδυάσει με διάφορους τρόπους μεθόδους αναζήτησης και μεθόδους αξιολόγησης.

Στην Εικόνα 13.19 παρουσιάζεται το παράθυρο επιλογής χαρακτηριστικών. Στα πεδία ″Attribute Evaluator″ και ″Search Method″, ο χρήστης ορίζει μια μέθοδο αξιολόγησης χαρακτηριστικών και μια μέθοδο αναζήτησης αντιστοίχως. Κάνοντας κλικ στο όνομα της μεθόδου, ανοίγει ένα παράθυρο στο οποίο παρέχονται πληροφορίες για τη μέθοδο, καθώς και δυνατότητες ρύθμισης των παραμέτρων της. Για παράδειγμα, εάν επιλεγεί μια μέθοδος αξιολόγησης wrapper, στο παράθυρο παραμέτρων ορίζεται η βασική μέθοδος κατηγοριοποίησης που θα χρησιμοποιηθεί για την αξιολόγηση των χαρακτηριστικών.

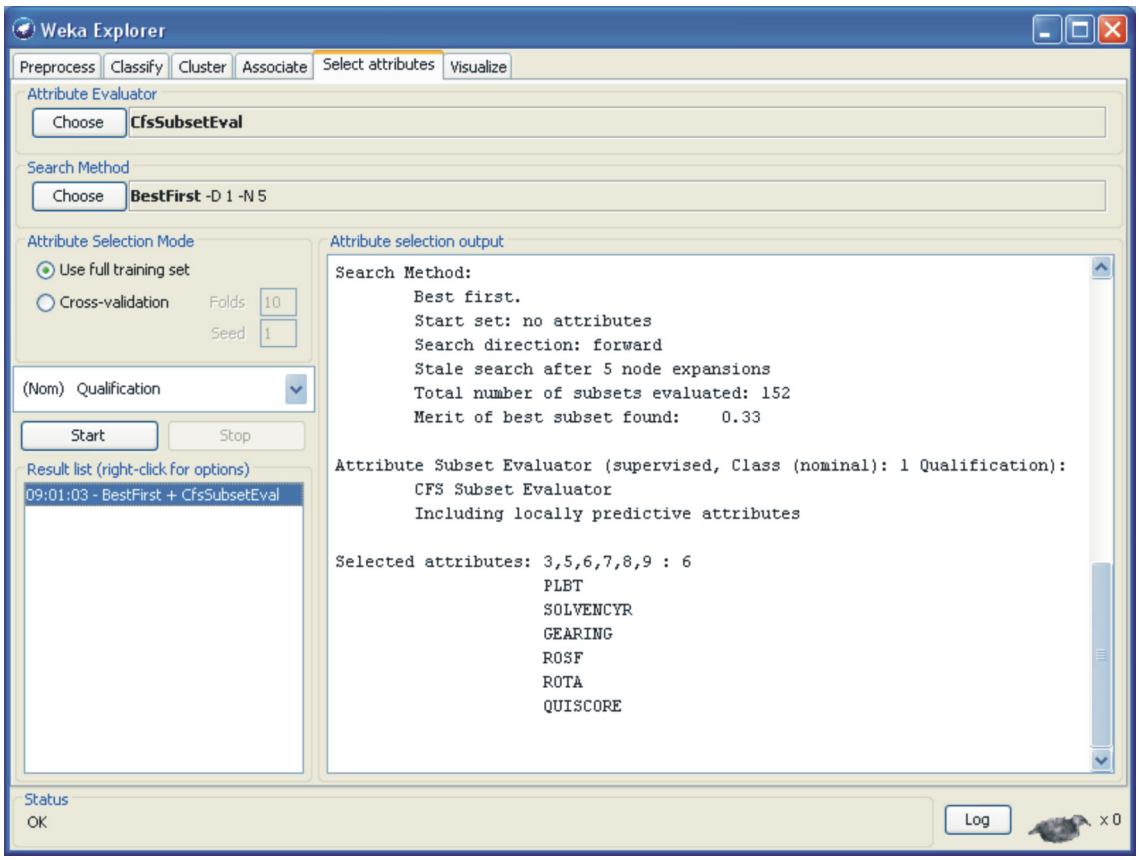
Στο παράδειγμα της Εικόνας 13.19 εφαρμόστηκε η μέθοδος αξιολόγησης CFS σε συνδυασμό με τη μέθοδο αναζήτησης BestFirst. Τα δεδομένα που χρησιμοποιηθήκαν αφορούν τις 150 επιχειρήσεις και το αποτέλεσμα του εξωτερικού τους ελέγχου. Συνολικά αξιολογήθηκαν 152 υποσύνολα χαρακτηριστικών. Το υποσύνολο χαρακτηριστικών που επιλέχθηκε περιλαμβάνει τους αριθμοδείκτες PLBT (Profit – Loss Before Taxation), SolvencyR(atio), Gearing, ROSF (Return on Shareholders' Funds), ROTA (Return on Total Assest) και Quiscore.

Ο χρήστης μπορεί να πειραματιστεί με διάφορες μεθόδους. Στο πεδίο ″Attribute selection output″ παρουσιάζονται τα αποτελέσματα της κάθε μεθόδου. Ο χρήστης, αφού επιλέξει τη μέθοδο επιλογής χαρακτηριστικών που θα χρησιμοποιήσει τελικά, μπορεί να μεταβεί στο tab ″Preprocess″ (Εικόνα 13.5), να ορίσει τη μέθοδο στο πεδίο ″Filter″ και να την εφαρμόσει κάνοντας κλικ στο κουμπί ″Apply″. Εναλλακτικά, μπορεί να μεταβεί στο tab ″Preprocess″ και να διαγράψει τα πεδία τα οποία δεν βρέθηκαν σημαντικά, επιλέγοντας τα και κάνοντας κλικ στο κουμπί ″Remove″.

