

WEKA ΠΑΡΑΔΕΙΓΜΑΤΑ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμηση Στατιστικής και Αναλογιστικών –
Χρηματοοικονομικών Μαθηματικών
Πανεπιστημιο Αιγαίου
Σαμος

2021

ΕΙΣΑΓΩΓΗ

Η weka είναι ένα software για εξόρυξη δεδομένων γραμμένο σε JAVA το οποίο περιέχει υλοποιημένες μεθόδους για:

- Προεπεξεργασία Δεδομένων
- Ταξινόμηση
- Συσταδοποίηση
- Εύρεση Κανόνων Συσχέτισης

ΕΙΣΑΓΩΓΗ

Το software είναι διαθέσιμο για εγκατάσταση από την ιστοσελίδα:

<http://www.cs.waikato.ac.nz/ml/weka/>

ΕΙΣΑΓΩΓΗ



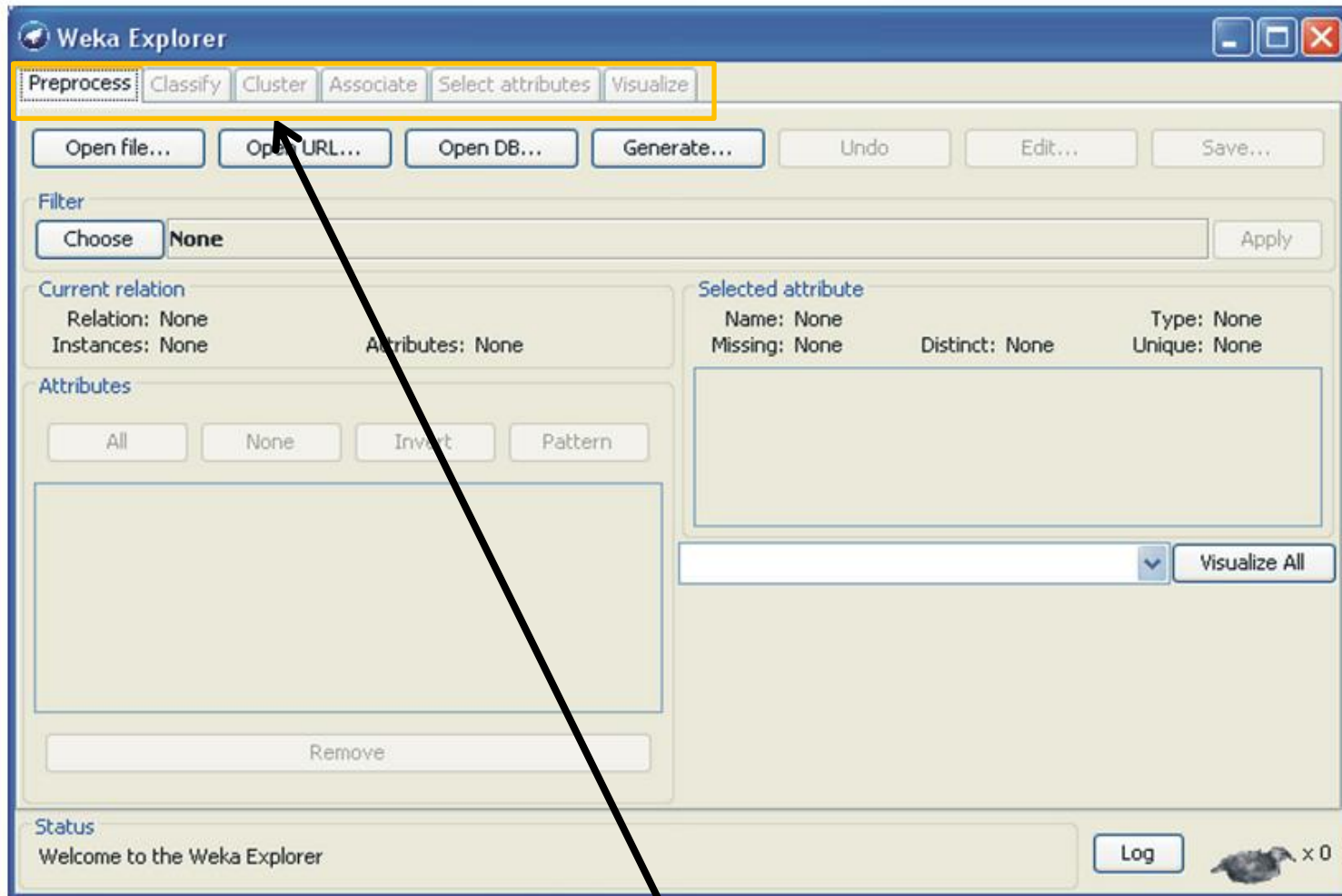
- Ο **Explorer** είναι η πιο δημοφιλής διεπαφή. Ο χρήσης μπορεί να εκτελέσει όλες τις κύριες εργασίες Εξόρυξης Δεδομένων, όπως κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, ανακάλυψη κανόνων συσχέτισης, προεπεξεργασία των δεδομένων και οπτικοποίηση.
- Ο **Experimenter** είναι ένα περιβάλλον για διεξαγωγή 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.

