WEKA ΠΑΡΑΔΕΙΓΜΑΤΑ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
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

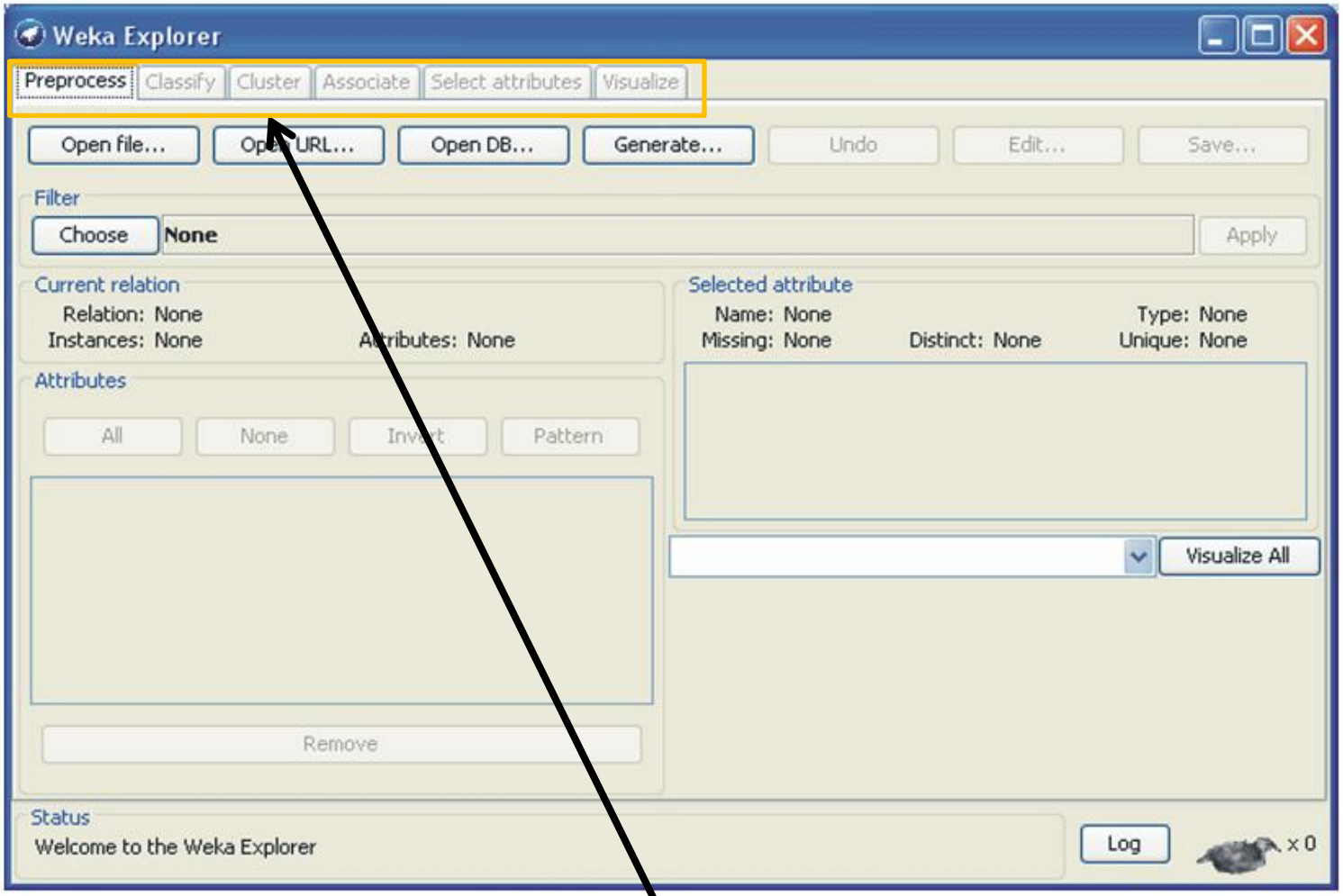
- Ο **Explorer** είναι η πιο δημοφιλής διεπαφή. Ο χρήσης μπορεί να εκτελέσει όλες τις κύριες εργασίες Εξόρυξης Δεδομένων, όπως κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, ανακάλυψη κανόνων συσχέτισης, προεπεξεργασία των δεδομένων και οπτικοποίηση.
- Ο **Experimenter** είναι ένα περιβάλλον για διεξαγωγή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

