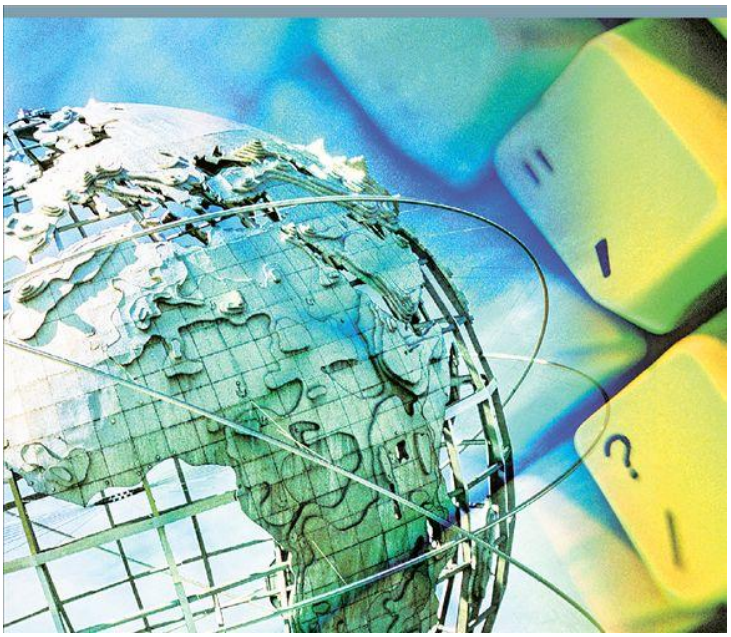
Α.Τ.Ε.Ι. ΠΑΤΡΑΣ

Σχολή Διοίκησης και Οικονομίας

Τμήμα Επιχειρηματικού Σχεδιασμού και Πληροφοριακών Συστημάτων

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Θέμα: «Κανόνες Συσχέτισης της Εξόρυξης Δεδομένων: Θεωρία και Αλγόριθμοι»



ΑΘΑΝΑΣΙΑΔΗ ΘΩΜΑΗ

ΠΑΠΑΧΡΙΣΤΟΠΟΥΛΟΥ ΙΩΑΝΝΑ

ΣΤΕΦΑΝΗ ΔΗΜΗΤΡΑ

ΕΠΟΠΤΕΥΩΝ ΜΑΣΤΡΟΓΙΑΝΝΗΣ ΝΙΚΟΛΑΟΣ

ΠΑΤΡΑ, Μάρτιος 2013

ΕΥΧΑΡΙΣΤΙΕΣ

Θα θέλαμε σε αυτή την παράγραφο να ευχαριστήσουμε θερμά το καθηγητή μας κύριο Μαστρογιάννη για την καλή συνεργασία σε αυτό μας το εγχείρημα όπως επίσης για την υποστήριξη, την βοήθεια και την καθοδήγηση του με τις πολύτιμες και εύστοχες υποδείξεις του κατά την διάρκεια και ολοκλήρωση αυτής της πτυχιακής εργασίας.

Αθανασιάδη Θωμαή
Παπαχριστοπούλου Ιωάννα
Στεφανή Δήμητρα

ΠΕΡΙΕΧΟΜΕΝΑ

Πρόλογος

Στον 21 αιώνα η τεχνολογία μέρα με την μέρα εξελίσσεται προσπαθώντας να κάνει την ζωή των ανθρώπων πιο εύκολη. Οι γρήγορες εξελίξεις οδηγούν στην εύρεση πιο αποδοτικών και αποτελεσματικών τρόπων συλλογής , αποθήκευσης καθώς και ανάλυσης των δεδομένων. Λόγω της ραγδαίας και απότομης εξέλιξης της τεχνολογίας θωρείται απαραίτητο και σημαντικό να πραγματοποιείται λεπτομερής ανάλυση και καλή ερμηνεία των δεδομένων που καταχωρούνται σε διάφορα αρχεία ή βάσεις δεδομένων. Γνωστός είναι ο σκοπός ότι εξάγουμε τα δεδομένα που είναι απαραίτητα και χρήσιμα για τη διαδικασία λήψης αποφάσεων πάνω σε ποικίλα θέματα. Η διαδικασία που χρησιμοποιείται για να πραγματοποιηθεί η ανάλυση και η ερμηνεία δεδομένων γίνεται με την εξαγωγή δεδομένων. Με τον όρο εξόρυξη δεδομένων χρησιμοποιείται ως συνώνυμο της ανακάλυψης γνώσης από βάσεις δεδομένων, καθώς επίσης και για αναφορά στις πραγματικές τεχνικές που χρησιμοποιούνται για την ανάλυση και εξαγωγή της από διάφορα σύνολα δεδομένων (Γιώργος Δάρδανος 2005). Η ανακάλυψη γνώσης από μια βάση δεδομένων (Knowledge Discovery in Databases – KDD) αναφέρεται σε ολόκληρη την διαδικασία εξόρυξης γνώσης από μεγάλα σύνολα δεδομένων. Με τους κανόνες συσχέτισης λαμβάνονται πληροφορίες πιθανόν χρήσιμες και παράλληλα εύκολα κατανοητές προς τον τελικό χρήστη. Μέσω των κανόνων αυτών εμφανίζονται τυχών ‘κρυμμένες’ συσχετίσεις μεταξύ των γνωρισμάτων από ένα σύνολο δεδομένων. Η εξαγωγή κανόνων συσχέτισης θεωρείται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων. με την χρήση αλγορίθμων επιτυγχάνεται η καλύτερη περιγραφή και η αξιολόγηση των δεδομένων για εξαγωγή συμπερασμάτων.

Συγκεκριμένα αποτελείται από τα ακόλουθα κεφάλαια:

Κεφάλαιο 1: Στο συγκεκριμένο κεφάλαιο θα αναφερθούμε γενικά για την εξόρυξη δεδομένων καθώς και την χρησιμότητα αυτής.

Κεφάλαιο 2: Στο σημείο αυτό θα αναφερθούμε γενικά για την συσχέτιση δεδομένων καθώς και για τους αλγόριθμους συσχέτισης ειδικότερα στον αλγόριθμο AprioriTID.

Κεφάλαιο 3: Το κεφάλαιο αυτό είναι άρρηκτα συνδεδεμένο με το προηγούμενο καθώς θα εφαρμόσουμε τον αλγόριθμο, περιγραφή δεδομένων , του λογισμικού και των αποτελεσμάτων.

Κεφάλαιο 4: Με την ολοκλήρωση του κεφαλαίου θα περιγράψουμε τα αποτελέσματα της εκπόνησης αυτής της εργασίας.

Κεφάλαιο 1

Εισαγωγή στην εξόρυξης δεδομένων

1.1 Γενικά

Τις τελευταίες δεκαετίες η διαδικασία της εξόρυξης γνώσης από βάσεις δεδομένων έχει αναπτυχτεί και συνεχίζει να αναπτύσσεται με ραγδαίο ρυθμό. Η ανάπτυξη αυτή έχει οδηγήσει στη παραγωγή μεγάλων όγκων πληροφοριών, τα δεδομένα αυτά αποθηκεύονται σε μεγάλες αποθήκες (βάσεις) δεδομένων, όπως δεδομένα διαδικτύου, αγορές σε πολυκαταστήματα/ super markets, τραπεζικές συναλλαγές, συναλλαγές πιστωτικών καρτών τα οποία αυτά δεδομένα είναι διαθέσιμα προς εξέταση. Αυτό συντέλεσε στη δημιουργία της διαδικασίας **εξόρυξης δεδομένων** (*Data Mining*). Η διαδικασία αποτελείται από τεχνικές στηριζόμενες σε αλγορίθμους ενώ βρίσκει χρήση σε τομείς όπως : στατιστικής, παγκοσμίου ιστού, στην ιατρική και τη μετεωρολογία. Έχοντας σκοπό την εύρεση χρήσιμων αποτελεσμάτων προς το τελικό χρήστη . Ο κυρίως λόγος που γίνεται αυτή η διαδικασία είναι για την εξαγωγή «γνώσης» που θα βοηθήσει στην πιο αποτελεσματική λήψη αποφάσεων.

Στο παγκόσμιο ιστό λόγο της χαοτικής και ραγδαίας ανάπτυξης οι δικτυακές εφαρμογές που διαχειρίζονται τις αποθήκες δεδομένων προσπαθώντας να βελτιώσουν τη ποιότητα των υπηρεσιών που προσφέρουν, κάνουν χρήση διαφόρων τεχνικών εξόρυξης δεδομένων. Ενώ παράλληλα η πρόοδος στη τεχνολογία τα τελευταία χρονιά έχει δημιουργήσει την ανάγκη δημιουργίας μεγάλων βάσεων δεδομένων λόγω της τεράστιας παραγωγής όγκου δεδομένων. Η προσπάθεια ανεύρεσης «χρήσιμης» γνώσης ως αποτέλεσμα από την ανάλυση των δεδομένων δημιούργησε την ανάγκη για την εύρεση νέας γενιάς εργαλείων και τεχνικών για την ανάλυση τους.

1.2 Ορισμός

Ένας νέος τομέας που μπαίνει στην ζωή του ανθρώπου. Μια αποθήκη δεδομένων από την οποία ο άνθρωπος ψάχνει να ανακαλύψει πρότυπα, πληροφορίες που παρουσιάζονται με διαφορετικές μορφές ώστε ο χρήστης ταχύτατα να μπορεί να αντλεί πληροφορίες που θα είναι χρήσιμες για αυτόν.

Η ανακάλυψη νέας και χρήσιμης πληροφορίας είναι ο σκοπός της εξόρυξης γνώσης από βάσεις δεδομένων, ωστόσο όμως οι τρόποι για την επίτευξη αυτού του στόχου διαφέρουν. Κάποιες από τις μεθόδους θα αναλυθούν παρακάτω, αναφέρονται κάποιες ενδεικτικά : τα δέντρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (neural networks), η Beyesian κατηγοριοποίηση. Οι μέθοδοι αυτοί χρησιμοποιούνται ώστε να βρεθούν συσχετίσεις και πρότυπα σε βάσεις δεδομένων.

Ο ορισμός αυτός αναφέρεται στα χρήσιμα πρότυπα ο βασικός σκοπός είναι η περιγραφή και η πρόβλεψη, η διερεύνηση και η ανάλυση των δεδομένων, με αυτοματοποιημένες και ήμι-αυτοματοποιημένες διαδικασίες, με στόχο την ανακάλυψη χρήσιμων προτύπων. Δηλαδή τα πρότυπα που υπάρχουν σε κάποιο σύνολο δεδομένων έτσι ώστε να μπορεί γίνει στο μέλλον κάποια πρόβλεψη για την συμπεριφορά ή την αξία κάποιων μεταβλητών. Η διαδικασία KDD είναι μια διαδικασία που αποτελείται από πολλά βήματα. Πιο αναλυτικά την προ- επεξεργασία των δεδομένων, την αναζήτηση των προτύπων και την αξιολόγηση της παραγόμενης γνώσης.