ΕΙΣΑΓΩΓΗ

Ανοίγοντας το πρόγραμμα, μέσω του μενού Application → Explorer → Open file δίνεται η δυνατότητα να επιλεγεί ένα σύνολο δεδομένων στο οποίο μπορούν να εφαρμοστούν τεχνικές που αφορούν :

- Preprocess
- Classify
- Cluster
- Associate
- Select Attributes
- Visualize

ΕΙΣΑΓΩΓΗ



ο παράθυρο της εφαρμογής περιλαμβάνει 6 tabs για προεπεξεργασία των δεδομένων, κατηγοριοποίηση, ανάλυση συστάδων, επιλογή γνωρισμάτων και οπτικοποίηση

ΕΙΣΑΓΩΓΗ

- Επιλέγοντας ένα σύνολο δεδομένων (αρχείο .arff), εμφανίζονται γραφικά τα δεδομένα για καθένα από τα γνωρίσματα ξεχωριστά καθώς και στατιστικές πληροφορίες για αυτά. Εάν στο σύνολο δεδομένων δίνεται και κάποια κλάση στην οποία ταξινομούνται, τα δεδομένα που ανήκουν στην ίδια κλάση εμφανίζονται με το ίδιο χρώμα

Weka 3.5.7 - Explorer

Program Applications Tools Visualization Windows Help

Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose None Apply

Current relation

Relation: iris Instances: 150 Attributes: 5

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> sepalength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petalength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

Selected attribute

Name: sepalength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

16 30 34 28 26 10 7

4.3 6.1 7.9

ΕΙΣΑΓΩΓΗ

- Τα αρχεία που περιέχουν το σύνολο δεδομένων πρέπει να έχουν συγκεκριμένο format και να αποθηκεύονται με την επέκταση .arff
- Στον φάκελο C:\Program Files\Weka-3-5\data περιέχονται κάποια παραδείγματα τέτοιων αρχείων.

ΕΙΣΑΓΩΓΗ

@relation heart-disease-simplified ← ΣΧΕΣΗ

@attribute age numeric ← ΠΕΔΙΑ

@attribute sex { female, male }

@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina }

@attribute cholesterol numeric

@attribute exercise_induced_angina { no, yes }

@attribute class { present, not_present }

@data ← ΔΕΔΟΜΕΝΑ

63,male,typ_angina,233,no,not_present

67,male,asympt,286,yes,present

67,male,asympt,229,yes,present

38,female,non_anginal,?,no,not_present

Στο Διαδίκτυο διατίθεται εφαρμογή μετατροπής αρχείων Excel σε αρχεία ARFF. Μπορείτε να προμηθευτείτε την εφαρμογή από [ιστοσελίδα 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:

Προ-επεξεργασία

Πεδία

Πληροφορίες Πεδίου

Ορισμός Κλάσης

Κατανομή τιμών

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None Apply

Current relation: Relation: Qualification Instances: 150 Attributes: 19

Attributes: All None Invert Pattern

No.	Name
1	Qualification
2	Turnover
3	PLBT
4	WORKING_CAP
5	SOLVENCYR
6	GEARING
7	ROSF
8	ROTA
9	QUISCORE
10	IFBIG
11	RETAINDPL
12	TOTASS
13	CURLIAB
14	LONGTERMLIAB