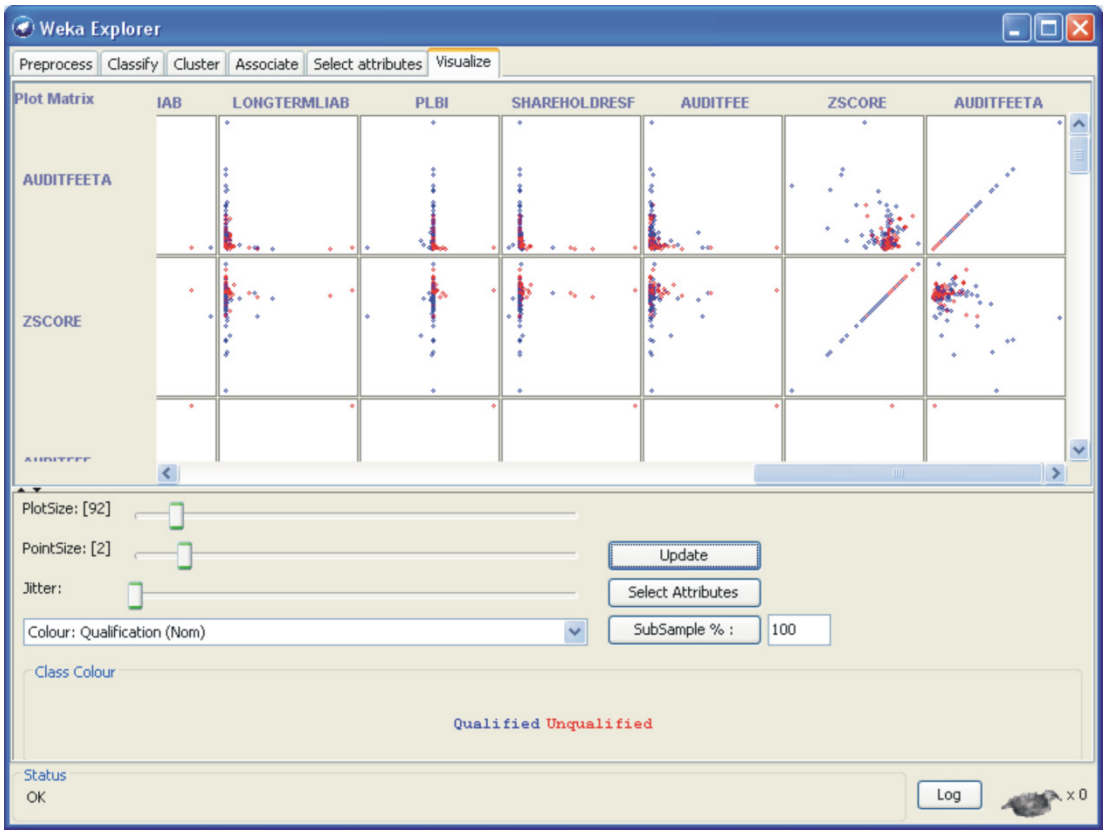
13.8 Οπτικοποίηση

Στο τελευταίο tab του WEKA Explorer παρέχονται εργαλεία οπτικοίησης των δεδομένων. Η οπτικοποίηση είναι πολύ χρήσιμη στην πράξη, καθώς επιτρέπει στον χρήστη να κατανοήσει με εύκολο και γρήγορο τρόπο τη διασπορά των παρατηρήσεων. Στην Εικόνα 13.20 παρουσιάζεται το tab ″Visualize″. Το παράθυρο αυτό περιέχει έναν πίνακα διαγραμμάτων διασποράς για όλα τα δυνατά ζεύγη των χαρακτηριστικών των δεδομένων. Στο παράδειγμα που απεικονίζεται χρησιμοποιήθηκαν τα δεδομένα των 150 επιχειρήσεων. Εάν στα δεδομένα υπάρχει γνώρισμα κλάσης, όπως στα δεδομένα του παραδείγματος, ο χρήστης ορίζει το πεδίο κλάσης και οι παρατηρήσεις χρωματίζονται ανάλογα. Στο παράδειγμα της Εικόνας 13.20 οι Qualified εταιρείες χρωματίζονται με μπλε χρώμα και οι Unqualified με κόκκινο. Επιπλέον, ο χρήστης μπορεί να αλλάξει το μέγεθος του πίνακα διαγραμμάτων και το μέγεθος των σημείων, καθώς και να επιλέξει γνωρίσματα ή/και υποσύνολο των παρατηρήσεων.