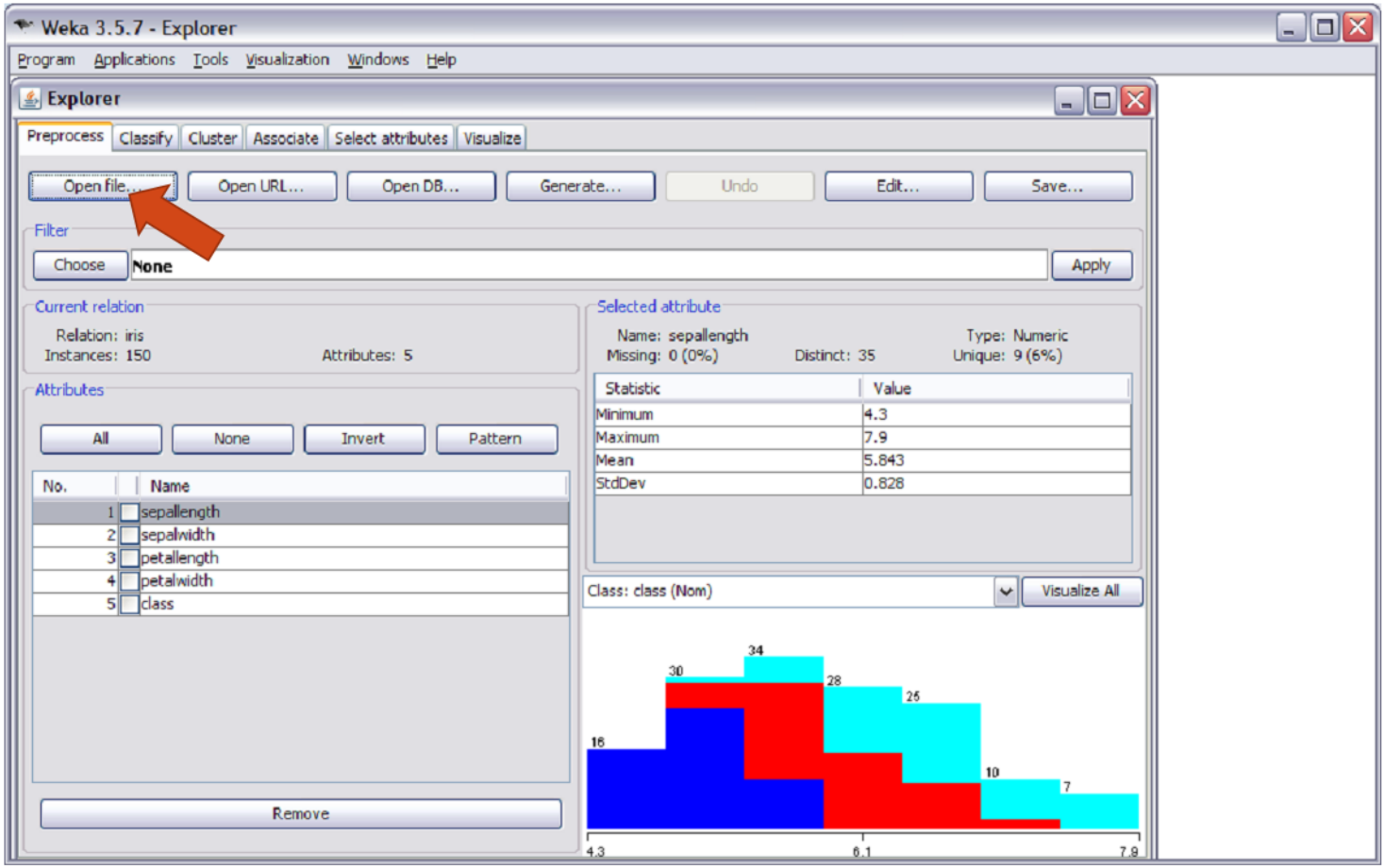
ΕΙΣΑΓΩΓΗ



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



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

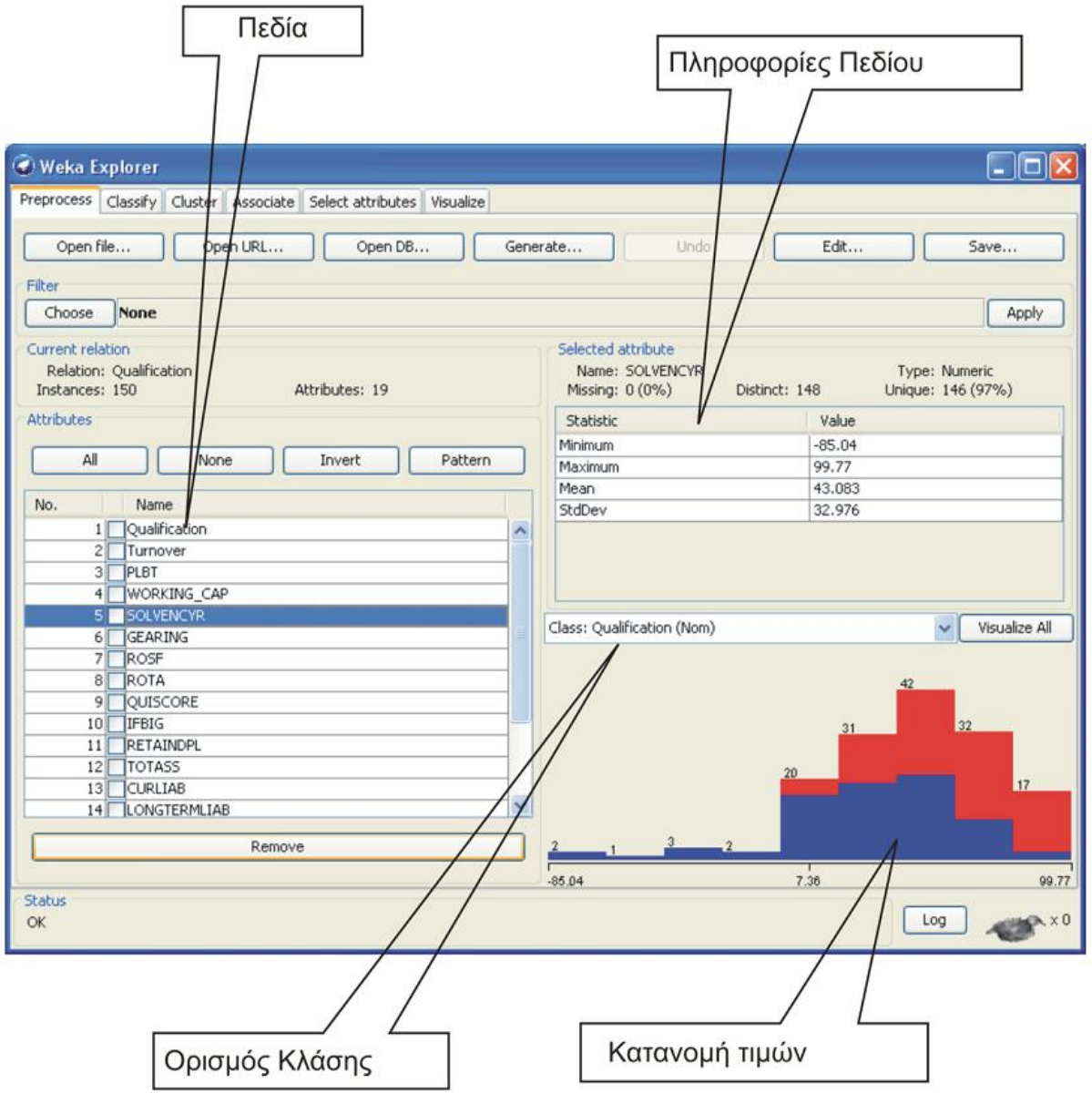
@attribute age numeric ← ΠΕΔΙΑ
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@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 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:

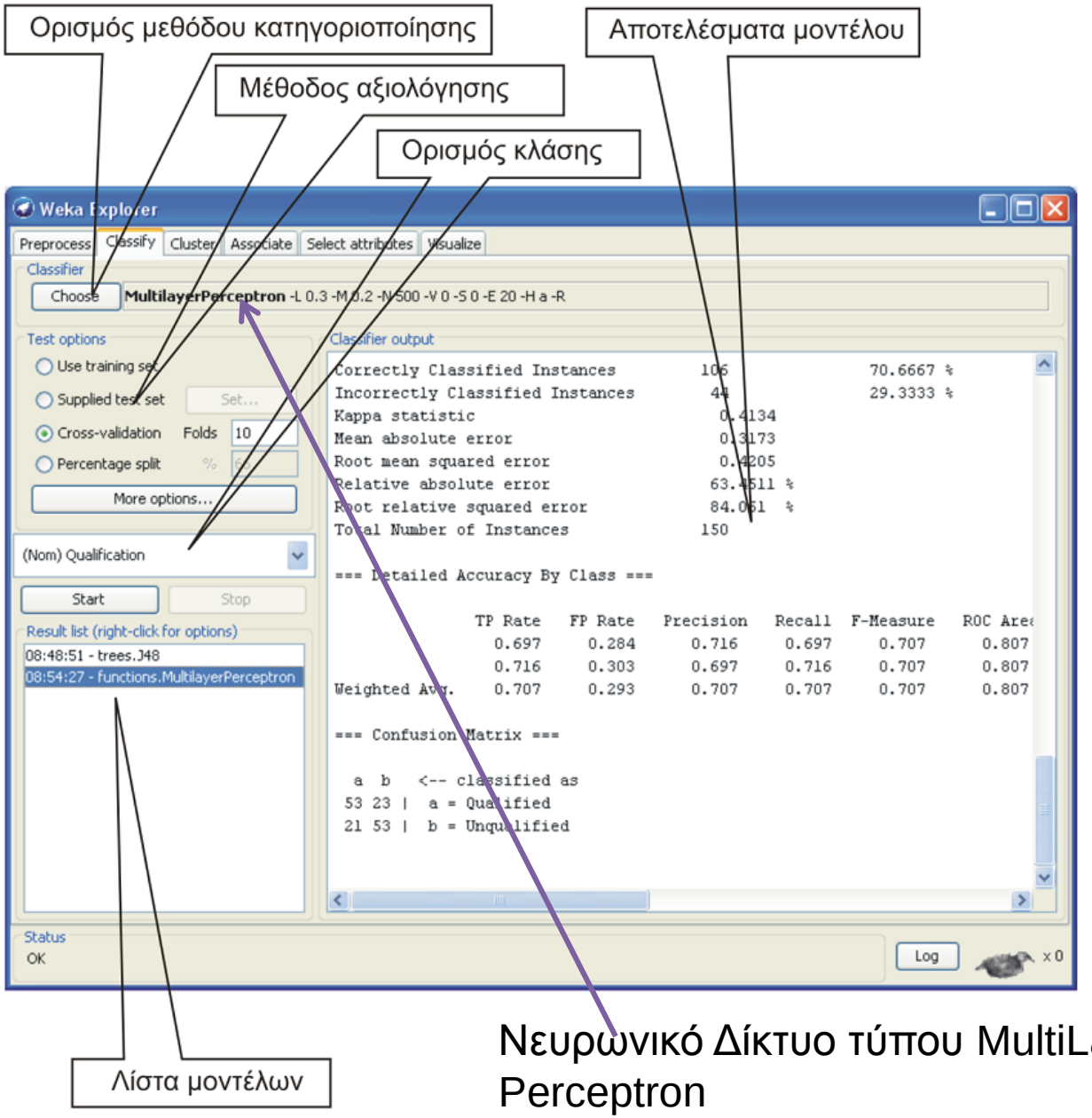
Προ-επεξεργασια



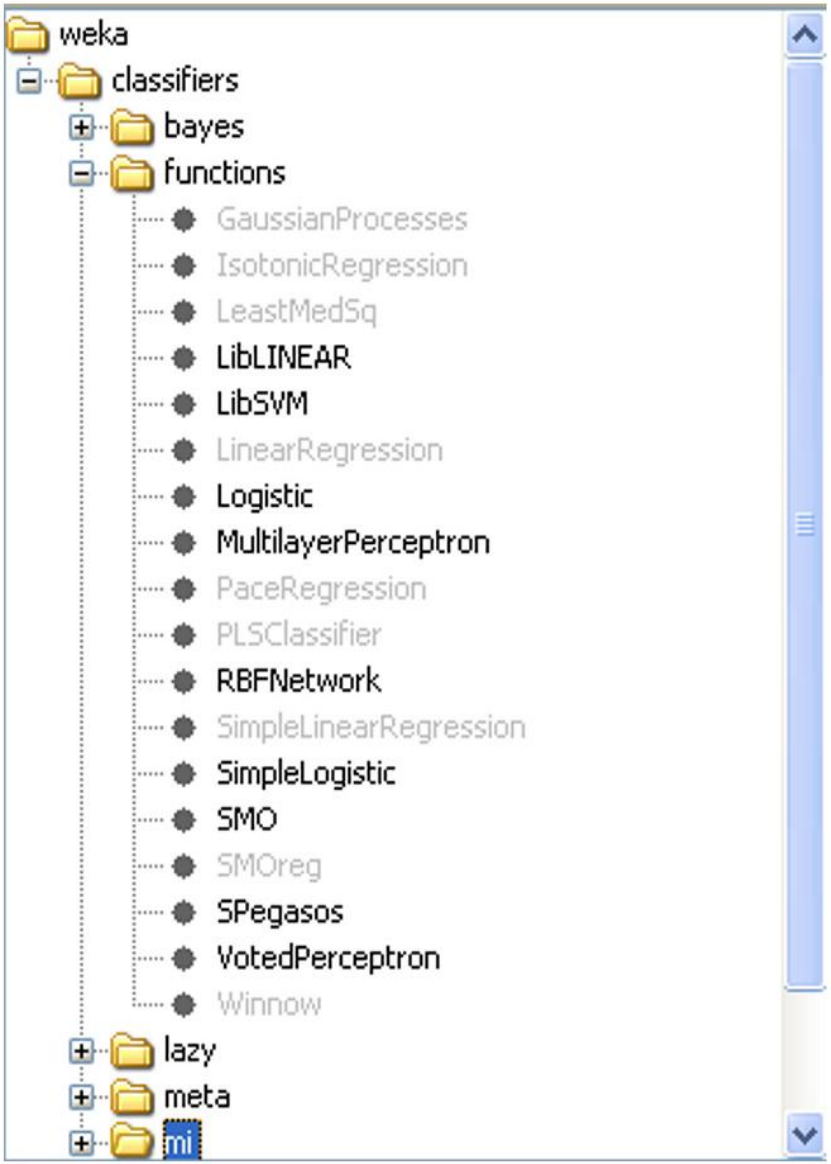
Κατηγοριοποίηση

Ορισμένες από τις κυριότερες μεθόδους κατηγοριοποίησης που περιλαμβάνονται είναι τα Μπαϋεσιανά Δίκτυα, οι Μηχανές Διανυσμάτων Υποστήριξης, η Λογιστική Παλινδρόμηση, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptronκαι ταΔένδρα Αποφάσεων