Remove

Status: OK Log x 0

Selected attribute: Name: SOLVENCYR Type: Numeric Missing: 0 (0%) Distinct: 148 Unique: 146 (97%)

Statistic	Value
Minimum	-85.04
Maximum	99.77
Mean	43.083
StdDev	32.976

Class: Qualification (Nom) Visualize All

Bin Range	Count
-85.04 - 7.38	2
7.38 - 43.083	1
43.083 - 78.78	3
78.78 - 114.48	2
114.48 - 150.18	20
150.18 - 185.88	31
185.88 - 221.58	42
221.58 - 257.28	32
257.28 - 292.98	17

Κατηγοριοποίηση

Ορισμένες από τις κυριότερες μεθόδους κατηγοριοποίησης που περιλαμβάνονται είναι τα Μπαΐεσιανά Δίκτυα, οι Μηχανές Διανυσμάτων Υποστήριξης, η Λογιστική Παλινδρόμηση, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron και τα Δένδρα Αποφάσεων

Κατηγοριοποίηση

Ορισμός μεθόδου κατηγοριοποίησης

Μέθοδος αξιολόγησης

Ορισμός κλάσης

Αποτελέσματα μοντέλου

The screenshot shows the Weka Explorer interface. The 'Classifier' tab is active, and 'MultilayerPerceptron' is selected. The 'Test options' section shows 'Cross-validation' with 'Folds' set to 10. The 'Classifier output' section displays the following results:

Metric	Value	Percentage
Correctly Classified Instances	106	70.6667 %
Incorrectly Classified Instances	44	29.3333 %
Kappa statistic	0.4134	
Mean absolute error	0.3173	
Root mean squared error	0.4205	
Relative absolute error	63.4511 %	
Root relative squared error	84.061 %	
Total Number of Instances	150	

Below the main output, there is a section for 'Detailed Accuracy By Class' and a 'Confusion Matrix'.

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Qualified	0.697	0.284	0.716	0.697	0.707	0.807
Unqualified	0.716	0.303	0.697	0.716	0.707	0.807
Weighted Avg.	0.707	0.293	0.707	0.707	0.707	0.807

The 'Confusion Matrix' section shows:

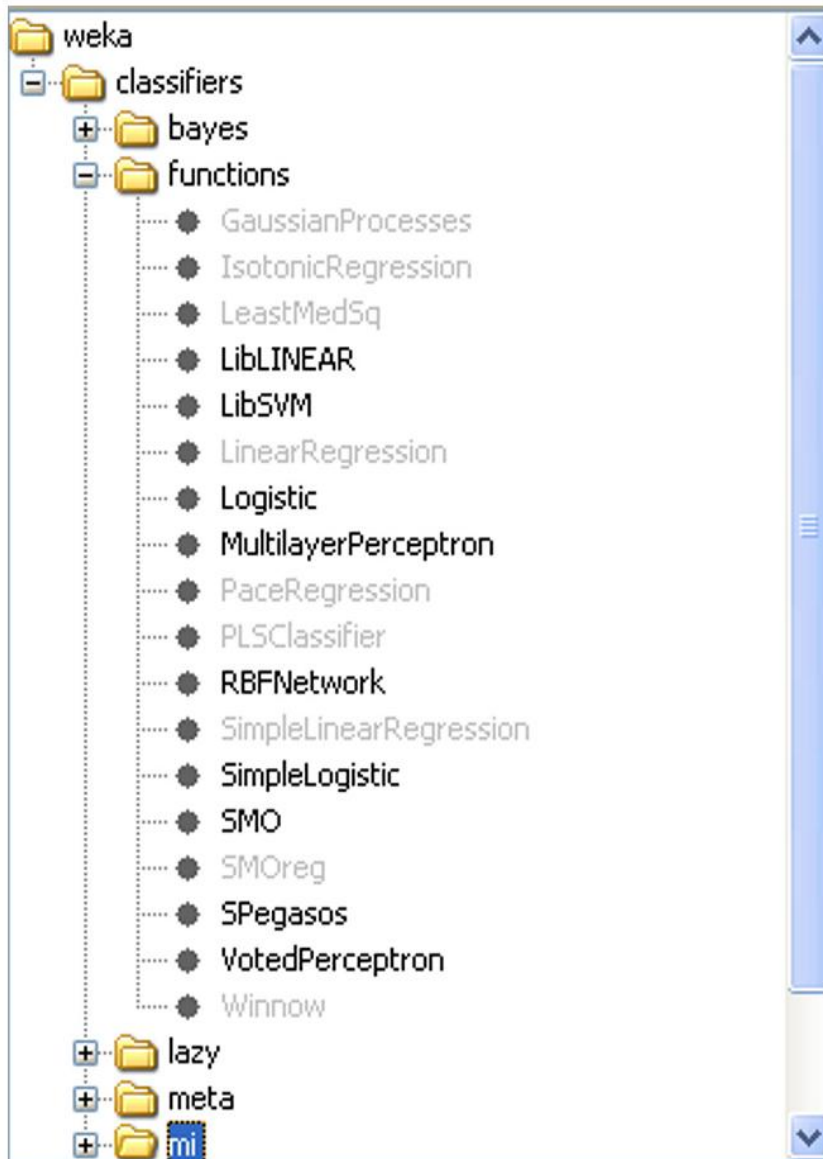
```
=== Confusion Matrix ===
 a b <-- classified as
53 23 | a = Qualified
21 53 | b = Unqualified
```