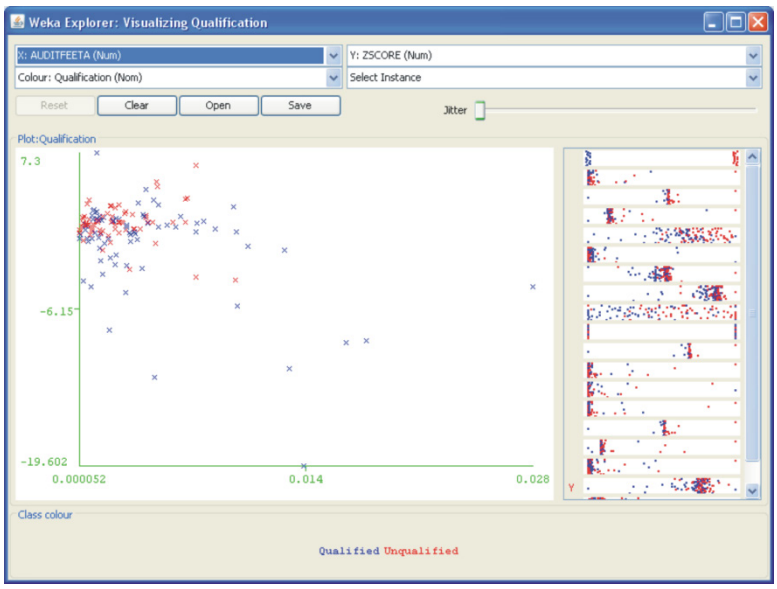
Ο χρήστης, κάνοντας κλικ σε ένα διάγραμμα, μπορεί να το προβάλει σε ξεχωριστό παράθυρο. Στην Εικόνα 13.21 παρουσιάζεται το παράθυρο για τις μεταβλητές AuditFeeTA (Audit Fees to Total Assets) και ZScore (Altman's). Ο χρήστης μπορεί να αλλάξει τις μεταβλητές των αξόνων X και Y από τα αντίστοιχα πεδία και να ορίσει το γνώρισμα κλάσης, ώστε οι παρατηρήσεις να χρωματιστούν ανάλογα. Στο δεξιό τμήμα του παράθυρου παρουσιάζονται οι κατανομές των παρατηρήσεων για σταθερή μεταβλητή στον άξονα Y και διάφορες μεταβλητές στον άξονα X.



Εικόνα 13.19 *Επιλογή χαρακτηριστικών*

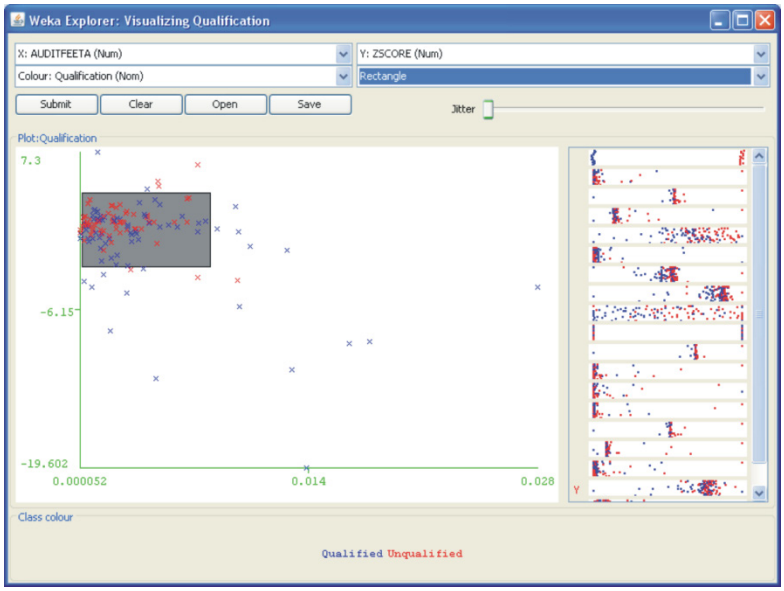


Εικόνα 13.20 *Πίνακας διαγραμμάτων διασποράς*



Εικόνα 13.21 *Διάγραμμα Διασποράς*

Ο χρήστης μπορεί να κάνει ζουμ σε ένα τμήμα του διαγράμματος. Αρχικά πρέπει να ορίσει στο πεδίο ″Select Instance″ την επιλογή ″Rectangle″ και στη συνέχεια να καθορίσει το τμήμα του διαγράμματος, όπως φαίνεται στην Εικόνα 13.22. Ακολούθως, με το πάτημα του κουμπιού ″Submit″ εμφανίζεται το επιλεγμένο τμήμα σε μεγέθυνση.



Εικόνα 13.22. *Ζουμ σε διάγραμμα διασποράς.*

13.9 Άλλες πηγές για το WEKA

Το WEKA είναι ένα πολύ δημοφιλές λογισμικό μηχανικής μάθησης και εξόρυξης δεδομένων. Ως αποτέλεσμα αυτού του γεγονότος, υπάρχει σημαντικός αριθμός βιβλίων, αλλά κυρίως ηλεκτρονικών πηγών, όπως παρουσιάσεων και tutorials, που αναφέρονται στο WEKA. Το κυριότερο βιβλίο είναι το Witten et al. (2011), το οποίο προέρχεται από την επιστημονική ομάδα που ανέπτυξε το WEKA. Επίσης το Kaluza (2013) ασχολείται αποκλειστικά με το WEKA. Συντομότερες παρουσιάσεις του WEKA γίνονται στο Maimon and Rokach (2010), στο Puthran and Shah (2012), στο Rochester (2013), στο Bouman and Van Dongen (2009) και στο Larose (2006).