1.3 Εξόρυξη γνώσης από βάσεις δεδομένων

Ο όρος εξόρυξης δεδομένων έχει διχάσει αρκετά το ερευνητικό κοινό ως προς το αν είναι κατάλληλος για να περιγράψει τη διαδικασία που αντιπροσωπεύει παρόλα αυτά έχει επικρατήσει η χρήση του ώστε να περιγράψει τη διαδικασία εύρεσης γνώσης που συνδέεται με τα δεδομένα. Η Εξόρυξη Γνώσεων από Δεδομένα (Data Mining) στοχεύει στην αυτόματη παραγωγή νέας γνώσης που δεν ήταν γνωστή πριν και θα μας είναι χρήσιμη για το μέλλον.

Ως Data Mining μπορούμε να ορίσουμε ως εξής: Date Exploration → Knowledge Mining = εξερεύνηση μεγάλων Βάσεων Δεδομένων επιχειρησιακών αρχείων με στόχο να εξάγουμε και να αποκτήσουμε από αυτά νέας γνώσης, χρήσιμη για τον χρήστη.

Με την χρήση αλγορίθμων παρουσιάζεται η γνώση που έχει εξορυχτεί από ένα σύνολο δεδομένων. Υπάρχει ποικιλία αλγορίθμων που ο καθένας έχει το δικό του τρόπο χρησιμότητας και κάθε τομέας χρησιμοποιεί τον αντίστοιχο αλγόριθμο που έχει σχεδιαστεί για το ανάλογο μέγεθος των δεδομένων.

Η επιλογή συνόλου δεδομένων από αποθήκη δεδομένων αναφέρεται σε μια αποθήκη με δεδομένα από πολλές ετερογενείς πηγές (εσωτερικές και εξωτερικές), οργανωμένα σε ένα ενιαίο σχήμα σε ένα υπολογιστικό σύστημα, τα οποία βοηθούν στην λήψη αποφάσεων. Παρέχουν την απαραίτητη υποδομή για την οργάνωση, αποθήκευση αλλά και την εξαγωγή ποσοτήτων από δεδομένα ακλουθώντας τα πρότυπα, έτσι ώστε να χρησιμοποιηθούν για την ανεύρεση χρήσιμης γνώσης. Υπάρχουν δύο είδη αποθήκευσης, οι εσωτερικές αποθήκες δεδομένων και οι εξωτερικές.

Για τη διαδικασία της εξόρυξης δεδομένων υπάρχουν αρκετοί αλγόριθμοι έκ των όποιων κάποιοι από αυτούς χρησιμοποιούν τεχνικές ή διαδικασίες από τομείς, όπως τη στατιστική, αλγορίθμους βάσεων δεδομένων κ.α.. Σημαντική διαφοροποίηση στους αλγορίθμους εξόρυξης δεδομένων είναι η έμφαση που έχει δοθεί στην εξελισημότητα τους, έχοντας άμεση σχέση με το μέγεθος του συνόλου των δεδομένων που εισάγονται για εξέταση.

Οι κανόνες συσχέτισης μας βοηθάνε ουσιαστικά και στο να βρούμε στοιχεία που αγοράζονται συχνά μαζί, σε ένα επίπεδο που μπορεί να υπάρξει συσχέτιση μεταξύ των προϊόντων αυτών. Οι κανόνες αυτοί χρησιμοποιούνται συχνά σε εμπορικά κέντρα, όπου τα στοιχεία τοποθετούνται κοντά το ένα στο άλλο έτσι ώστε οι χρήστες να αγοράζουν περισσότερα στοιχεία. Ένα παράδειγμα που είναι πολύ γνωστό είναι η περίφημη ιστορία των Wal-Mart (μπύρα-πάνες), όπου οι Wal-Mart μελέτησαν τα δεδομένα και βρήκαν ότι Αμερικανοί άνδρες που αγοράζουν πάνες, επίσης, έχουν την τάση να αγοράζουν και μπύρα. Έτσι, η Wal-Mart τοποθέτησαν δίπλα στις πάνες, τις μπύρες και οι πωλήσεις μπύρας αυξήθηκαν. Βλέπουμε λοιπόν ότι η δύναμη της εξόρυξης δεδομένων είναι μεγάλη και χάρη σε αυτήν έγινε μια τέτοια πρόβλεψη. Έναν άλλο παράδειγμα είναι στο Google οπού έχουμε την αυτόματη συμπλήρωση όταν πληκτρολογούμε μια λέξη. Επίσης στο Amazon μας προτείνουν στοιχεία με βάση την τρέχουσα θέση που βρισκόμαστε εκείνη την στιγμή και περιηγούμαστε στην αγορά.

1.4 Ποιότητα της γνώσης

Σημαντικό κομμάτι της εξόρυξης γνώσης είναι και η ποιότητα των αποτελεσμάτων που παράγονται εάν ληφθεί υπόψη ότι είναι δυνατόν να υπάρχουν εάν όχι εκατομμύρια, χιλιάδες

διαφορετικά πρότυπα. Το πρότυπο που δημιουργείται μπορεί να είναι ενδιαφέρον (interesting) όταν είναι έγκυρο, εύκολο προς κατανόηση και μπορεί να φανεί χρήσιμο. Ωστόσο η ποιότητα των προτύπων εξαρτάται άμεσα από την ποιότητα των δεδομένων τα οποία εξετάζονται.

Ο όρος «Ποιότητα (quality)» στην εξόρυξης δεδομένων μπορεί να περιγράφει από τα παρακάτω:

- Παρουσίαση της πραγματικής γνώσης που περιλαμβάνονται στα δεδομένα που αναλύονται. Όπως θα αναφερθούμε πιο κάτω τα δεδομένα μπορεί να κρύβουν πολλές και ενδιαφέρουσες πληροφορίες που μέσω των μεθόδων εξόρυξης γνώσης μπορούν να αποκαλυφθούν. Την εποχή που διανύουμε είναι πιο σημαντικό από ποτέ η εκμετάλλευση αυτής της κρυφής γνώσης.

- Συντονισμός (tuning) αλγορίθμων. Η δυνατότητα των αλγορίθμων και των τεχνικών που χρησιμοποιούνται, κάτω από διαφορετικές περιπτώσεις να εξάγουν διαφορετικά αποτελέσματα μπορεί να γίνει σημαντικό πρόβλημα, ως προς το πόσο κατάλληλη είναι η επιλογή της μεθόδου που θα χρησιμοποιηθεί για να παράγει κάποιο αποτέλεσμα.

- Επιλογή των πιο αντιπροσωπευτικών και με ενδιαφέρον για τα δεδομένα προτύπων. Κατά τη διάρκεια της εξόρυξης γνώσης παράγονται πολλά πρότυπα για αυτό το λόγο η επιλογή του προτύπου που είναι πιο αντιπροσωπευτικό σε κάθε περίπτωση γίνεται μια δύσκολη εργασία.

Η ποιότητα των προτύπων που δημιουργούνται από την εξόρυξης δεδομένων εξαρτάται από το στόχο που θέλει να επιτύχει ο ερευνητής.

Λαμβάνοντας υπόψη ότι η αξιολόγηση της ποιότητας είναι σημαντικό πρόβλημα, που πρέπει να επιλυθεί, οι τεχνικές της χρησιμότητας (usefulness) και της σχετικότητας (relevance) έλκουν πολύ το ενδιαφέρον του ερευνητή.

1.5 Διαχείριση της αβεβαιότητας στη διαδικασία εξόρυξης γνώσης

Η διαδικασία εξόρυξης γνώσης προσπαθεί να ανακαλύψει πρότυπα τα οποία να παρουσιάζουν ενδιαφέρον στο σύνολο των δεδομένων. Υπάρχουν αρκετοί τρόποι να αναπαρασταθεί η γνώση αυτή και εξαρτώνται άμεσα από τις τεχνικές που χρησιμοποιούνται. Οι τρόποι αυτοί ενδεικτικά φαίνονται παρακάτω ενώ θα αναλυθούν στη συνέχεια: οι κανόνες συσχέτισης, η κατηγοριοποίηση και η συσταδοποίηση.

Η διαχείριση της αβεβαιότητας (uncertainty) στη διαδικασία της εξόρυξης γνώσης είναι μια πλευρά η οποία είναι εξίσου σημαντική. Τα στοιχειά των δεδομένων της διαδικασίας KDD συνήθως τοποθετούνται σε ομάδες με μοναδικό σκοπό να ανήκουν σε μια και μονό κατηγορία. Η κατηγοριοποίηση γίνεται σε κατηγορίες που έχουν ήδη οριστεί οι οποίες πολλές φορές δε περιγράφουν με ακρίβεια τα χαρακτηριστικά των δεδομένων. Το γεγονός αυτό μπορεί να οδηγήσει σε ανακριβή ή μη άρτια εξαγωγή γνώσης.

Το θέμα της διαχείρισης της αβεβαιότητας κεντρίζει το ενδιαφέρων των ερευνητών για μελέτη. Στη συνέχεια φαίνονται κάποια από τα σημαντικότερα σημεία για να γίνει κατανοητή η αβεβαιότητα στην ομαδοποίηση των δεδομένων:

- *Οι κατηγορίες δεν επικαλύπτονται*. Κάθε αντικείμενο των δεδομένων πρέπει να τοποθετηθεί σε μια και μόνο κατηγορία. Ωστόσο ένα αντικείμενο θα μπορούσε ανήκει σε περισσότερες από μια κατηγορίες.
- *Όλα τα στοιχειά αντιμετωπίζονται ανάλογα στην διαδικασία κατηγοριοποίηση*. Κάθε αντικείμενο πρέπει να ενταχτεί σε μια από τις υπάρχουσες κατηγορίες. Αυτό από μόνο του οδηγεί στη βεβαιότητα ότι κάποια γενική κατηγορία δε θα μπορούσε να περιγράφει κάθε αντικείμενο το ίδιο αξιόπιστα.

Το θέμα αυτό της αβεβαιότητας μπορεί να μετριαστεί με τη χρήση της κατηγοριοποίησης σε πιθανοτικές και ασαφείς προσεγγίσεις.

Η ασαφής προσέγγιση εκτιμά το βαθμό που ένα αντικείμενο ανήκει σε μια κατηγορία ενώ οι πιθανοτικές το αν το αντικείμενο ανήκει στη κατηγορία ή όχι.