The 'Result list' at the bottom left shows a list of models, with '08:54:27 - functions.MultilayerPerceptron' selected.

Λίστα μοντέλων

Νευρωνικό Δίκτυο τύπου MultiLayer Perceptron

Κατηγοριοποίηση



Το WEKA περιλαμβάνει μεγάλο αριθμό μεθόδων κατηγοριοποίησης. Οι μέθοδοι είναι ομαδοποιημένες σε κατηγορίες, οι οποίες παρουσιάζονται σε μορφή δένδρου

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".

Ορισμός μεθόδου κατηγοριοποίησης

Μέθοδος αξιολόγησης

Ορισμός κλάσης

Αποτελέσματα μοντέλου

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose MultilayerPerceptron - L 0.3 - M 0.2 - N 500 - V 0 - S 0 - E 20 - H a - R

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split %

More options...

Classifier output

Correctly Classified Instances 106 70.6667 %

Incorrectly Classified Instances 44 29.3333 %

Kappa statistic 0.4134

Mean absolute error 0.3173

Root mean squared error 0.4205

Relative absolute error 63.4511 %

Root relative squared error 84.061 %

Total Number of Instances 150

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Qualified	0.697	0.284	0.716	0.697	0.707	0.807
Unqualified	0.716	0.303	0.697	0.716	0.707	0.807
Weighted Avg.	0.707	0.293	0.707	0.707	0.707	0.807

=== Confusion Matrix ===

a b <-- classified as

53 23 | a = Qualified

51 53 | b = Unqualified

Result list (right-click for options)