Εκτός των βιβλίων, υπάρχει ένας πολύ μεγάλος αριθμός ηλεκτρονικών πηγών για την παρουσίαση του

WEKA και την εκπαίδευση χρηστών σε αυτό. Σε αυτές τις ηλεκτρονικές πηγές περιλαμβάνονται ιστοσελίδες, παρουσιάσεις σε μορφή αρχείων Power Point και pdf, καθώς και εκπαιδευτικά βίντεο, διαθέσιμα κυρίως μέσω του YouTube. Επιλεκτικά αναφέρουμε τις ιστοσελίδες Weka 3- Data Mining with Open Source Machine Learning Software in Java (Cs.waikato.ac.nz, 2015), IBM developerWorks : Open Source : Technical Library (Ibm.com, 2015) της IBM, WEKA Tutorial (Cs.utexas.edu, 2015)του Πανεπιστημίου του Τέξας, Data mining techniques using WEKA (Slideshare.net, 2012) στο δίκτυο slideshare και Technology Forge – WEKA Tutorials (Technologyforge.net, 2015). Στην ιστοσελίδα 100 Best WEKA Tutorial Videos (Meta-guide.com, 2015) ο ενδιαφερόμενος αναγνώστης μπορεί να βρει πολλά εκπαιδευτικά βίντεο για το WEKA. Στο διαδίκτυο υπάρχουν πολλές πρόσθετες ηλεκτρονικές πηγές για το WEKA, εύκολα προσβάσιμες μέσω του Google.

13.10 Ελεύθερα λογισμικά Επιχειρηματικής Ευφυΐας και Εξόρυξης Δεδομένων

Το WEKA δεν είναι το μοναδικό ελεύθερο λογισμικό Εξόρυξης Δεδομένων. Στο σημείο αυτό θα επιχειρήσουμε μια επιγραμματική παρουσίαση άλλων ελεύθερων λογισμικών Επιχειρηματικής Ευφυΐας και Εξόρυξης Δεδομένων. Ο ενδιαφερόμενος αναγνώστης μπορεί να απευθυνθεί στις σχετικές ιστοσελίδες για πρόσθετες πληροφορίες και για να προμηθευτεί το λογισμικό που επιθυμεί.

Στην κατηγορία των ελεύθερων λογισμικών Επιχειρηματικής Ευφυΐας, μερικές από τις πιο γνωστές περιπτώσεις, σε τυχαία σειρά παρουσίασης, είναι οι ακόλουθες:

- IBM Watson Analytics. Η IBM προσφέρει μια ελεύθερη και βασισμένη στο νέφος έκδοση του Watson Analytics με σημαντικές δυνατότητες. Το λογισμικό επιτρέπει προεπεξεργασία δεδομένων, προγνωστική ανάλυση, οπτικοποίηση και δημιουργία αναφορών. Υπάρχει περιορισμός στον όγκο των δεδομένων που μπορεί να χειριστεί.
- Microsoft Power BI. Λογισμικό νέφους από την Microsoft. Επιτρέπει ενοποίηση δεδομένων από πολλές πηγές, διεξαγωγή αναλύσεων με εύχρηστα εργαλεία drag-and-drop και δημιουργία εξατομικευμένων ταμπλό. Υπάρχει περιορισμός στον όγκο των δεδομένων.
- SAP Lumira Cloud. Όπως υποδηλώνει και το όνομα, πρόκειται για λογισμικό νέφους. Διαθέτει αυξημένες δυνατότητες οπτικοποίησης και συνεργασίας. Επίσης, προσφέρει δωρεάν αποθηκευτικό χώρο 1 GB για τα δεδομένα.
- Pentaho Community Edition. Λογισμικό ανοικτού κώδικα, το οποίο περιλαμβάνει δυνατότητες ETL, OLAP, εργαλεία Εξόρυξης Δεδομένων, δημιουργία αναφορών και ταμπλό, διαχείριση μεταδεδομένων κλπ.
- Jaspersoft. Εταιρεία λογισμικού Επιχειρηματικής Ευφυΐας, η οποία προσφέρει και ελεύθερες εκδόσεις των λογισμικών της. Τα λογισμικά επιτρέπουν τη διεξαγωγή αναλύσεων, τη δημιουργία αναφορών και την αναλυτική επεξεργασία (OLAP).
- Jedox Base Business Intelligence. Ένα ακόμα λογισμικό ανοικτού κώδικα, το οποίο παρέχει δυνατότητες OLAP. Ένα σημαντικό χαρακτηριστικό του είναι ότι προσφέρει πρόσθετα (Add-Ins), τα οποία επεκτείνουν τις δυνατότητες του Excel.
- SpagoBI. Ελεύθερο λογισμικό για αναλυτική επεξεργασία (OLAP) και δημιουργία αναφορών και ταμπλό. Διαθέτει αυξημένες δυνατότητες οπτικής ανάλυσης των δεδομένων.
- KNIME. Λογισμικό Επιχειρηματικής Ευφυΐας ανοικτού κώδικα με αυξημένες δυνατότητες Εξόρυξης Δεδομένων και Μηχανικής Μάθησης. Διαθέτει εργαλεία για εργασίες ETL και προεπεξεργασία δεδομένων.
- Tableau Public. Η Tableau είναι μια αναγνωρισμένη εταιρεία λογισμικού επιχειρηματικής ευφυΐας. Η έκδοση Tableau Public διατίθεται ελεύθερα και διαθέτει εξελιγμένα εργαλεία οπτικοποίησης, δημιουργίας ταμπλό και επικοινωνίας της πληροφορίας.