Κατηγοριοποίηση



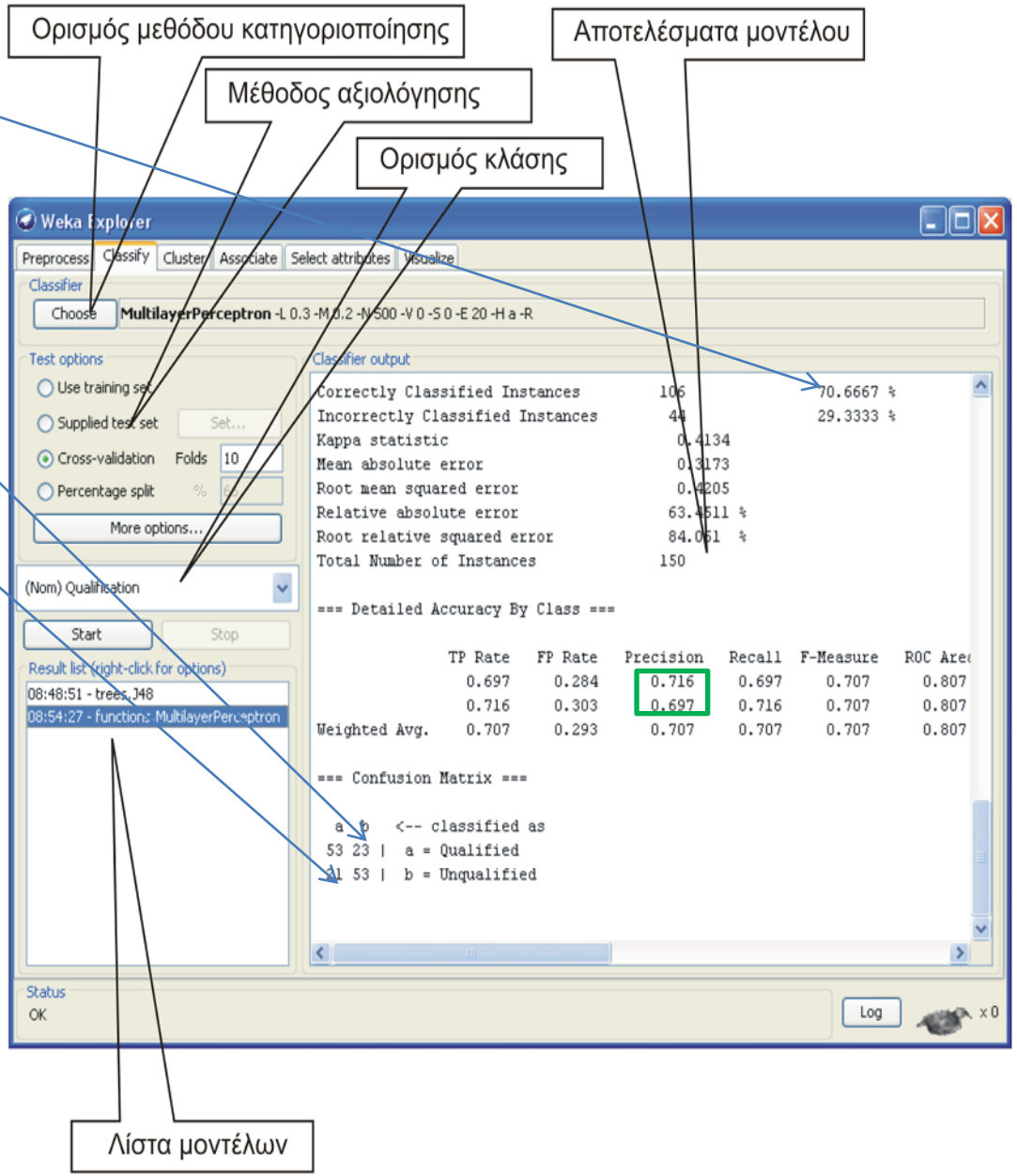
Κατηγοριοποίηση



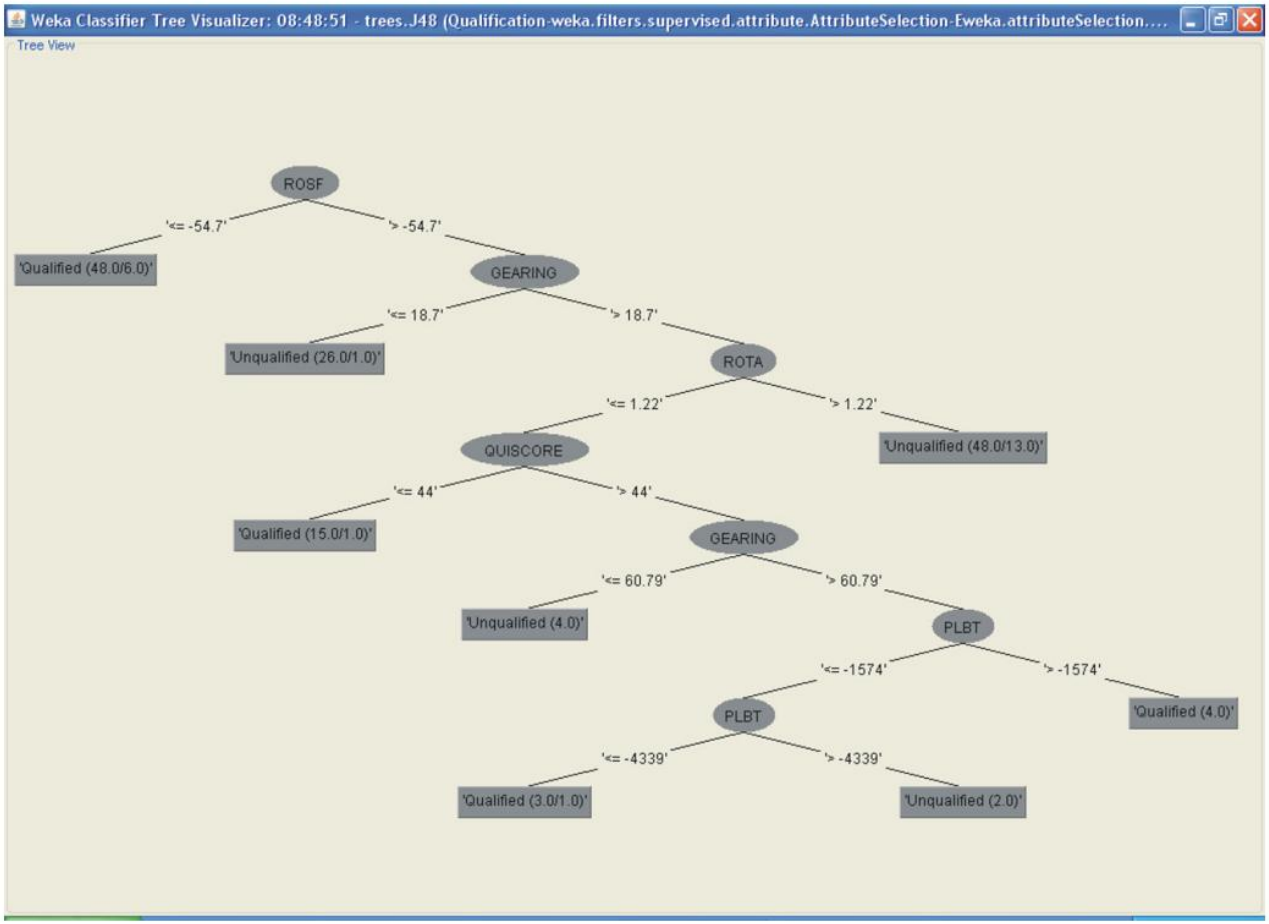
Το WEKA περιλαμβάνει μεγάλο αριθμό μεθόδων κατηγοριοποίησης. Οι μέθοδοι είναι ομαδοποιημένες σε κατηγορίες, οι οποίες παρουσιάζονται σε μορφή δένδρου

Νευρωνικό δίκτυο

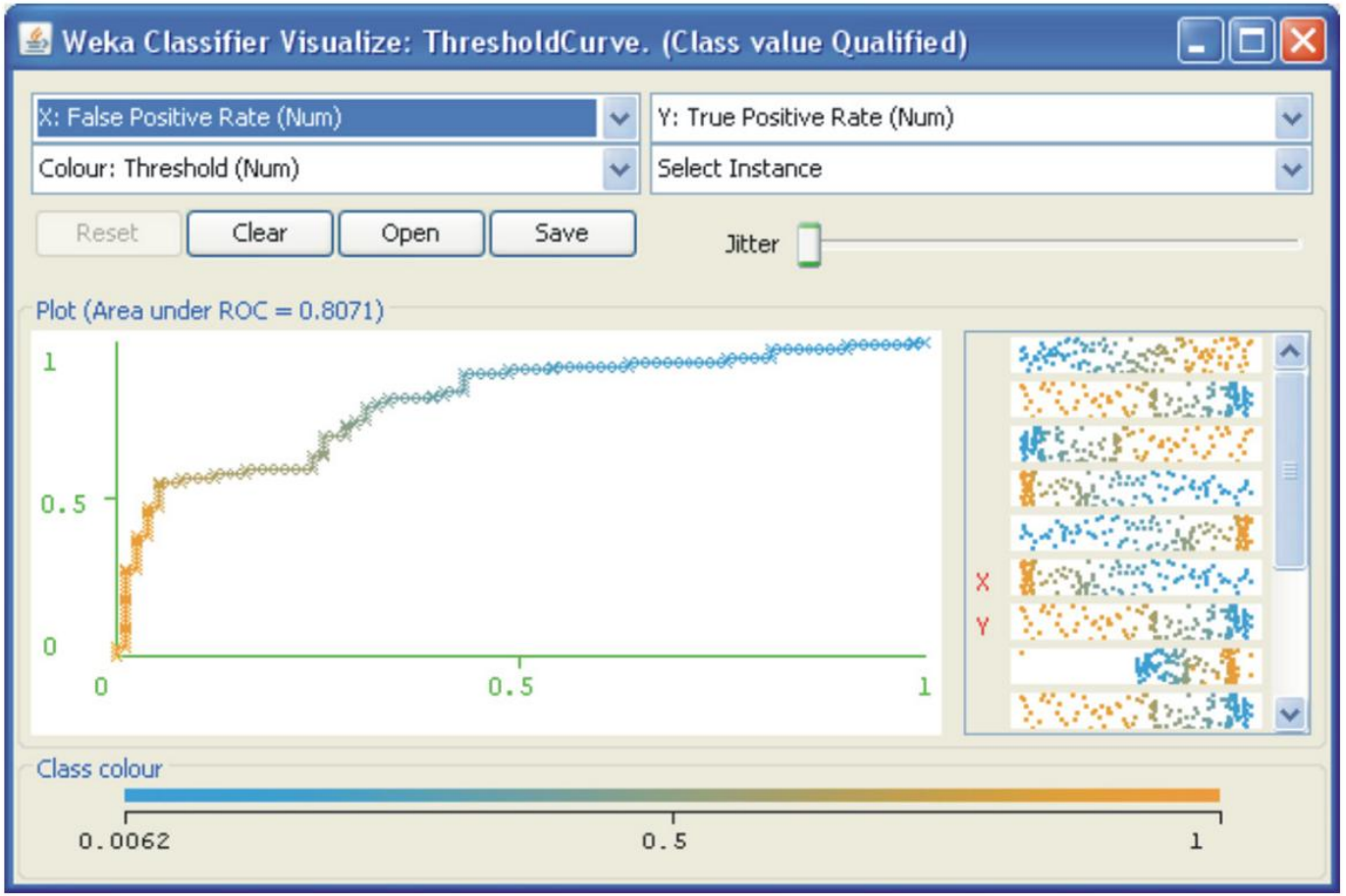
προέβλεψε σωστά τις 106 παρατηρήσεις (ποσοστό 70,6667%). Αναλυτικότερα, προέβλεψε σωστά τις 53 από τις 76 "Qualified"εταιρείες (ποσοστό 69.7%) και τις 53 από τις 74 "Unqualified"εταιρείες (ποσοστό 71.6%). Τα στοιχεία αυτά παρουσιάζονται στο confusion matrix και στην αναλυτική ακρίβεια ανά κλάση, στη στήλη "TP Rate".