1.6 Η χρησιμότητα της εξόρυξης δεδομένων

Ο λόγος που χρησιμοποιείται η Εξόρυξη Δεδομένων είναι για την ανάλυση βάσεων δεδομένων και να γίνει χρήση της εξαγόμενης γνώσης στη λήψη αποφάσεων:

i. Ανάλυση αγοράς και διαχείριση:

Στόχος της αγοράς

Διαχείριση πελατειακών σχέσεων

Ανάλυση καλαθιού αγοράς

Τμηματοποίηση της αγοράς

Εφαρμογή*: Τα καταστήματα μελέτησαν τον τρόπο επιλογής και αγοράς των προϊόντων των πελατών τους. Για παράδειγμα μελέτησαν πως οι καταναλωτές αγοράζουν πατατάκια αγοράζουν και μπύρες για αυτόν τον λόγο τα καταστήματα τοποθέτησαν μαζί αυτά τα προϊόντα.*

ii. Ανάλυση εταιρειών και διαχείριση ρίσκου:

Προβλέψεις

Διατήρηση πελατολογίου

Βελτιωμένη χρηματοδότηση (π.χ. τράπεζες)

Έλεγχος ποιότητας

Ανάλυση ανταγωνιστικότητας

Εφαρμογή: *Οι τράπεζες κατασκευάζουν δέντρα αποφάσεων από ιστορικά στοιχεία τραπεζικών δανείων για την παραγωγή αλγορίθμων όπου τα αποτελεσμάτων αυτών θα δίνουν πληροφορίες αν ο υποψήφιος πελάτης ικανοποιεί τις απαιτήσεις για να πάρει δάνειο.*

iii. Εντοπισμός απάτης και διαχείριση:

Άλλες εφαρμογές που χρησιμοποιούν Εξόρυξη Δεδομένων:

Εξόρυξη κειμένου

Ευφυής απαντήσεις σε ερωτήματα

Εφαρμογή: *Χρησιμεύει στον εντοπισμό ανθρώπων που σκηνοθετούν ατυχήματα για να λάβουν το χρηματικό ποσό από τις ασφαλιστικές τους εταιρείες, γιατροί που γράφουν φάρμακα ή εξετάσεις χωρίς ο ασθενής να τα χρειάζεται.*

1.7 Εξόρυξης γνώσης από το παγκόσμιο ιστό

Ο παγκόσμιος ιστός (ΠΙ-WWW) τα τελευταία χρόνια αναπτύσσετε ραγδαία και έχει γίνει το πιο δημοφιλές μέσω επικοινωνίας. Κάθε μέρα αυξάνεται κατά εκατομμύρια νέες ιστοσελίδες και εφαρμογές. Ακόμα, τα τελευταία χρόνια αναπτύσσεται ραγδαία το ηλεκτρονικό εμπόριο ένας πρακτικός και γρήγορος τρόπος συναλλαγών, αγοράς και πώλησης καθώς και online υπηρεσίες. .Ο αυξανόμενος αυτός ρυθμός της χρήσης του παγκόσμιου ιστού οδηγεί στην υιοθέτηση νέων τεχνικών επεξεργασίας των δεδομένων. Το κύριο μέλημα είναι ο τρόπος που θα εντοπιστεί η σχετική πληροφορία που αναζητά ο χρήστης.

Με τον όρο *Παγκόσμιος Ιστός* εννοούμε το μέσο για την ανάκτηση της πληροφορίας που διατίθεται μέσω του Διαδικτύου. Η προσπάθεια εξόρυξης γνώσης από το παγκόσμιο ιστό (web mining) είναι μια περίπλοκη διαδικασία που εμπλέκει τεχνικές και από άλλες περιοχές ανάκτηση πληροφοριών, τεχνητή νοημοσύνη βάσεις δεδομένων κ.α.. Επίσης υπάρχει δυνατότητα να επιχειρηθεί μελλοντική πρόβλεψη για τις ανάγκες των χρηστών

Η εξόρυξης δεδομένων από τον Παγκόσμιο Ιστό όπως έχει οριστεί από τους Χαλκίδη και Βαζιργιάννη «*η χρήση τεχνικών εξόρυξης για την αυτόματη ανακάλυψη και εξαγωγή δεδομένων από κείμενα και υπηρεσίες του παγκοσμίου ιστού».*

Η συνεχής χρήση του ΠΙ σήμερα κρύβει ανεπεξέργαστη πληροφορία, αφού τα δεδομένα που υπάρχουν δεν είναι δομημένα και ο χρήστης συχνά δεν αντλεί την διαθέσιμη πληροφορία

που του εμφανίζεται. Έτσι μελετώντας τις κινήσεις στα αρχεία που επισκέπτεται και την συμπεριφορά του χρήστη μπορούμε να εξάγουμε σημαντικές πληροφορίες που θα βοηθήσουν στην περεταίρω μελέτη. Η διαδικασία αυτή μας βοηθάει με την επεξεργασία των δεδομένων που έχουμε συλλέξει από διάφορες ιστοσελίδες, να αντιμετωπίσουμε προβλήματα που μπορεί να εμφανίζονται είτε τώρα είτε μεταγενέστερα. Στην αρχή γίνεται η μορφοποίηση των δεδομένων ώστε να μπορούν εύκολα να επεξεργαστούν. Αυτό επιτυγχάνεται με τρείς τεχνικές ομαδοποίησης.

Υπάρχουν τρεις κατηγορίες που χωρίζονται τα δεδομένα του ΠΙ:

- Δεδομένα περιεχομένου: δομημένα δεδομένα σε εικόνες, κείμενα.
- Δεδομένα δομής: δεδομένα στον γράφο δηλαδή ιστοσελίδες που συνδέονται με υπερσυνδέσμους.
- Δεδομένα χρήσης: δεδομένα από την πλοήγηση του χρήστη στο web.

Στόχος της εξόρυξης γνώσης από τα δεδομένα είναι να κατανοηθεί η σχέση μεταξύ αυτών, να ταξινομηθούν ώστε αυτόματα να μπορεί να πραγματοποιείται η εξαγωγή γνώσης.

1.9 Ανακάλυψη γνώσης από βάσεις δεδομένων

Η ανακάλυψη γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases – KDD) ουσιαστικά αναφέρεται στη διαδικασία της εξόρυξης γνώσεις από αρκετά μεγάλες βάσεις δεδομένων, ουσιαστικά μας βοηθάει να μελετήσουμε και να βγάλουμε σημαντικά αποτελέσματα από μεγάλα σύνολα δεδομένων τα οποία είναι αρκετά περίπλοκα μεταξύ τους. Αναφέρεται στη πληροφορία που είναι αρκετά χρήσιμη και εξάγεται από ολόκληρη τη διαδικασία και θα προσελκύσει το ενδιαφέρον διαφορετικών ερευνητικών τομέων αξιολογώντας την εξαγόμενη γνώση. Η ανακάλυψη γνώσης θεωρείται συχνά ως συνώνυμο με την εξόρυξη γνώσης.

Για να διαχειριστούν περίπλοκα δεδομένα μελετητές, αναλυτές, στατιστικολόγοι και γενικά ερευνητές από διάφορους κλάδους, ανέπτυξαν και συνδύασαν μεθοδολογία και αλγόριθμους για την σωστή και αποτελεσματική διαχείριση των μεγάλων όγκων δεδομένων που συσσωρεύονται καθημερινά. Έτσι βλέπουμε ότι σε πολλούς επιστημονικούς τομείς όπως είναι η τεχνολογία των βάσεων δεδομένων, όπου τα αρχεία είναι καταχωρημένα σε μέσα

αποθήκευσης όπου βρίσκονται μεγάλοι όγκοι δεδομένων οι οποίοι αυξάνονται ραγδαία βέβαια, είναι χρήσιμο να ανακαλύπτεται χρήσιμη γνώση που θα διευκολύνει τον τελικό χρήστη.

1.9.1 Ορισμός

Η ονομασία αυτή της KDD (Knowledge Discovery in Databases) χρησιμοποιείται από το 1989 (πρώτο συνέδριο KDD) με στόχο να φανεί ότι η γνώση είναι το τελικό προϊόν μιας ανακάλυψης καθοδηγούμενης από τα δεδομένα. Ο όρος KDD χρησιμοποιείται για την περιγραφή ολόκληρης της διαδικασίας ανακάλυψης γνώσης από ένα σύνολο δεδομένων.

«KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων , ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα» (Frawley, Piatesky -Shashiro & Matheus 1991)

Με τον όρο δεδομένα εννοούμε στοιχεία συσχετίσεων από αριθμούς, κείμενα, γεγονότα που καθημερινά συσσωρεύονται σε βάσεις δεδομένων επιχειρήσεων, οργανισμών, τράπεζες, επιστημονικών τομέων. Τα πρότυπα είναι σύνολο αντικειμένων που επαναλαμβάνονται και με την βοήθεια συναρτήσεων περιλαμβάνουν πληροφορίες από τα μη επεξεργασμένα δεδομένα. Σε πολλούς τομείς συναντάμε πρότυπα όπως είναι στην αστρονομία, στην βιολογία σε διάφορες εφαρμογές όπως βιομετρικά, υπολογιστική ανατομία, εγκεφαλική δραστηριότητα. Τα πρότυπα αυτά χρειάζονται να είναι έγκυρα και συνεπείς στα πρώτα δεδομένα που έχουν συλλεχτεί, να έχουν κάποιο ενδιαφέρον ώστε ο χρήστης να μπορεί να πάρει κάποια απόφαση από όλη αύτη την διαδικασία. Ένα πρότυπο είναι ενδιαφέρον όταν είναι εύκολα κατανοητό σε ένα νέο σετ δεδομένων, πιθανόν χρήσιμο, καινούριο ή επικυρώνει μια υπόθεση που επιθυμεί να ελέγξει ο χρήστης. Ένα ενδιαφέρον πρότυπο αναπαριστά γνώση. Με την αξιολόγηση των προτύπων που θα εξαχθούν από την διαδικασία θα μπορεί ο χρήστης να λάβει κάποια απόφαση χρήσιμη προς αυτόν. Ο σκοπός των προτύπων είναι να προβλέψουν αλλά και να επεξηγήσουν με κατανοητό τρόπο την εξορυγμένη γνώση.