Πέραν των Λογισμικών Επιχειρηματικής Ευφυΐας, τα οποία είναι ειδικά σχεδιασμένα για επιχειρήσεις, προσφέρονται και πολλά ελεύθερα λογισμικά Εξόρυξης Δεδομένων. Αρκετά από αυτά τα λογισμικά διατίθενται σε μορφή υποπρογραμμάτων και η χρήση τους απαιτεί ικανότητες προγραμματισμού. Ωστόσο, ορισμένα διαθέτουν γραφικό περιβάλλον εργασίας και μπορούν να χρησιμοποιηθούν από τελικούς χρήστες. Το WEKA είναι μια τέτοια περίπτωση, αλλά δεν είναι η μοναδική:

- Το Orange είναι λογισμικό εξόρυξης δεδομένων και μηχανικής μάθησης. Η χρήση του είναι δυνατή και με οπτικό προγραμματισμό. Διαθέτει ιδιαίτερες δυνατότητες οπτικοποίησης δεδομένων. Απευ-

θύνεται σε αρχάριους αλλά και έμπειρους αναλυτές.

- Το Rattle GUI είναι ένα γραφικό περιβάλλον για τη χρήση της γλώσσας R. Η R είναι μια ισχυρή και ταχέως ανερχόμενη γλώσσα προγραμματισμού για Εξόρυξη Δεδομένων, στατιστική ανάλυση και δημιουργία γραφικών. Η απευθείας χρήση της απαιτεί γνώσεις προγραμματισμού.
- Το Rapid Miner Studio είναι εμπορικό λογισμικό Εξόρυξης Δεδομένων. Διατίθεται μια ελεύθερη έκδοση με περιορισμένες (σε σχέση με την εμπορική έκδοση) δυνατότητες. Διαθέτει μεγάλη ποικιλία μεθόδων για κατηγοριοποίηση, παλινδρόμηση, ανάλυση κανόνων συσχέτισης, ανάλυση συστάδων κλπ.
- Το TANAGRA είναι ένα ελεύθερο ακαδημαϊκό λογισμικό για εκπαίδευση και έρευνα. Περιλαμβάνει πολλές μεθόδους για επιβλεπόμενη μάθηση, ανάλυση συστάδων, κανόνες συσχέτισης, στατιστική ανάλυση κλπ. Στα πλεονεκτήματα του περιλαμβάνεται το φιλικό και εύχρηστο γραφικό περιβάλλον εργασίας.
- Το Alteryx Project Edition παρέχει τη δυνατότητα στους αναλυτές να το χρησιμοποιήσουν για την εκτέλεση ενός έργου ελεύθερα. Το λογισμικό Alteryx Designer περιλαμβάνει εργαλεία ολοκλήρωσης δεδομένων, καθώς και προγνωστικής και χωρικής ανάλυσης.
- Το λογισμικό CMSR Data Miner είναι ελεύθερο για ακαδημαϊκή χρήση και περιλαμβάνει πληθώρα μεθόδων Εξόρυξης Δεδομένων, όπως Νευρωνικά Δίκτυα, Δένδρα Αποφάσεων, μεθόδους ανάλυσης συστάδων, εργαλεία οπτικοποίησης κλπ.
- Το KEEL σχεδιάστηκε για ερευνητικούς και εκπαιδευτικούς σκοπούς και δίνει έμφαση σε εξελικτικούς αλγορίθμους. Διαθέτει ένα απλό γραφικό περιβάλλον εργασίας. Περιλαμβάνει μεθόδους προεπεξεργασίας δεδομένων, στατιστικής ανάλυσης, μηχανικής μάθησης κλπ. και διευκολύνει τη διεξαγωγή πειραμάτων.

13.11 Πηγές για ελεύθερα σύνολα δεδομένων

Είναι προφανές ότι για τη διεξαγωγή αναλύσεων χρειάζονται καταρχήν δεδομένα. Δυστυχώς όμως, η διαθεσιμότητα δεδομένων δεν είναι πάντα εξασφαλισμένη. Οι επιχειρήσεις μπορούν εύκολα να αντλήσουν δεδομένα από τα πληροφοριακά τους συστήματα ή να προμηθευτούν δεδομένα από εξωτερικές πηγές έναντι αμοιβής. Για τον ακαδημαϊκό ερευνητή, η πρόσβαση σε δεδομένα είναι δυσκολότερη. Ειδικά τα οικονομικά δεδομένα είναι πολύ πιθανόν να είναι διαβαθμισμένα ως εμπιστευτικά και να μην είναι δημοσίως διαθέσιμα.

Μια πολύ συνηθισμένη πηγή δεδομένων για ακαδημαϊκή έρευνα είναι εξειδικευμένες βάσεις δεδομένων, που περιέχουν οικονομικά και άλλα στοιχεία επιχειρήσεων. Οι βάσεις αυτές είναι εμπορικό προϊόν, χρησιμοποιούνται από τράπεζες, χρηματιστηριακές εταιρείες και άλλους οργανισμούς και διατίθενται έναντι καθόλου ευκαταφρόνητης αμοιβής. Για τα εκπαιδευτικά ιδρύματα και για ακαδημαϊκούς σκοπούς προσφέρονται σε πολύ πιο προσιτές τιμές. Είναι πολύ διαδεδομένη τακτική διεθνώς να παρέχουν τα ΑΕΙ, μέσω των βιβλιοθηκών τους, πρόσβαση σε τέτοιες βάσεις δεδομένων. Οι ερευνητές μπορούν να μελετήσουν τα δεδομένα και να εξάγουν από τις βάσεις τα δεδομένα που τους ενδιαφέρουν, συνήθως σε μορφή αρχείων Excel. Στη συνέχεια μπορούν να τα αναλύσουν με όποιο λογισμικό επιθυμούν. Η πιο διαδεδομένη βάση δεδομένων είναι η Compustat, η οποία προσφέρεται από τον γνωστό οίκο εκτίμησης πιστοληπτικής ικανότητας Standard & Poor's. Μια άλλη διαδεδομένη βάση δεδομένων είναι η Amadeus, η οποία διατίθεται από το Bureau Van Dijk.