Δένδρο Αποφάσεων



Καμπύλες ROC



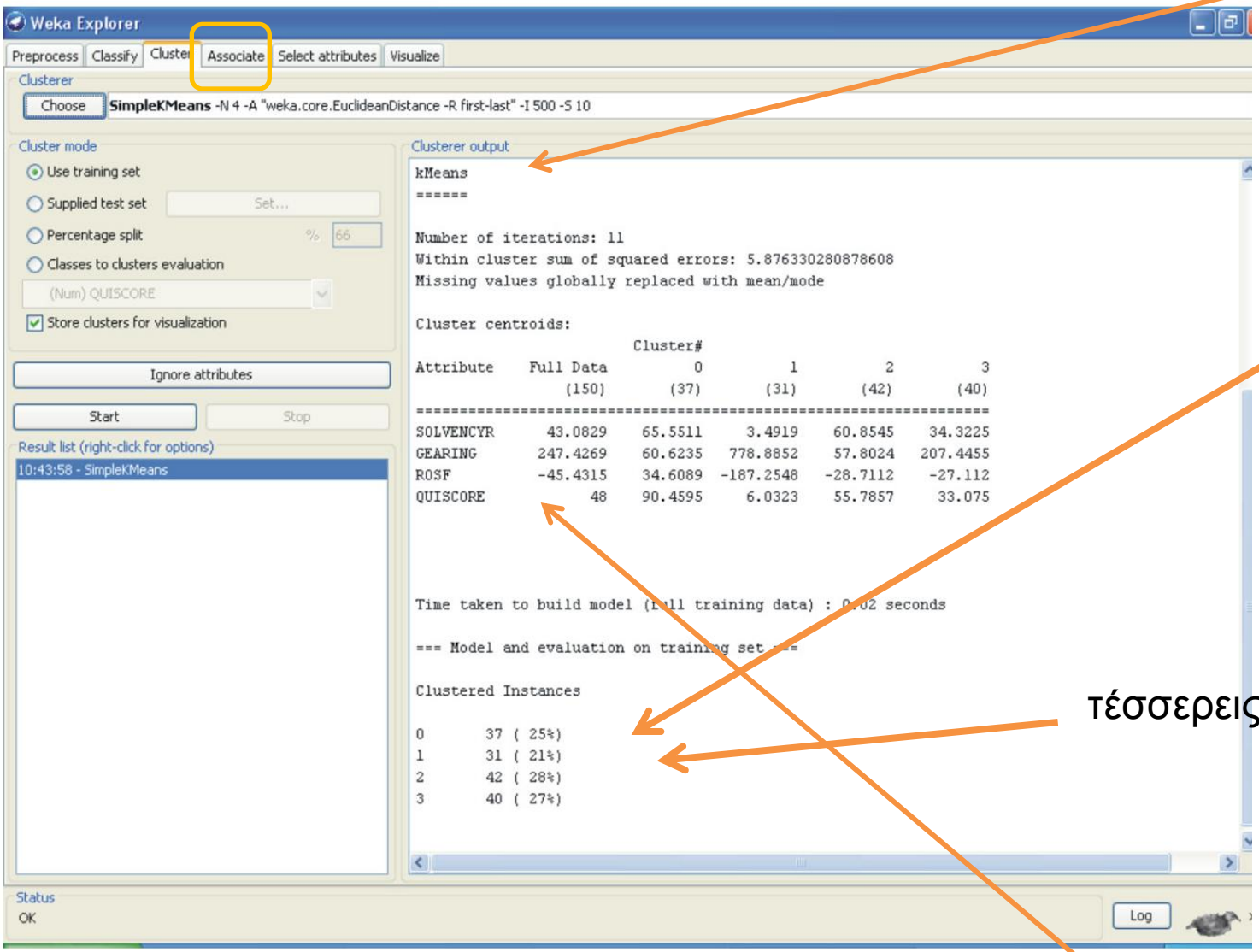
Ανάλυση σε συστάδες

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

150 επιχειρήσεις

Ανάλυση σε συστάδες

k-Means