1.10 Διαδικασία KDD

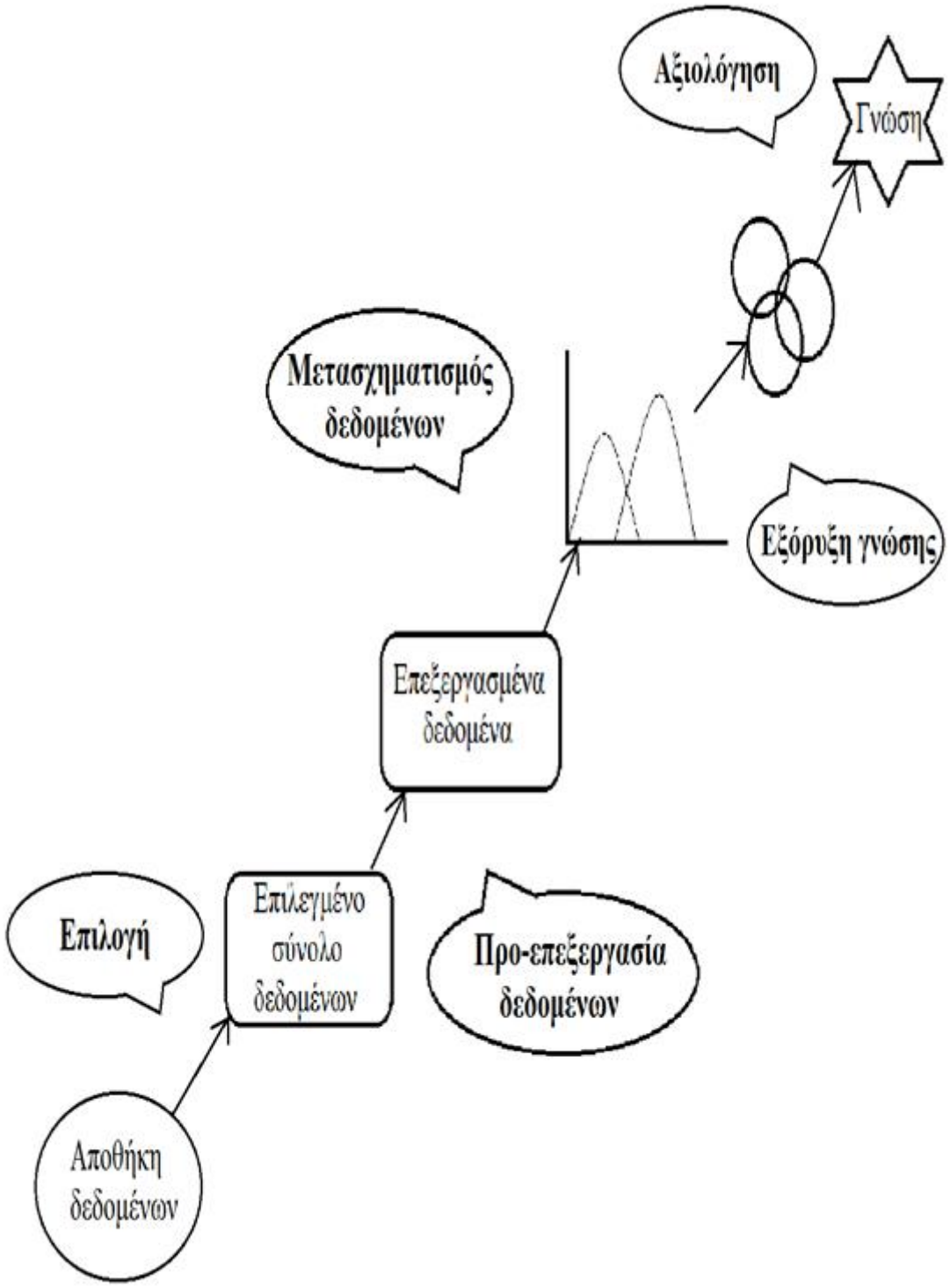
Η διαδικασία της KDD είναι μία αυτοματοποιημένη διαδικασία, μέσω της οποίας γίνεται προσπάθεια ανάλυσης και μοντελοποίησης μεγάλων αποθηκών δεδομένων. Αν επεξεργαστούμε μια τεράστια βάση δεδομένων υπάρχει μεγάλη πιθανότητα να ανακαλύψουμε την ύπαρξη «κρυμμένης γνώσης». Δηλαδή, μπορεί να βρούμε συσχετίσεις αλληλεξάρτησης ή ομαδοποίησης μεταξύ των δεδομένων, πράγματα τα οποία μπορεί να μην είναι άμεσα εμφανή. Αυτή η κρυμμένη γνώση θεωρείται ότι δεν είναι εκ των προτέρων διαθέσιμη, αλλά μπορεί να αποδειχθεί πολύ χρήσιμο αν την επεξεργαστούμε.

Η διαδικασία της ανακάλυψης γνώσης, η επεξεργασία δηλαδή των τεράστιων αυτών αποθηκών δεδομένων, η εύρεση έγκυρων και πρωτότυπων προτύπων μέσα από πολύ μεγάλους και περίπλοκους πίνακες δεδομένων, με στόχο τα πρότυπα που θα προκύψουν να είναι χρήσιμα και κατανοητά καθώς και η ανάλυση της γνώσης, επαναλαμβάνεται αρκετές φορές για να ανακαλύψουμε χρήσιμες γνώσεις που θα βοηθήσουν. Την KDD μπορούμε να την εφαρμόσουμε στον χώρο των επιχειρήσεων αναφέρουμε τις δραστηριότητες σε marketing, επενδύσεις, προσδιορισμό απειλών, βιομηχανική παραγωγή, τηλεπικοινωνίες, καθαρισμό δεδομένων.

Υπάρχουν πολλά μοντέλα της διαδικασίας αυτής, με πιο διαδεδομένο το μοντέλο KDD των 6 βημάτων, τα οποία είναι:

1) Εκμάθηση και καθορισμός περιοχής
2) Προ-επεξεργασία δεδομένων
3) Μετασχηματισμός δεδομένων
4) Εξόρυξη γνώσης
5) Αξιολόγηση
6) Παρουσίαση και χρήση της ανακαλυφθείσας γνώσης

Η διαδικασία KDD θεωρείται διαλογική και επαναληπτική, δηλαδή μπορούμε αν θέλουμε αλλά και αν μας απαιτηθεί να επιστρέψουμε σε ένα προηγούμενο βήμα. Το κυριότερο και πιο σημαντικό μοντέλο της διαδικασίας KDD είναι η διαδικασία εξόρυξης δεδομένων. Γενικά η εξόρυξη δεδομένων αλληλεπιδρά με την διαδικασία ανακάλυψης γνώσης. Το παρακάτω σχήμα μας δείχνει σχηματικά βήμα προς βήμα την διαδικασία ανακάλυψης χρήσιμης πληροφορίας από βάσεις δεδομένων.



«Εικόνα 1: Βήματα της διαδικασίας ανακάλυψης γνώσης»

1.10.1 Βήματα της KDD

Τα βήματα της διαδικασίας KDD είναι τα εξής:

- Εκμάθηση και καθορισμός περιοχής (Data cleaning):

Αρχικά εξετάζεται ο τομέας που πρόκειται να εφαρμοστεί καθώς και ο στόχος που θέλουμε να επιτύχουμε. Επίσης μπορούμε να κάνουμε υποθέσεις σχετικά με τα αποτελέσματα καθώς και να τα προβλέψουμε. Αφού καθορίσουμε την περιοχή που θα εξετάσουμε επιλέγουμε ένα σύνολο δεδομένων για την ανάλυση. Επιλέγουμε δηλαδή ένα σύνολο δεδομένων, μεταβλητών που θα τα διαχειριστούμε ώστε να καταφέρουμε να ανακαλύψουμε χρήσιμη πληροφορία με την βοήθεια διάφορων εργαλείων και ειδικών μέσων. Γίνεται η αφαίρεση των διάφορων που παράγουν θόρυβο, καθώς και τα άσχετα δεδομένα

- Προ-επεξεργασία δεδομένων (Data integration):

Το επόμενο βήμα της διαδικασίας KDD μετά την επιλογή των δεδομένων που θα χρειαστούμε είναι ο καθαρισμός και η προ-επεξεργασία αυτών των δεδομένων. Σε αυτό το στάδιο γίνονται οι βασικές λειτουργίες με την βοήθεια διάφορων μεθόδων που θα αναφερθούν στις επόμενες σελίδες, που θα βοηθήσουν στην συλλογή χρήσιμων πληροφοριών όπως είναι η μέτρηση και η αφαίρεση του θορύβου ή των outliers από τα δεδομένα, καθώς και ποιες στρατηγικές για τον τρόπο διαχείρισης των δεδομένων που υπολείπονται. Αναφέροντας τον όρο θόρυβο εννοούμε τα λάθη που μπορεί να υπάρχουν στα χαρακτηριστικά και τους αριθμούς των δεδομένων (ονόματα, κωδικοί) καθώς και στοιχεία που μπορεί να μην έχουν καμία σαφήνεια μεταξύ τους. Έτσι σε αυτό το βήμα, την προ-επεξεργασία, τα δεδομένα διαμορφώνονται με τέτοιο τρόπο ώστε να καταφέρει ο χρήστης να βγάλει σωστά και έγκυρα συμπεράσματα. Τα δεδομένα είναι συχνά ανομοιογενή επειδή πηγάζουν από πολλές διαφορετικές πηγές, ενώνονται σε μία βάση δεδομένων.

Λόγω του κινδύνου του θορύβου που κρύβεται στα δεδομένα, το βήμα αυτό αποτελεί το 60% όλης της διαδικασίας KDD και αυτό γιατί ο μη σωστός καθορισμός των στοιχείων μας παράγει μη-ποιοτικά δεδομένα τα οποία μπορούν να δημιουργήσουν πρόβλημα σε όλη την διαδικασία. Με τον όρο ποιοτικά δεδομένα εννοούμε τα κατηγορικά δεδομένα, δεδομένα που

έχουν μη μετρήσιμα χαρακτηριστικά με στόχο την εξήγηση, την ερμηνεία και την κατανόηση τους για να ανακαλύψουμε χρήσιμες πληροφορίες.

Μέθοδοι για την προ-επεξεργασία των δεδομένων:

Ο καθαρισμός των δεδομένων : Μια αυτοματοποιημένη μέθοδος που χρησιμεύει σε πολλές περιπτώσεις με δεδομένα που έχουν συλλεχτεί περιλαμβάνουν λάθη ή δεν παρουσιάζουν συμβατότητα μεταξύ τους λόγω του ότι δεν είναι άμεσα διαθέσιμα. Ο θόρυβος που εμφανίζεται εξαιτίας προβλημάτων, τεχνικών ή άλλων προβλημάτων που παρουσιάζονται κατά την διάρκεια εισαγωγής των δεδομένων οι λάθος τιμές που μπορεί να πάρουν τα δεδομένα ή η μεγάλη απόκλιση των δεδομένων μεταξύ τους, πρέπει να αφαιρεθεί. Σε μια επιχείρηση μπορεί για παράδειγμα να υπάρχουν διπλές εγγραφές για ένα πελάτη ή να μην υπάρχει τιμή και έτσι να εμφανίζεται ο όρος «unknown» που σημαίνει ότι το κελί είναι κενό και δεν περιέχει καμία τιμή.

Για να επιλύσουμε ένα τέτοιο πρόβλημα μπορούμε να χρησιμοποιήσουμε διάφορες τεχνικές όπως:

- Μέθοδος ταξινόμησης σε κουτιά: διαμερίζουμε τα δεδομένα σε ίσα κουτιά. Υπάρχουν δύο ειδών διαμερισμών. α) Διαμερισμός ίσου πλάτους: είναι μία άμεση τεχνική γιατί τα δεδομένα χωρίζονται σε ίση απόσταση και οι τιμές των δεδομένων παρουσιάζουν μια κλιμάκωση και β) Διαμερισμός ίσου βάθους: έχουμε τον ίδιο αριθμό δεδομένων και έτσι πετυχαίνουμε ομαλή κλιμάκωση.

- Ομαδοποίηση των δεδομένων: Ομαδοποίηση των δεδομένων με ομοια χαρακτηριστικά

- Παλινδρόμηση οπού με την βοήθεια της γραμμικής συνάρτησης Y= a + bX τα δεδομένα παρουσιάζονται σε γραφική παράσταση και μπορούμε να υπολογίσουμε και να προβλέψουμε για παράδειγμα το κέρδος σε μια εταιρία.