Ένας άλλος τρόπος δημιουργίας συνόλων δεδομένων είναι η συλλογή τους από την ερευνητική ομάδα. Η συνηθέστερη μέθοδος είναι η αποστολή ερωτηματολογίων. Επίσης, οι ερευνητές μπορούν να αντλήσουν δεδομένα από την ιστοθέση της Επιτροπής Κεφαλαιαγοράς, από τις δημοσιευμένες οικονομικές καταστάσεις των επιχειρήσεων και άλλες παρόμοιες πηγές. Η διαδικασία συλλογής δεδομένων είναι ιδιαίτερα χρονοβόρα και σχεδόν αναπόφευκτα οδηγεί στη δημιουργία σχετικά μικρών συνόλων δεδομένων.

Ο ευκολότερος τρόπος προμήθειας δεδομένων για επεξεργασία είναι μέσω του Διαδικτύου. Διάφοροι φορείς προσφέρουν ελεύθερα σύνολα δεδομένων για εξυπηρέτηση επιστημονικών και ερευνητικών σκοπών. Δυστυχώς, τα σύνολα οικονομικών δεδομένων δεν είναι τόσο συχνά όσο τα μηχανολογικά, μετεωρολογικά και άλλα δεδομένα. Ορισμένες από τις πλέον χρησιμοποιούμενες πηγές για προμήθεια δεδομένων είναι οι ακόλουθες:

- Ίσως η πιο γνωστή συλλογή συνόλων δεδομένων είναι η UCI Machine Learning Repository, που διατίθεται από το University College Irvine. Η συλλογή διαθέτει περισσότερα από 300 σύνολα δεδομένων, ορισμένα από τα οποία (περίπου 16) υπάγονται στην κατηγορία "Business". Τα σύνολα δεδομένων προορίζονται για εκπαίδευση και έρευνα στη Μηχανική Μάθηση. Υπάρχουν σύνολα

κατάλληλα για κατηγοριοποίηση, παλινδρόμηση και ανάλυση συστάδων. Η ιστοθέση παρέχει πληροφορίες σχετικά με τα δεδομένα, όπως το πλήθος στηλών και γραμμών, το έτος δημιουργίας, τον τύπο των δεδομένων, εργασίες για τις οποίες είναι κατάλληλα κλπ.

- Το WEKA, με τη διαδικασία εγκατάστασης, αποθηκεύει στον υποφάκελο ″data″ ορισμένα σύνολα δεδομένων. Επίσης, στην ιστοθέση του WEKA υπάρχει ειδική ιστοσελίδα με συλλογές συνόλων δεδομένων. Για τους χρήστες του WEKA, οι συλλογές αυτές έχουν το πλεονέκτημα ότι είναι σε μορφότυπο ARFF και δεν χρειάζονται μετατροπή, όπως συμβαίνει με άλλες συλλογές δεδομένων, οι οποίες διατίθενται από άλλες πηγές.
- Η ιστοθέση KDnuggets αναφέρεται στην εξόρυξη δεδομένων και προσφέρει μεγάλο αριθμό πολύ χρήσιμων πληροφοριών για θέματα, όπως λογισμικά, θέσεις εργασίας κλπ. Σε ειδική ιστοσελίδα της παρέχει κατάλογο πηγών για ελεύθερα σύνολα δεδομένων.
- Η κυβέρνηση των ΗΠΑ διαθέτει μέσω ειδικής ιστοθέσης της περισσότερα από 100.000 ελεύθερα σύνολα δεδομένων. Τα δεδομένα αυτά προέρχονται από διάφορες κρατικές υπηρεσίες και δεν είναι πάντα απευθείας αξιοποιήσιμα για πειράματα εξόρυξης δεδομένων. Σε αρκετές περιπτώσεις ο χρήστης θα πρέπει να υποβάλει τα δεδομένα σε προεπεξεργασία. Ωστόσο, η συλλογή είναι πολύ μεγάλη και ο υπομονετικός ερευνητής πιθανόν να ανακαλύψει σύνολα δεδομένων κατάλληλα για τα πειράματα τα οποία επιθυμεί να εκτελέσει.

Βιβλιογραφία/Αναφορές

Bouman, R., & Van Dongen, J. (2009). *Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL*. Indianapolis, IN: Wiley Publishing Inc.

Cs.utexas.edu. (2015). *WEKA Tutorial*. Retrieved 27 September, 2015, from http://www.cs.utexas.edu/users/ml/tutorials/Weka-tut/.

Cs.waikato.ac.nz. (2015). *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. Retrieved 27 September, 2015, from http://www.cs.waikato.ac.nz/ml/weka/documentation.html.

Hall, M. A. (1999). *Correlation-Based Feature Selection for Machine Learning* (Ph.D.). University of Waikato.

Ibm.com. (2015). *IBM developerWorks : Open source : Technical library*. Retrieved 27 September, 2015, from http://www.ibm.com/developerworks/views/opensource/libraryview.jsp?search_by=data+mining+weka.

Kaluza, B. (2013). *Instant WEKA How-to*. Birmingham, UK: Packt Publishing Ltd.

Larose, D. T. (2006). *Data Mining Methods and Models*. Hoboken, NJ: John Wiley & Sons Inc.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer + Business Media.