08:48:51 - trees_348

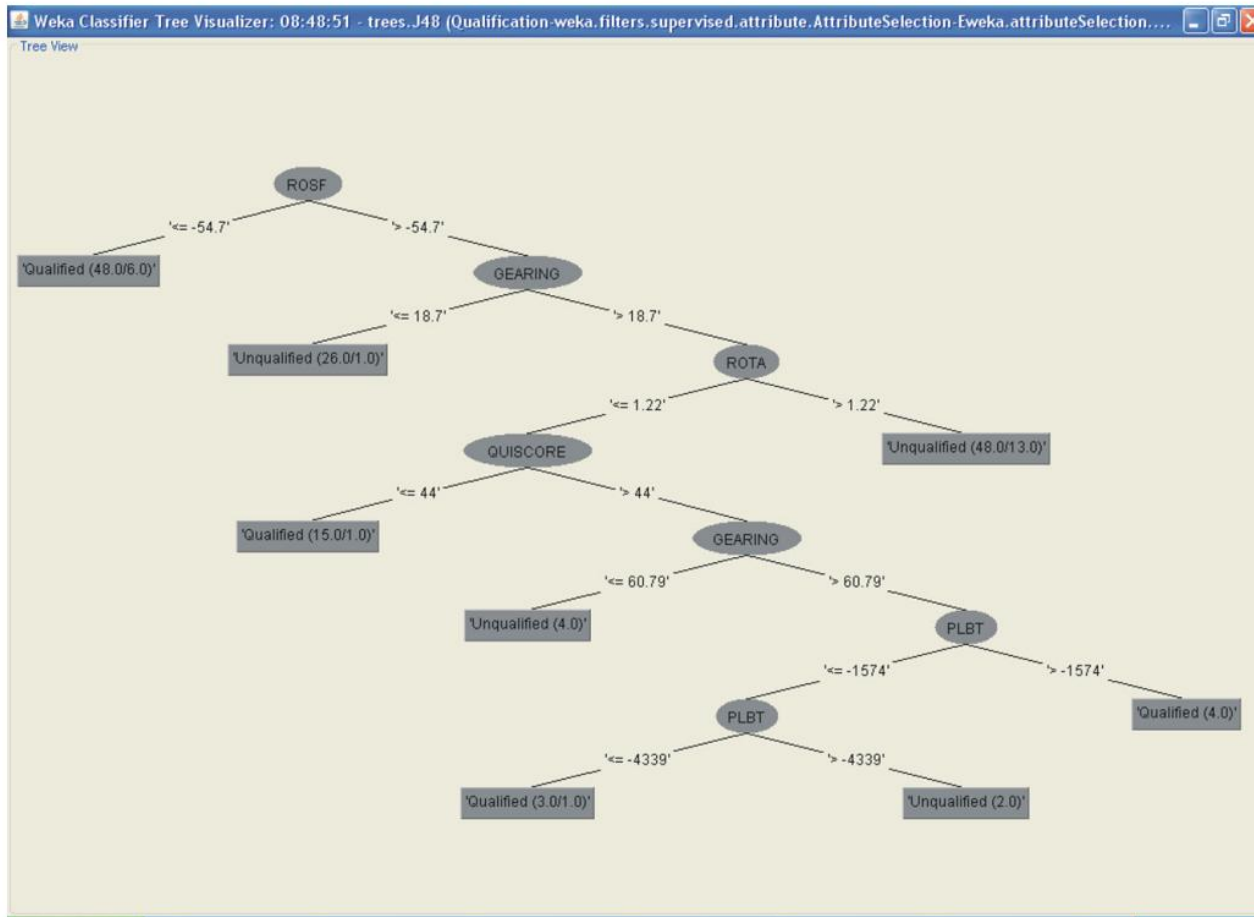
08:54:27 - functions: MultilayerPerceptron