Μείωση δεδομένων : λόγω των μεγάλων όγκων δεδομένων που καθημερινά αποθηκεύονται δημιουργούνται προβλήματα αφού ο χώρος δεν επαρκή και παρουσιάζεται μεγάλη πολυπλοκότητα. Για να λυθεί αυτό το πρόβλημα μπορούμε να συναθροίσουμε τα δεδομένα σε μικρές ομάδες, να μειώσουμε τις διαστάσεις και να επιλέξουμε ένα δείγμα με το οποίο θα κάνουμε την ανάλυση ή μπορούμε να συμπιέσουμε τα δεδομένα ώστε να πετύχουμε μείωση του μήκους. Ακόμα μια μέθοδος είναι αυτή της ανάλυσης κύριων συνιστωσών δηλαδή αν έχουμε μεγάλο όγκο δεδομένων μας βοηθάει να επιλύσουμε OLAP συστήματα γιατί μπορούμε να παρουσιάσουμε σε πολυδιάστατη μορφή, σε μορφή κύβων τα δεδομένα και να εξασφαλίσουμε χώρο αλλά και χρόνο.

- Μετασχηματισμός δεδομένων

Μετασχηματίζονται ή μειώνονται τα δεδομένα με την μείωση του αριθμού των μεταβλητών, την δομή και την μορφή των δεδομένων. Τα επιλεγμένα δεδομένα τροποποιούνται ώστε η μορφή τους να είναι κατάλληλη για την διαδικασία της εξόρυξης. Βρίσκονται χρήσιμα χαρακτηριστικά και επιλέγουμε με ποιο τρόπο θα λειτουργήσει η εξόρυξη γνώσης, για να είναι πιο εύκολο να εξορύξουμε τα δεδομένα. Η εξομάλυνση του θορύβου από τα δεδομένα, η κανονικοποίηση, η χρήση των συστημάτων OLAP αλλά και η δημιουργία νέων τιμών με την βοήθεια παλαιότερων καταχωρίσεων είναι μέθοδοι με τους οποίους μπορούμε να πετύχουμε τον μετασχηματισμό.

Στη συνέχεια περιγράφονται τα βήματα που ακολουθούνται για την εύρεση και την παρουσίαση της εξαχθείσας γνώσης.

i. Εξόρυξη γνώσης:
   Δημιουργούμε το βέλτιστο μοντέλο με το οποίο θα αναπαριστάτε με τη βοήθεια αλγορίθμων, δέντρων απόφασης, κανόνων συσχέτισης, παλινδρόμησης η γνώση που θα εξαχθεί από την επεξεργασία των δεδομένων και θα την χρησιμοποιήσουμε για την εξόρυξη χρήσιμης γνώσης.
ii. Αξιολόγηση προτύπων (Pattern evaluation):
   Γίνεται η αναγνώριση των προτύπων, η αφαίρεση προτύπων που δεν μας είναι χρήσιμα, βρίσκουμε ποια είναι αυτά τα πρότυπα τα οποία παρουσιάζουν μια βεβαιότητα και μια ακρίβεια έτσι ώστε να έχουμε τα πρότυπα που παρουσιάζουν την γνώση που πραγματικά θέλουμε, βάσει συγκεκριμένων μέτρων αξιολόγησης (evaluation measures).

iii. Παρουσίαση της γνώσης:

Το τελευταίο βήμα της διαδικασίας KDD στο οποίο παρουσιάζεται η εξορυγμένη γνώση και αναπαριστάτε μέσω διάφορων τεχνικών απεικόνισης. Σε αυτό το στάδιο γίνεται και έλεγχος για τυχόν συγκρούσεων με άλλη εξορυγμένη γνώση από προηγούμενη εξόρυξη. Η γνώση έχει ανακαλυφθεί παρουσιάζεται στον χρήστη, ο οποίος μπορεί να ερμηνεύσει τα αποτελέσματα είναι που είναι καταγραμμένα λεπτομερώς και αποτελούν μια νέα βάση δεδομένων που μπορεί να χρησιμοποιηθεί σε μελλοντικά στάδια.

1.10.2 Στόχος της διαδικασίας

Στόχος όλης αυτής της διαδικασίας είναι με την ανάπτυξη διάφορων μεθόδων και τεχνικών να αξιολογηθούν σωστά τα δεδομένα ώστε να ερμηνευτούν και να προκύψει χρήσιμη αλλά και κατανοητή προς τον χρήστη γνώση, τόσο σε πρακτικό όσο και σε θεωρητικό επίπεδο. Ο χρήστης θα μπορεί να πάρει αποφάσεις και τα δεδομένα θα μπορούν να αξιοποιηθούν σε μελλοντικές περιστάσεις.

Προσδιορίζοντας τους στόχους που θέλουμε να πετύχουμε λαμβάνοντας υπόψη βέβαια και τους περιορισμούς καθώς και τους πόρους που έχουμε στην διάθεση μας, καταφέρνουμε να καθορίσουμε το πρόβλημα που θέλουμε να επιλύσουμε. Με τον σωστό έλεγχο των δεδομένων πριν την εξόρυξη οι στόχοι μας θα είναι πιο φερέγγυοι και έτσι οι ειδικοί θα μπορούν να ανακαλύψουν χρήσιμα αποτελέσματα από την αξιολόγηση των πρότυπων.

Αποθηκεύουν πλήθος στοιχείων με την εφαρμογή αλγορίθμων σε πραγματικές βάσεις δεδομένων. Εφαρμογές και τεχνικές εξόρυξης γνώσης. Επεξεργασία χιλιάδων δεδομένων που καθημερινά αποθηκεύονται για να παραχθεί χρήσιμη πληροφορία. Συνήθως ο όγκος των δεδομένων που αποθηκεύονται έχουν μέγεθος terabyte, έτσι με την διαδικασία εξόρυξης γνώσης καταφέρνουμε να επιλέξουμε ομαδοποιώντας τα δεδομένα και συσχετίζοντας μεταξύ τους να ανακαλύψουμε την χρήσιμη γνώση. Επιχειρήσεις που στόχο έχουν πως θα ωφεληθούν ώστε να πετύχουν το κέρδος καθώς και ερευνητικά κέντρα που στόχο έχουν την ανακάλυψη γνώσης.

Μπορούμε πολλές φορές να συνδυάσουμε κάποια από τα παραπάνω βήματα. Για παράδειγμα, τα βήματα του καθαρισμού και της ενσωμάτωσης των δεδομένων, μπορούν να

συνδυαστούν έτσι ώστε να δημιουργηθεί μια αποθήκη δεδομένων. Ακόμα συνδυάσουμε τα βήματα της επιλογής και της τροποποίησης των δεδομένων. Αυτή η συγχώνευση διάφορων βημάτων μας δείχνει ότι η εξόρυξη δεδομένων είναι διαδικασία-κλειδί για την ανεύρεση γνώσης.

Τέλος όπως αναφέραμε η διαδικασία ανεύρεσης γνώσης είναι επαναληπτική, ο χρήστης μπορεί να εκμεταλλευτεί αυτή την δυνατότητα ώστε να τροποποιήσει τα μέτρα αξιολόγησης, να τελειοποιήσει την διαδικασία της εξόρυξης, να επιλέξει νέα δεδομένα, να τροποποιήσει ακόμα περισσότερο τα δεδομένα που υπάρχουν ή να ενσωματώσει στη βάση νέα δεδομένα από καινούργιες πηγές με τελικό στόχο την εξαγωγή πιο κατάλληλων και αποδοτικά αποτελεσμάτων.

1.11 Τεχνικές εξόρυξης

Τα τελευταία χρόνια έχουν εξελιχθεί διάφορες τεχνικές και μέθοδοι εξόρυξης δεδομένων. Λόγω των τεράστιων βάσεων δεδομένων χρησιμοποιούνται διαφορετικά κριτήρια κατηγοριοποίησης των μεθόδων και των συστημάτων εξόρυξης δεδομένων ανάλογα με τις τεχνικές που θα χρησιμοποιηθούν, το είδος των βάσεων δεδομένων και το είδος της γνώσης που θα εξαχθεί.

Ένα σύστημα βάσεων δεδομένων κατηγοριοποιείται με βάση τους τύπους των συστημάτων των βάσεων δεδομένων, την εξαγόμενη γνώση και τις τεχνικές που χρησιμοποιούνται για την εξόρυξη δεδομένων. Οι στόχοι για την εξόρυξη δεδομένων είναι η πρόβλεψη και η περιγραφή.

Ο στόχος της πρόβλεψης είναι ο υπολογισμός της μελλοντικής αξίας ή η συμπεριφορά κάποιων μεταβλητών που παρουσιάζουν ενδιαφέρον και οι οποίες βασίζονται στη συμπεριφορά άλλων μεταβλητών.

Η περιγραφή δίνει έμφαση στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με έναν κατανοητό και αξιοποιήσιμο τρόπο .

Η περιγραφή είναι πολύ πιο σημαντική από την πρόβλεψη διότι η περιγραφή επικεντρώνεται στην ανακάλυψη προτύπων με τρόπο κατανοητό και αξιοποιήσιμο με σκοπό να εξαχθούν συμπεράσματα όπου θα συμβάλλουν στην λήψη αποφάσεων.

Υπάρχει ένας αριθμός μεθόδων εξόρυξης δεδομένων όπου όλοι οι μέθοδοι έχουν σκοπό τον προσδιορισμό, περιγραφή και την αξιολόγηση χρήσιμων και κατανοητών προτύπων ώστε να ληφθούν οι κατάλληλες αποφάσεις.

Οι μέθοδοι οι οποίοι θα αναλύσουμε εκτενέστερα είναι η κατηγοριοποίηση και η συστηματοποίηση.

1.11.1 Κατηγοριοποίηση

Η κατηγοριοποίηση είναι μια σημαντική διαδικασία της εξόρυξης δεδομένων που έχει σαν σκοπό τη δημιουργία του μοντέλου που χρησιμοποιείται να κατηγοριοποιηθούν τα δεδομένα των οποίων η κατηγοριοποίηση δεν είναι γνωστή.

Η κατηγοριοποίηση αναλύεται ως μια διαδικασία που αποτελείται από δύο βήματα.

Το πρώτο βήμα ονομάζεται εκμάθηση. Στην εκμάθηση αναλύεται ένα σύνολο κατηγοριών δεδομένων εκπαίδευσης (training data). Τα δεδομένα εκπαίδευσης μελετώνται από έναν αλγόριθμο κατηγοριοποίησης με σκοπό τη δημιουργία μοντέλου. Τα στοιχεία επιλέγονται τυχαία από ένα πληθυσμό δειγμάτων. Το μοντέλο ονομάζεται κατηγοριοποιητής και μελετάται με τους κανόνες κατηγοριοποίησης, δέντρα αποφάσεων και διάφορους μαθηματικούς τύπους.