Meta-guide.com. (2015). *100 Best Weka Tutorial Videos | Meta-Guide.com*. Retrieved 27 September, 2015, from http://meta-guide.com/videography/100-best-weka-tutorial-videos/.

Puthran, S., & Shah, K. (2012). *Intrusion Detection System using Datamining Techniques*. Saarbrücken: LAP LAMBERT Academic Publishing.

Rochester, E. (2013). *Clojure Data Analysis Cookbook*. Birmingham, UK: Packt Publishing Ltd.

Slideshare.net. (2012). *Data mining techniques using weka*. Retrieved 27 September, 2015, from http://www.slideshare.net/rathorenitin87/data-mining-techniques-using-weka.

Technologyforge.net. (2015). *Technology Forge - WEKA Tutorials*. Retrieved 27 September, 2015, from http://www.technologyforge.net/WekaTutorials/.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann Publishers.

Πίνακας Όρων

Ελληνικός Όρος	Αγγλικός Όρος
Ακάθαρτα Δεδομένα	Dirty Data
Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων	Knowledge Discovery in Data Bases
Ανάλυση Διακύμανσης	Analysis of Variance
Ανάλυση Καταναλωτικού Καλαθιού	Market Basket Analysis
Ανάλυση Κυρίων Συνιστωσών	Principal Components Analysis
Ανάλυση Παλινδρόμησης	Regression Analysis
Ανάλυση Συναισθήματος	Sentiment Analysis
Ανάλυση Συστάδων	Cluster Analysis ή Clustering
Αναλυτική Κάθοδος	Drill Down
Αναλυτική των Επιχειρήσεων	Business Analytics
Αντίστροφη Μετάδοση Σφάλματος	Backpropagation
Απλή Σύνδεση	Simple Linkage
Αποθήκη Δεδομένων	Data Warehouse
Απώλεια Πελατών	Customer Churn
Αρχιτεκτονική Δικτύου	Network Architecture
Αυτοοργανούμενοι Χάρτες	Self Organizing Maps
Βηματική Οπίσθια Εξάλειψη	Stepwise Backward Elimination
Βηματική Πρόσθια Επιλογή	Stepwise Forward Selection
Γενετικοί Αλγόριθμοι	Genetic Algorithms
Γνώρισμα	Attribute
Δένδρα Αποφάσεων	Decision Trees
Διαγράμματα Επιρροής	Influence Diagramms
Διαιρετικές Μέθοδοι	Divisive Methods
Διακριτοποίηση Δεδομένων	Data Discretization
Διασταυρούμενες Πωλήσεις	Cross Selling
Διασταυρούμενη Επικύρωση 10 Τμημάτων	10 fold cross validation
Διατακτικά Γνωρίσματα	Ordinal Attributes
Διαχείριση Αλλαγής	Change Management
Διαχωριστικές Μέθοδοι	Partitioning Methods
Διεπαφή	Interface
Διερευνητική Ανάλυση Δεδομένων	Exploratory Data Analysis
Διοίκηση Επιχειρησιακής Απόδοσης	Corporate Performance Management
Διοίκηση Μέσω Εξαιρέσεων	Management by Exception
Δωμάτια Αποφάσεων	Decision Rooms
Εικονογραφικές Τεχνικές	Iconic Display Techniques
Εισερχόμενο Μάρκετινγκ	Inbound Marketing
Εκπαίδευση μέσω Παραδειγμάτων	Learning by Examples
Εκπαίδευση μέσω Παρατήρησης	Learning by Observation
Ελεγκτικό Πρότυπα	Statement of Auditing Standards
Εξαγωγή, Μετασχηματισμός, Φόρτωση	Extract, Transform, Load
Εξερχόμενο Μάρκετινγκ	Outbound Marketing
Εξόρυξη Δεδομένων	Data Mining
Εξόρυξη Επιχειρηματικών Διαδικασιών	Business Process Mining
Εμπιστοσύνη	Confidence
Επεκτασιμότητα	Scalability