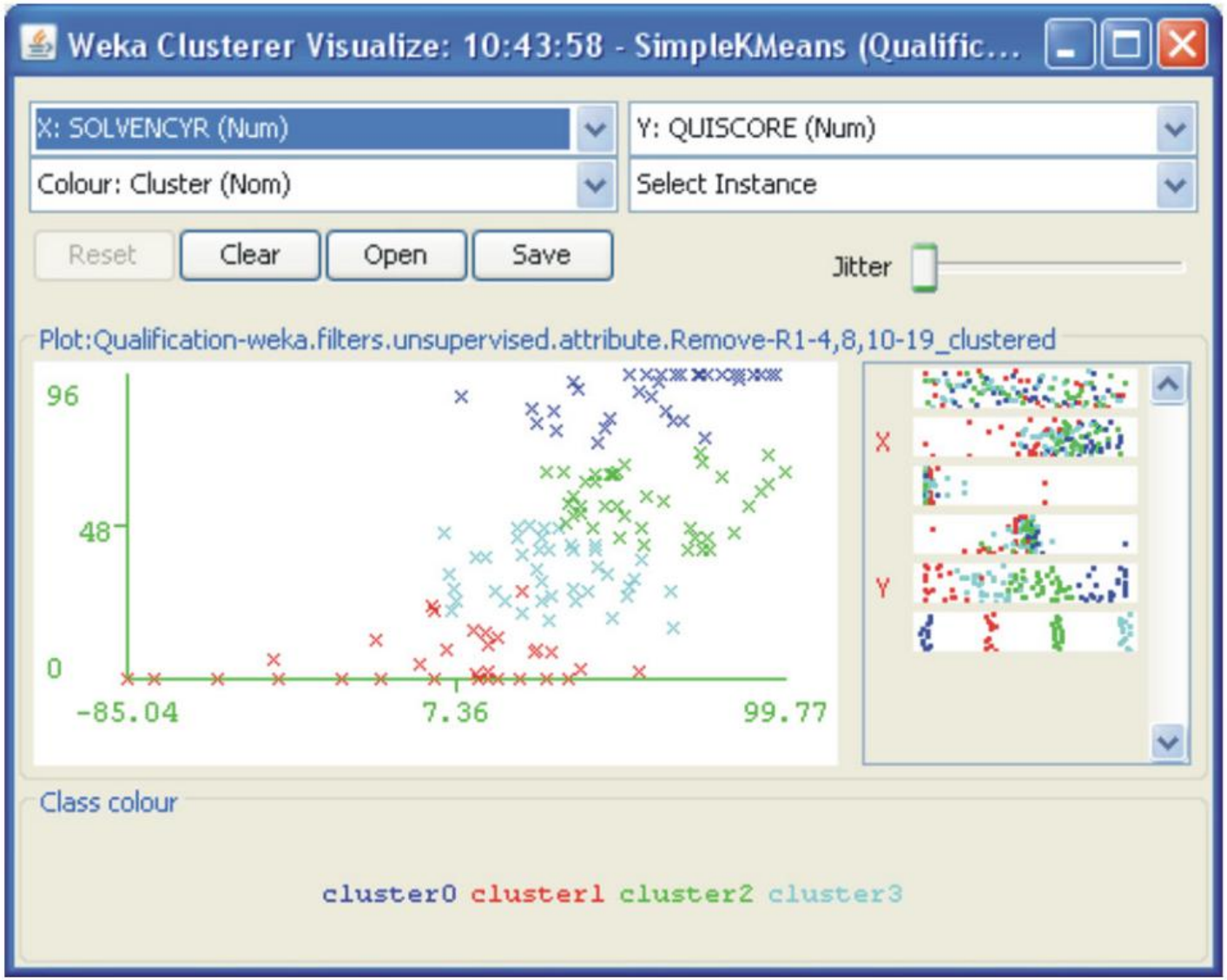


Η πρώτη συστάδα περιλαμβάνει 37 παρατηρήσεις, η δεύτερη 31, η τρίτη 42 και η τέταρτη 40.

τέσσερεις συστάδες

αριθμοδείκτης αξιοπιστίας

Ανάλυση σε συστάδες



Οπτικοποιηση