Το δεύτερο βήμα είναι η κατηγοριοποίηση όπου πρέπει να αξιολογηθεί το μοντέλο. Η αξιολόγηση μπορεί να πραγματοποιηθεί με διάφορες μεθόδους. Ουσιαστικά κατηγοριοποιούνται τυχαία δείγματα με δικά μας δεδομένα και μετέπειτα το μοντέλο συγκρίνει την κατηγορία που ανήκουν τα δεδομένα με την πρόβλεψη που έγινε μέσω του μοντέλου μας. Αν το μοντέλο είναι αξιόπιστο τότε χρησιμοποιείται για την ταξινόμηση μελλοντικών δειγμάτων.

Στο παρακάτω πίνακα παρουσιάζεται ένα παράδειγμα κατηγοριοποίησης.

Πίνακας δεδομένων κατηγοριοποίησης στις κατηγορίες

Όνομα	Φύλο	Οικογενειακή κατασταση	Ιδιοκτήτης	Φορολογητέο εισόδημα	Πληρωτέοι φόροι
Κατερίνα	Θ	Παντρεμένη	Ναι	50000	Ναι
Κώστας	Α	Ανύπαντρος	Όχι	100000	Όχι
Αναστασία	Θ	Ανύπαντρη	Ναι	40000	Όχι
Νίκος	Α	Παντρεμένος	Όχι	35000	Ναι
Μαρία	Θ	Παντρεμένη	Όχι	400000	Όχι
Χαρά	Θ	Ανύπαντρη	Ναι	135000	Ναι
Γιώργος	Α	Ανύπαντρος	Όχι	20000	Ναι
Εβίτα	Θ	Παντρεμένη	Ναι	34000	Όχι
Ρούλα	Θ	Ανύπαντρη	Ναι	456000	Ναι
Πασχάλης	Α	Ανύπαντρος	Όχι	23000	Όχι
Πέτρος	Α	Ανύπαντρος	Ναι	399000	Ναι

Οι πιο γνωστές μέθοδοι κατηγοριοποίησης ή αλλιώς μπορεί να χρησιμοποιηθεί το συνώνυμο ταξινόμηση είναι η Bayesian κατηγοριοποίηση, δέντρα απόφασης, νευρωνικά δίκτυα, Support Vector Machines(SVMs).

Bayesian κατηγοριοποίηση

Ο σκοπός της Bayesian κατηγοριοποίησης είναι η κατηγοριοποίηση ενός δείγματος X σε μια από τις κατηγορίες $C_1, C_{2 \ldots} C_n$. Αυτό πραγματοποιείται με τη χρήση ενός μοντέλου πιθανότητας που ορίζεται σύμφωνα με τη θεωρία Bayes. Κάθε κατηγόρια χαρακτηρίζεται από μια εκ των προτέρων πιθανότητα παρατήρησης της κλάσης C_i όπου ανήκει το δείγμα που εξετάζεται.

Ο Naïve Bayesian είναι ο πιο απλός κατηγοριοποιητής όπου υποθέτει πως η επίδραση ενός χαρακτηριστικού σε μια υποτιθέμενη ή δεδομένη κατηγορία είναι ανεξάρτητη από τις τιμές άλλων χαρακτηριστικών.

Ένας ακόμη κατηγοριοποιητής είναι ο Bayesian Belief Networks όπου λαμβάνει υπόψη τις σχέσεις εξαρτήσεων που μπορεί να υπάρχουν μεταξύ των μεταβλητών.

Δέντρα αποφάσεων

Στη κατηγοριοποίηση έχουμε δεδομένα τα οποία έχουν εξαρχής γνωστές τάξεις. Είναι μια πολύ διαδεδομένη τεχνική εξόρυξης που ταξινομεί τα δεδομένα στις υπάρχουσες τάξεις και δημιουργεί πρότυπα. Ο αλγόριθμος κατηγοριοποίησης χρησιμοποιεί τα δεδομένα για να καθορίσει το σύνολο των παραμέτρων που χρειάζονται για περαιτέρω διάκριση δεδομένων. Στη συνέχεια κωδικοποιεί τα δεδομένα – χαρακτηριστικά σε ένα μοντέλο, που ονομάζεται κατηγοριοποιητής. Αφού δημιουργηθεί ένας αποτελεσματικός ταξινομητής, χρησιμοποιείται σαν πρόβλεψη, ώστε να ταξινομήσει νέα δεδομένα στις τάξεις. Οι αλγόριθμοι κατηγοριοποίησης διακρίνονται σε αλγόριθμους που παράγουν δέντρα απόφασης, σε λογιστική παλινδρόμηση, σε ταξινομητές Bayes, βασισμένους σε νευρωνικά δίκτυα και σε ταξινομητές SVM (Support Vector Machines). Τα δέντρα αποφάσεων παράγουν μία οπτική παρουσίαση των κανόνων, γεγονός το οποίο συμβάλλει σημαντικά στη διάδοση τους ως μέθοδο για ταξινόμηση. Τα δέντρα αποφάσεων είναι δυνατόν να χρησιμοποιηθούν στην ταξινόμηση, στην παλινδρόμηση, αλλά και για τη μείωση του όγκου δεδομένων μέσω του μετασχηματισμού τους σε μία πιο συμπιεσμένη μορφή, διατηρώντας όμως τα βασικά χαρακτηριστικά των δεδομένων. Τα δέντρα αποφάσεων αποτελούν την πιο διαδεδομένη μέθοδο για ταξινόμηση και γι’ αυτό παρουσιάζονται αναλυτικότερα παρακάτω.

Το δέντρο απόφασης κατασκευάζεται από το σύνολο εκπαίδευσης δηλαδή από ένα σύνολο δεδομένων/ εγγραφών. Κάθε εγγραφή χαρακτηρίζεται από το σύνολο χαρακτηριστικών και την τάξη. Η λογική της κατασκευής ενός δέντρου αποφάσεων είναι η σωστή και ακριβής σχέση των χαρακτηριστικών αυτών και της τάξης. Ένα δέντρο αποφάσεων περιέχει μηδενικούς ή περισσότερους ενδιάμεσους κόμβους και έναν ή περισσότερους τερματικούς κόμβους. Κάθε ενδιάμεσος κόμβος αποτελείται από δύο ή περισσότερους κόμβους-παιδιά. Όλοι οι ενδιάμεσοι κόμβοι περιέχουν διαιρέσεις, οι οποίες ελέγχουν την τιμή της έκφρασης των χαρακτηριστικών. Τέλος, ένας τερματικός κόμβος αποτελείται από μία τιμή τάξης.

Οι βασικοί αντικειμενικοί σκοποί των ταξινομητών δέντρων αποφάσεων είναι να ταξινομήσουν σωστά όσο το δυνατόν περισσότερο ποσοστό από το σύνολο εκπαίδευσης και να γενικεύσουν πέρα από το δείγμα εκπαίδευσης, έτσι ώστε ένα νέο και άγνωστο δείγμα

εκπαίδευσης να μπορεί να ταξινομηθεί με όσο το δυνατό μεγαλύτερη ακρίβεια, να μπορούν να ενημερώνονται, όταν διατεθούν περισσότερα δεδομένα και να έχουν όσο πιο απλή δομή γίνεται.

Στη στατιστική η τεχνική της εξαγωγής δέντρων αποφάσεων ξεκίνησε με τη δημιουργία ιεραρχικής ταξινόμησης για διερεύνηση ερευνητικών δεδομένων. Διάφορα στατιστικά προγράμματα, όπως το AID, το MAID, το THAID, και το CHAID κατασκεύασαν δυαδικά διαχωριστικά δέντρα, τα οποία αποσκοπούσαν στην ανακάλυψη των σχέσεων μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Στην αναγνώριση προτύπων τα δέντρα αποφάσεων χρησιμοποιήθηκαν στην επεξήγηση εικόνων από απομακρυσμένους δορυφόρους, όπως ο LANDSAT στη δεκαετία του 1970. Στην επιστήμη της μηχανικής μάθησης τα δέντρα αποφάσεων χρησιμοποιήθηκαν, προκειμένου να αποφευχθεί το «μποτιλιάρισμα» (bottleneck) της απόκτησης γνώσης για έμπειρα συστήματα. Τέλος, στη διαδοχική διάγνωση σφαλμάτων (sequential fault diagnosis) οι αλγόριθμοι που χρησιμοποιούνται παίρνουν συχνά τη μορφή δέντρων αποφάσεων.

Τα δέντρα απόφασης για την κατασκευή τους χρησιμοποιούν προ-κατηγοριοποιημένα δεδομένα και θεωρούνται από το χρήστη εύκολα στην κατανόηση. Κάθε κόμβος προσδιορίζει τον έλεγχο ενός χαρακτηριστικού και το κάθε κλαδί αντιπροσωπεύει την πιθανή τιμή για το συγκεκριμένο χαρακτηριστικό.

Η διαδικασία ξεκινά με τη ρίζα του δέντρου όπου εξετάζουμε τα χαρακτηριστικά που καθορίζονται από τον κόμβο και στη συνέχεια προσδιορίζουμε τους εσωτερικούς κόμβους ώστε να καταλήξουμε στην τελική μας απόφαση.Ενώ θεωρείται εύκολο στην κατανόηση καθώς και η γραφική τους απεικόνιση υπάρχει ένα μειονέκτημα ότι τα όταν τα δεδομένα είναι αριθμητικά τα δέντρα απόφασης πολύπλοκα και πολύ δύσκολα στην κατανόηση.

Οι αλγόριθμοι που χρησιμοποιούνται για την κατασκευή των δέντρων απόφασης είναι ID3, SLIQ, C4.5 και SPRINT.

Νευρωνικά δίκτυα

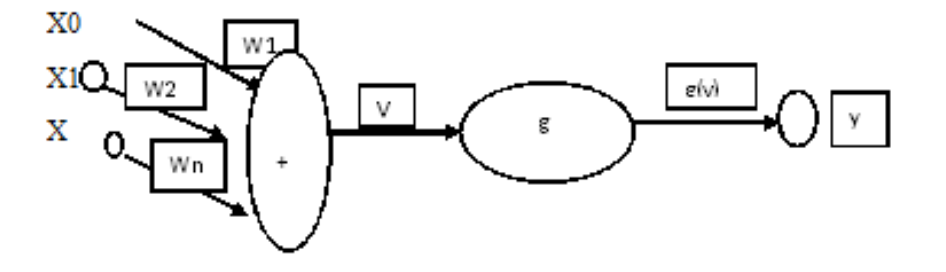
Τα νευρωνικά δίκτυα είναι ένα άλλο είδος προσέγγισης της κατηγοριοποίησης και βασίζεται στα νευρωνικά δίκτυα. Είναι μια δομή που αποτελείται από ένα δίκτυο νευρώνων οι οποίοι συνδέονται μεταξύ τους. Η πιο διαδεδομένη κατηγορία νευρωνικών δικτύων είναι τα λεγόμενα δίκτυα πρόσθιας τροφοδότησης, τα οποία επιτρέπουν την κίνηση των δεδομένων

μόνο προς μια κατεύθυνση, δηλαδή από μια είσοδο προς μια έξοδο. Δίκτυα που σχηματίζουν κυκλικές δομές ονομάζονται ανατροφοδοτούμενα νευρωνικά δίκτυα.