Status OK

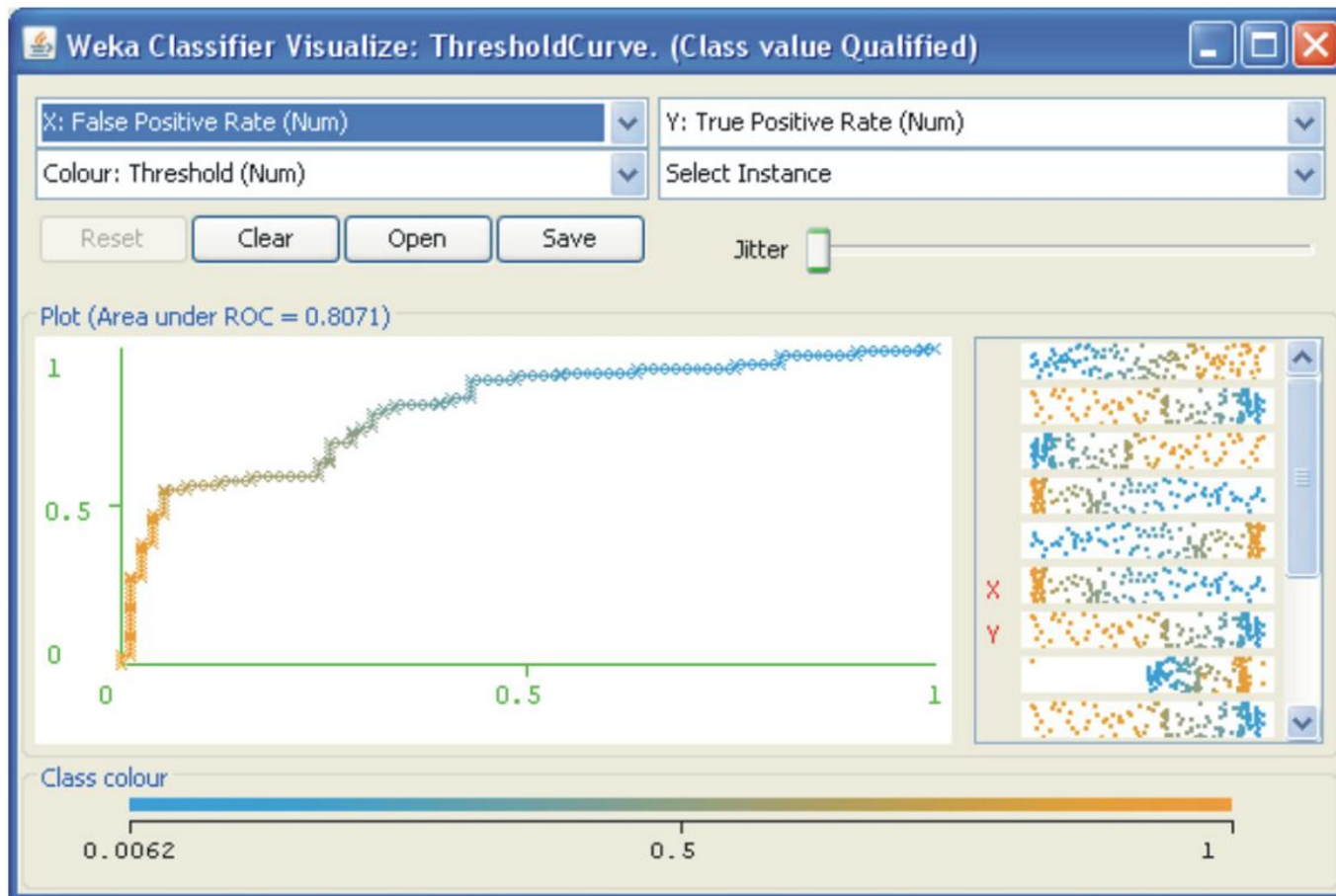
Log x 0

Λίστα μοντέλων

Δένδρο Αποφάσεων



Καμπύλες ROC



Ανάλυση σε συστάδες

- Ανάμεσα στους αλγορίθμους που διατίθενται περιλαμβάνονται ο k-Means, η Συσσωρευτική Ιεραρχική ΑΣ, η Expected Maximazation (EM)

150 επιχειρήσεις

Ανάλυση σε συστάδες

k-Means

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Clusterer: Choose SimpleKMeans -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode:

- Use training set
- Supplied test set
- Percentage split
- Classes to clusters evaluation
- Store clusters for visualization

Clusterer output:

```

kMeans
=====
Number of iterations: 11
Within cluster sum of squared errors: 5.876330280878608
Missing values globally replaced with mean/mode

Cluster centroids:
Attribute  Full Data      Cluster#
              (150)          (37)      (31)      (42)      (40)
-----
SOLVENCYR  43.0829    65.5511    3.4919    60.8545    34.3225
GEARING    247.4269   60.6235    778.8852   57.8024    207.4455
ROSF       -45.4315   34.6089   -187.2548  -28.7112   -27.112
QUISCORE   48          90.4595    6.0323    55.7857    33.075

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances
0      37 ( 25%)
1      31 ( 21%)
2      42 ( 28%)
3      40 ( 27%)
  
```

Result list (right-click for options):