Επεξεργασία Φυσικής Γλώσσας	Natural Language Processing
Επιβλεπόμενη Μάθηση	Supervised Learning
Επικύρωση Μοντέλου	Model Validation
Επιλογή Χαρακτηριστικών	Feature Selection
Επιχειρηματικές Διαδικασίες	Business Processes
Επιχειρηματική Ευφυΐα Ως Υπηρεσία	Business Intelligence As A Service
Ευαίσθητη ως προς το Κόστος Εκπαίδευση	Cost Sensitive Learning
Ευρετικές Μέθοδοι	Heuristics
Ευφυή Συστήματα Υποστήριξης Αποφάσεων	Intelligent Decision Support Systems
Ιδιοδιάνυσμα	Eigenvector
Ιδιοτιμή	Eigenvalue
Ιεραρχικές Μέθοδοι	Hierarchical Methods
Καθαρισμός Δεδομένων	Data Cleansing
Κάθετος Τεμαχισμός	Dice
Καθοδηγούμενο από τα Γεγονότα Μάρκετινγκ	Event Driven Marketing ή Triggered Marketing
Κανόνας Συσχέτισης	Association Rule
Κανονικοποίηση Δεδομένων	Data Normalization
Κανονιστική Συμμόρφωση	Regulatory Compliance
Κατηγοριοποίηση	Classification
Κατάρα των Διαστάσεων	Curse of Dimesionality
Κατηγορικά Δεδομένα	Categorical Data
Κέρδος Πληροφορίας	Information Gain
Κόκκωση	Granularity
Κυβοειδές	Cuboid
Κύκλος Ζωής Ανάπτυξης Συστήματος	System Development Life Cycle
Κύριοι Δείκτες Επίδοσης	Key Performance Indicators
Κυρτό Περίβλημα	Convex Hull
Λογικές Αποφάσεις	Rational Decisions
Λογισμικό Ομάδων	Groupware
Λογιστική (ή Λογαριθμική) Παλινδρόμηση	Logistic Regression
Λόγος Κέρδους	Gain Ratio
Μέθοδοι Βασισμένες σε Μοντέλα	Model Based Methods
Μέθοδοι Βασισμένες στην Πυκνότητα	Density Based Methods
Μέθοδοι Πλέγματος	Grid Based Methods
Μείωση Διαστάσεων	Dimensionality Reduction
Μερική Υλοποίηση	Partial Metarialization
Μετασχηματισμός Δεδομένων	Data Transformation
Μη Επιβλεπόμενη Μάθηση	Unsupervised Learning
Νευρωνικά Δίκτυα	Neural Networks
Οκνηροί Κατηγοριοποιητές	Lazy Classifiers
Ονομαστικά Γνωρίσματα	Nominal Attributes
Οριζόντιος Τεμαχισμός	Slice
Οριοθετημένη Λογική	Bounded Rationality
Παλινδρόμηση	Regression
Παλινδρόμηση Διανυσμάτων Υποστήριξης	Support Vector Regression
Παρατήρηση με ακραίες τιμές	Outlier
Περιγραφική Ανάλυση	Descriptive Analytics
Περιστροφή	Pivot

Πίνακας Διαστάσεων	Dimension Table
Πίνακας Σύγχυσης	Confusion Matrix
Πίνακας Συμβάντων	Fact Table
Πίνακας Συνδιασποράς	Covariance Matrix
Πλεονασμός Δεδομένων	Data Redundancy
Πλήρης Σύνδεση	Complete Linkage
Πλήρης Υλοποίηση	Full Materialization
Πολυσυγγραμικότητα	Multicollinearity
Πρατήριο Δεδομένων	Data Mart
Προγνωστική Ανάλυση	Predictive Analytics
Πύλη	Portal
Στοιχειοσύνολο	Itemset
Στοχευμένη Διαφήμιση	Target Marketing
Συναθροιστική Άνοδος	Roll Up
Συνδιασπορά	Covariance
Σύνολο Εκπαίδευσης	Training Set
Σύνολο Επικύρωσης	Validation Set
Συνάρτηση Ακτινωτής Βάσης	Radial Base Function
Σύνδεση Μέσου Όρου	Average Link
Συντελεστής Προσδιορισμού	Coefficient of Determination
Συσσωρευτικές Μέθοδοι	Agglomerative Methods
Συστήματα Αναλυτικής Επεξεργασίας Άμεσης Επικοινωνίας	On Line Analytical Processing
Συστήματα Διαχείρισης Εφοδιαστικής Αλυσίδας	Supply Chain Management Systems
Συστήματα Διαχείρισης Σχέσεων Πελατών	Customer Relationship Management Systems
Συστήματα Επεξεργασίας Συναλλαγών με Άμεση Επικοινωνία	On Line Transaction Processing
Συστήματα Σχεδιασμού Επιχειρησιακών Πόρων	Enterprise Resources Planning Systems
Συστήματα Υποστήριξης Διοίκησης	Executive Support Systems
Συστήματα Υποστήριξης Ομαδικών Αποφάσεων	Group Decision Support Systems
Συστήματα Υποστήριξης Ομάδων	Group Support Systems
Συχνότητα Εμφάνισης	Frequency, ή Support Count, ή Count
Σφάλμα Κβάντωσης	Quantization Error
Σχήμα Αστέρα	Star Schema
Σχήμα Αστερισμού	Constellation Schema
Σχήμα Χιονονιφάδας	Snowflake Schema
Κατηγοριοποιητές Βασισμένοι σε Παραδείγματα	Instance Based Classifiers
Τεχνικές Γεωμετρικού Μετασχηματισμού	Geometrically Transformed Techniques
Τεχνικές Εικονοστοιχείων	Iconic Display Techniques
Τεχνικές Στοίβας	Stacked Display Techniques
Τεχνικές Δυναμικής Προβολής	Dynamic Projection Techniques
Τεχνικές Διαδραστικής Επιλογής	Interactive Filtering Techniques
Τεχνικές Διαδραστικής Διαβάθμισης Λεπτομέρειας	Interactive Zooming Techniques
Τεχνικές Διαδραστικής Στρέβλωσης	Interactive Distortion Techniques
Τεχνικές Διαδραστικής Σύνδεσης και Χρωματισμού	Interactive Linking and Brushing Techniques
Τεχνητό Ανοσοποιητικό Σύστημα	Artificial Immune System
Τμηματοποίηση Αγοράς	Market Segmentation
Τοπολογία Δικτύου	Network Topology
Τραχέα Σύνολα	Rough Sets

Υλοποιημένες Όψεις	Materialized Views
Υπερεπίπεδο	Hyperplane
Υπερπροσαρμογή στα Δεδομένα	Data Overfitting
Υποκατάστατα Κλειδιά	Surrogate Keys
Υπολογιστική Νέφους	Cloud Computing
Υποστήριξη	Support
Χαμένες Τιμές	Missing Values
Χαρακτηριστικά	Features
Ψευδομεταβλητές	Dummy Variables