Η διαδικασία αυτή ξεκινά αναγνωρίζοντας τα χαρακτηριστικά εισόδου και εξόδου και με την κατάλληλη τοπολογία κατασκευάζεται το δίκτυο ώστε να επιλεχτεί το σωστό σύνολο εκπαίδευσης. Στη συνέχεια γίνεται έλεγχος του δικτύου όπου τελικά θα παραχθεί το μοντέλο που θα εφαρμόζεται για προβλέψεις κατηγοριών (έξοδοι) και των μη-κατηγοριοποιημένων δειγμάτων (είσοδοι).

Τα νευρωνικά δίκτυα αποτελούνται από "νευρώνες" με βάση τη νευρωνική δομή του εγκεφάλου. Τα λάθη από την αρχική κατηγοριοποίηση της πρώτης εγγραφής επανατροφοδοτούνται στο δίκτυο ενώ γίνεται η επεξεργασία των στοιχείων οπού "μαθαίνουν" την πραγματική κατηγοριοποίηση της κάθε εγγραφής και με αυτόν τον τρόπο την δεύτερη φορά χρησιμοποιούνται για την τροποποίηση των αλγορίθμων δικτύων. Η διαδικασία αυτή επαναλαμβάνεται. Οι νευρώνες ενός δικτύου χωρίζονται σε τρείς βασικές κατηγορίες: τους νευρώνες εισόδου, οι οποίοι δέχονται τις πληροφορίες που θα υποστούν επεξεργασία, τους νευρώνες εξόδου ,στους οποίους καταλήγουν τα αποτελέσματα της παραπάνω επεξεργασίας, και τους ενδιάμεσους νευρώνες, οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου. Οι τελευταίοι εναλλακτικά ονομάζονται και κρυφοί νευρώνες .Ουσιαστικά, οι νευρώνες σε ένα δίκτυο είναι αφενός ένα σύνολο εισερχόμενων τιμών και των αντίστοιχων βαρών τους, και αφετέρου μια συνάρτηση που αθροίζει τα παραπάνω βάρη, αντιστοιχώντας τα αποτελέσματα σε ένα νευρώνα εξόδου.

Δομή νευρωτικού δικτύου



«Εικόνα 2: Δομή νευρωτικού δικτύου»

Όπου x_i εισερχόμενες τιμές , w_i τα συσχετιζόμενα βάρη, g η συνάρτηση που αθροίζει τα βάρη και η y είναι η έξοδος, όπως φαίνεται στην Εικόνα 2.

Ο κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διάφορες πηγές, πραγματοποιεί τον υπολογισμό με βάση τις εισόδους και παράγεται μια έξοδος.

Τεχνική κοντινότερων γειτόνων

Στη μέθοδο αυτή τα στοιχεία κατηγοριοποιούνται βάση των προηγούμενων ταξινομήσεων, όπου τα στοιχεία ήταν παρόμοια με τα δικά μας. Με αυτόν τον τρόπο παράγονται συνεχείς και επικαλυπτόμενες γειτονιές. Το ποσοστό σφάλματος είναι ασύμπτωτο και είναι ανεξάρτητο από το μέτρο απόστασης. Αν κάθε ένα από τα αντικείμενα αυτά είναι προσκολλημένα σε μια κλάση, τότε ο καθορισμός της κλάσης στην οποία θα ανατεθεί ένα μη ταξινομημένο αντικείμενο, γίνεται μέσα από την παρατήρηση των κλάσεων στις οποίες είναι αντιστοιχισμένα τα κοντινότερα σε αυτό αντικείμενα. Ο αλγόριθμος kNN βρίσκει τα k κοντινότερα αντικείμενα, του υπό ταξινόμηση αντικειμένου, και το ταξινομεί στην πιο συνηθισμένη κλάση των k αυτών αντικειμένων.

Support Vector Machines (SVMs)

Είναι μία μέθοδος που ελαχιστοποιεί τον εμπειρικό κίνδυνο και ταυτόχρονα στοχεύει στην ελαχιστοποίηση του ανώτερου ορίου του σφάλματος γενίκευσης. Η βασική SVM παίρνει ένα σύνολο δεδομένων εισόδου και προβλέπει, για κάθε δεδομένη είσοδο, η οποία από τις δύο δυνατές τάξεις αποτελεί την έξοδο, καθιστώντας το ένα μη πιθανολογική δυαδικό γραμμικό ταξινομητή. Ένα μοντέλο SVM είναι μια αναπαράσταση των παραδειγμάτων ως σημεία στο χώρο, χαρτογραφούνται έτσι ότι τα παραδείγματα των ξεχωριστών κατηγορίες χωρίζονται από μια σαφή κενό που είναι τόσο ευρύ όσο το δυνατόν.

Η κάθε παρατήρηση αποτελείται από ένα ζευγάρι τη μορφής yi ανήκει Rn της συσχετιζόμενης κατηγορίας yi. Μια μηχανή διανυσμάτων υποστήριξης παράγει ένα ταξινομητή, το επονομαζόμενο βέλτιστο υπέρ- επίπεδο διαχωρισμού, μέσα από την μη γραμμική απεικόνιση των εισερχόμενων ανωτέρω διανυσμάτων στον πολύ-διάστατο χώρο των χαρακτηριστικών που περιγράφουν τα διανύσματα αυτά. Η βασική έννοια γύρω από την οποία δομείται μια μηχανή διανυσμάτων υποστήριξης είναι αυτή του περιθωρίου, σε κάθε μια

από τις πλευρές ενός καθορισμένου υπέρ-επιπέδου που χωρίζει τα δεδομένα ενός συνόλου εκπαίδευσης. Συγκεκριμένα μια μηχανή διανυσμάτων υποστήριξης κατασκευάζει ένα γραμμικό μοντέλο για την εκτίμηση του συνόλου των παραμέτρων a της συνάρτησης απόφασης f (x, a) , έτσι ώστε η τελευταία να πραγματοποιήσει την αντιστοίχιση i j x → y (η f (x, a) ονομάζεται μηχανή εκπαίδευσης).

Αν τα δεδομένα εκπαίδευσης είναι γραμμικά διαχωρισμένα, τότε η μηχανή διανυσμάτων υποστήριξης εκπαιδεύει μηχανές για την εκτίμηση ενός βέλτιστου υπέρ-επιπέδου που διαχωρίζει τα δεδομένα, στην μεγαλύτερη δυνατή απόσταση ανάμεσα σε αυτό και τα κοντινότερα αντικείμενα εκπαίδευσης. Τα αντικείμενα εκπαίδευσης που είναι πιο κοντά στο βέλτιστο υπέρ-επίπεδο διαχωρισμού ονομάζονται διανύσματα υποστήριξης, η δε λύση αναπαρίσταται σαν ένας γραμμικός συνδυασμός των παραπάνω αντικειμένων. Σε πιο γενικές περιπτώσεις που τα δεδομένα δεν είναι γραμμικά διαχωρισμένα, η μηχανή διανυσμάτων υποστήριξης χρησιμοποιεί μη γραμμικές μηχανές για την εύρεση ενός υπέρ-επίπεδο που ελαχιστοποιεί το πλήθος των λαθών για το σύνολο εκπαίδευσης.

1.11.2 Συσταδοποίηση

Η συσταδοποίηση είναι μια από τις σημαντικές διεργασίες στην εξόρυξη δεδομένων και αφορά την ανακάλυψη συστάδων καθώς και τον προσδιορισμό κατανομών ή προτύπων. Πιο αναλυτικά η συσταδοποίηση είναι η ανακάλυψη ομάδων ή αλλιώς οι ονομαζόμενες συστάδες αντικειμένων όπου τα αντικείμενα πρέπει να είναι όμοια στην κάθε μία συστάδα είναι όπως αναφέρεται η «*μη εποπτευμένη μάθηση*» στην αναγνώριση προτύπων, αριθμητική ταξινόμηση (numerical taxonomy), στην οικολογία και στην τμηματοποίηση (partition).

Τα βήματα τα οποία αποτελείται η διαδικασία της συσταδοποίησης είναι:

Βήμα 1: Επιλογή χαρακτηριστικών γνωρισμάτων

Στο βήμα αυτό πραγματοποιείται η επιλογή των κατάλληλων γνωρισμάτων με στόχο να κωδικοποιηθεί όσο το δυνατόν περισσότερη πληροφορία σε σχέση με την εργασία που μας ενδιαφέρει.

Βήμα 2: Επιλογή αλγορίθμου συσταδοποίησης

Σε αυτό το σημείο γίνεται η κατάλληλη επιλογή του αλγορίθμου μέσα σε έναν ποικίλο αριθμό αλγορίθμων ανάλογα με τα δεδομένα μας. Μετά την επιλογή του αλγορίθμου, ο αλγόριθμος καθορίζεται και χαρακτηρίζεται από το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης.

- Μέτρο γειτνίασης: Προσδιορίζεται πόσο όμοια είναι τα δύο αντικείμενα.

- Κριτήριο συσταδοποίησης: Αναφερόμαστε στην συνάρτηση κόστους ή κάποιου άλλου τύπου που χρησιμοποιούνται για να εκφραστεί το κριτήριο συσταδοποίησης.

Βήμα 3: Έλεγχος και ερμηνεία αποτελεσμάτων

Στο τελικό μας βήμα γίνεται η αξιολόγηση και η ερμηνεία των αποτελεσμάτων. Με τα κατάλληλα κριτήρια και τις τεχνικές εξακριβώνεται η ακρίβεια των αποτελεσμάτων.

Εφαρμογές συσταδοποίησης

Η συσταδοποίηση μπορεί να εφαρμοστεί σε διάφορες περιπτώσεις όπως:

Ο επιχειρηματικός κλάδος: Στον συγκεκριμένο κλάδο η τεχνική της συσταδοποίησης μπορεί να χρησιμοποιηθεί για την δημιουργία πελατειακών ομάδων με βάση τα αγοραστικά πρότυπα. Με αυτόν τον τρόπο μπορούν να αναπτυχθούν οι κατάλληλες στρατηγικές για την καλύτερη εξυπηρέτηση των πελατών μας.

Χωρική ανάλυση στοιχείων: Τα χωρικά δεδομένα θεωρούνται οι δορυφορικές εικόνες, ιατρικό εξοπλισμό και γεωγραφικά συστήματα πληροφοριών. Η συσταδοποίηση είναι χρήσιμη στην κατανόηση και την ανάλυση αυτών των δεδομένων.

Παγκόσμιος ιστός: Γίνεται συσταδοποίηση των δεδομένων ώστε να βρεθούν συστάδες χρηστών που επισκέπτονται κάποιο ιστοσελίδα.

Για παράδειγμα έχουμε μια βάση δεδομένων ενός καταστήματος με ρούχα και περιλαμβάνει εγγραφές ρούχων που προτιμώνται περισσότερο. Μέσω της συσταδοποίησης δημιουργούμε συστάδες πελατών που έχουν κοινές αγοραστικές προτιμήσεις.