10:43:58 - SimpleKMeans

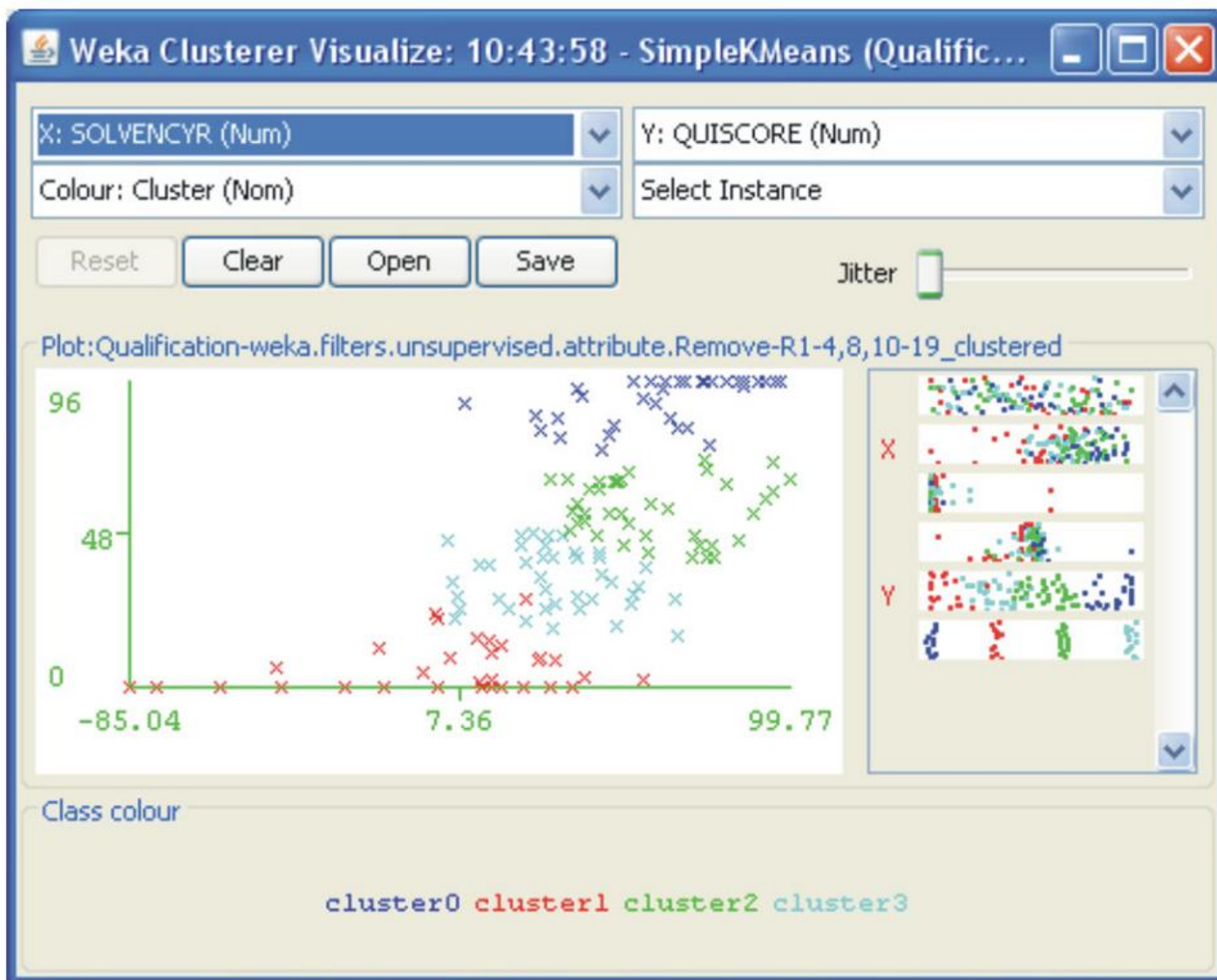
Status: OK

Η πρώτη συστάδα περιλαμβάνει 37 παρατηρήσεις, η δεύτερη 31, η τρίτη 42 και η τέταρτη 40.

τέσσερις συστάδες

αριθμοδείκτης αξιοπιστίας

Ανάλυση σε συστάδες



Οπτικοποίηση