Για να μπορέσουμε να ανακαλύψουμε τις αγοραστικές προτιμήσεις των πελατών μας χρησιμοποιούνται κάποιοι αλγόριθμοι όπου ταξινομούνται με βάση:

- ✓ τον τύπο δεδομένων που θα εισαχθούν
- ✓ την μέθοδο που καθορίζει τη συσταδοποίηση του συνόλου των διδόμενων
- ✓ τη θεωρία και τις θεμελιώδεις έννοιες στις οποίες είναι βασισμένες οι τεχνικές ανάλυσης συστάδας

Είδη συσταδοποίησης

❖ **Διαιρετική συσταδοποίηση:**

Χρησιμοποιείται στην αποσύνθεση συνόλων δεδομένων όπου θα εξετάσουμε σε ένα σύνολο ομάδων που δεν θα υπάρχει συσχέτιση μεταξύ τους. Οι αλγόριθμοι αυτοί στοχεύουν στην ελαχιστοποίηση των ανόμοιων δειγμάτων μέσα σε μια συστάδα και στην μεγιστοποίηση την ανομοιότητας μεταξύ διαφορετικών συστάδων.

Βασικοί αλγόριθμοι είναι:

Ο K-Means ακολουθεί μια επαναληπτική διαδικασία με σκοπό τον διαχωρισμό δεδομένων σε k συστάδες. Η διαδικασία αποτελείται από δύο βήματα:

- Πρώτα ορίζονται τα k κέντρα των c συστάδων και ανατίθενται κάθε στοιχείο του συνόλου των δεδομένων στη συστάδα της οποίας τα κέντρα είναι πιο κοντά και ξανά υπολογίζονται τα κέντρα. Αν η σειρά των δεδομένων δεν έχει κάποια ιδιαίτερη σημασία τότε παίρνουμε k-εγγραφές προς εξέταση.

- Στο δεύτερο βήμα υπολογίζονται η απόσταση κάθε στοιχείου του συνόλου δεδομένων από το κέντρο της κάθε ομάδας.

Παραλλαγές του αλγορίθμου k-means:

Υπάρχουν διάφορες παραλλαγές του αλγορίθμου όπου διαφέρουν στον τρόπο επιλογής των αρχικών k κέντρων, στον υπολογισμό της ομοιότητας και στη στρατηγική που χρησιμοποιούν για τον υπολογισμό των κέντρων των συστάδων.

- Αλγόριθμος ISO DATA: Αναζητά με βάση κάποιο κόστος εκτέλεσης για την αναζήτηση των καλύτερων αριθμών συστάδων.
- Αλγόριθμος Fuzzy C-Means: Με βάση την θεωρία ασαφούς λογικής και επεκτείνει τον αλγόριθμο K-Means.
- Αλγόριθμος SAS PROC FASTCLUS: Γίνετε έλεγχος ελάχιστων στοιχείων σε μια συστάδα και καθορίζετε η απόσταση κάθε στοιχείου μιας συστάδας από το κέντρο της συστάδας.

❖ **Η ασαφής συσταδοποίηση:** Έχει σαν θεωρία πως ένα δεδομένο μπορεί να χρησιμοποιηθεί σε παραπάνω από μία συστάδα.

❖ **Η μη ασαφής συσταδοποίηση:** Είναι ακριβώς αντίθετο με την ασαφής συσταδοποίηση ένα δεδομένο είτε ανήκει σε μια συστάδα είτε όχι.

❖ **Η συσταδοποίηση βασισμένη στα δίκτυα Kohonen:** Βασίζεται στη θεωρία νευρωτικών δικτύων.

❖ **Ιεραρχική συσταδοποίηση:** Διασπώνται μεγάλες συστάδες σε μικρότερες ή οι μικρές συστάδες ομαδοποιούνται σε μία.

Οι αλγόριθμοι διαιρούνται σε συσσωρευτικούς και διαιρετικούς αλγορίθμους. Οι συσσωρευτικοί ιεραρχικοί αλγόριθμοι δημιουργούν μια ακολουθία σχημάτων συσταδοποίησης μειώνοντας τον αριθμό των συστάδων. Οι διαιρετικοί ιεραρχικοί αλγόριθμοι σε αντίθεση με τους διαιρετικούς δημιουργούν μια ακολουθία σχημάτων συσταδοποίησης αυξάνοντας τον αριθμό συστάδων.

Βασικοί ιεραρχικοί αλγόριθμοι:

Cure: Είναι ιεραρχικός αλγόριθμος ο οποίος εφαρμόζεται για τη συσταδοποίηση μεγάλων βάσεων δεδομένων συνδυάζοντας τεχνικές δειγματοποίησης και τμηματοποίησης.

Birch: Χρησιμοποιείται η λεγόμενη ιεραρχική δεδομένων το CF-tree. Τμηματοποιεί τα στοιχεία με τρόπο αυξητικό και δυναμικό τρόπο.

- **Συσταδοποίηση βασισμένη στην πυκνότητα:** Οργανώνονται τα γειτονικά αντικείμενα ενός συνόλου δεδομένου σε συστάδες με βάση κάποια κριτήρια πυκνότητας.

DBSCAN: Ο συγκεκριμένος αλγόριθμος συσταδοποίησης βασίζετε στην πυκνότητα. Ουσιαστικά η περιοχή γύρω από το αντικείμενο της κάθε συστάδας θα περιέχει τον ελάχιστο αριθμό από αντικείμενα.

DEN CLUE: Ένας δεύτερος αλγόριθμος συσταδοποίησης που βασίζετε στην πυκνότητα και έχει σαν σκοπό την μοντελοποίηση της συνολικής πυκνότητας των σημείων με αναλυτικό τρόπο.

- **Συσταδοποίηση βασισμένη σε πλέγμα:** αναλύονται τα χωρικά δεδομένα.

STING: Ο αλγόριθμος αυτός επιτρέπει τη χρήση πληροφοριών συσταδοποίησης στην αναζήτηση των ερωτήσεων.

Wave Cluster: Βασίζεται σε τεχνικές επεξεργασίας σημάτων οι οποίες μετατρέπουν τα χωρικά δεδομένα στο πεδίο συχνοτήτων.

- **Συσταδοποίηση υποχωρών:** Μελετάει τα προβλήματα που προκύπτουν από τις υψηλές διαστάσεις των δεδομένων.

CLIQUE: Ο αλγόριθμος εξετάζει διαφορετικά υποχώρα για διαφορετικές συστάδες και ανακαλύπτει τις πυκνές περιοχές σε κάθε υπόχωρο. Για να προσεγγίσει την πυκνότητα των σημείων, το διάστημα εισόδου χωρίζεται στα κελιά.

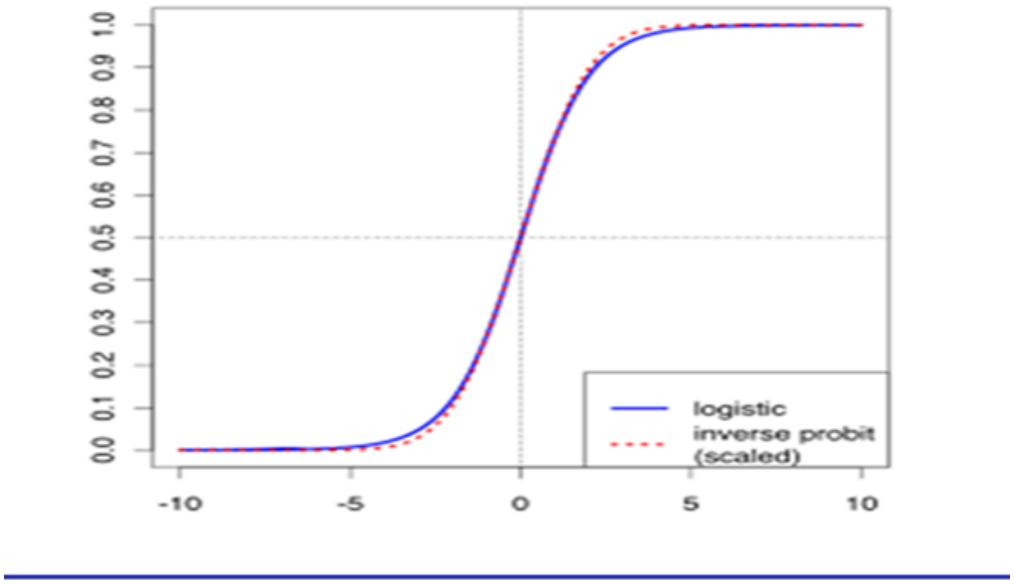
PROCLUS: Αναζητά υποσύνολα διαστάσεων ώστε τα σημεία δεδομένων να είναι πυκνά συσταδοποιημένα στους αντίστοιχους υποχώρους.

1.12 Παλινδρόμηση

Η παλινδρόμηση συμπεριφέρεται σαν συνάρτηση πρόβλεψης στην όποια καταχωρούνται τα προς εξέταση δεδομένα και πραγματοποιείται πρόβλεψη πραγματικού αριθμού. Ακόμη μπορούμε να προβλέπουμε τα κέρδη μιας επιχείρησης, τις πωλήσεις, τις τιμές των ακινήτων, τη θερμοκρασία, το τετραγωνικό μήκος σε πόδια και διάφορα άλλα πράγματα. Η διαδικασία εύρεσης ενός προτύπου αρχίζει με ένα σύνολο δεδομένων μέσα στο οποίο οι μεταβλητές στόχοι είναι ήδη γνωστές και συνιστούν τα ιστορικά δεδομένα όπου τα μισά χρησιμοποιούνται για την πρόβλεψη και τα υπόλοιπα για τη δοκιμή του προτύπου.

Το λογιστικό μοντέλο είναι ένα μη γραμμικό μοντέλο τα σφάλματα του οποίου δεν υπακούν στην κανονική κατανομή και η μεταβλητή απόκρισης είναι διακριτή. Η λογιστική παλινδρόμηση χρησιμοποιείται όταν επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού ή ενός συμβάντος. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή (Y) είναι δίτιμη (δηλαδή παίρνει την τιμή 0 όταν απουσιάζει το χαρακτηριστικό ή την τιμή 1 όταν υπάρχει το χαρακτηριστικό).

Ένα παράδειγμα της εφαρμογής της είναι η εξέταση της εμφάνισης στεφανιαίας νόσου σε ένα νοσοκομείο σε ένα δείγμα ανδρών σε σχέση με την ηλικία, εάν καπνίζουν, τη συστολική και διαστολική πίεση του αίματος, τα επίπεδα της χοληστερόλης και το βάρος τους. Σε αυτή την περίπτωση, κωδικοποίησαν με 0 τα άτομα που δεν έχουν πάθει έμφραγμα τα τελευταία 10 χρόνια και με 1 τα άτομα που έχουν υποστεί έμφραγμα, όπως φαίνεται στην Εικόνα 3.



«Εικόνα 3»

Το γραμμικό μοντέλο είναι αδύνατο να χρησιμοποιηθεί όταν η μεταβλητή Y είναι δυαδική και έχουμε τα εξής τρία προβλήματα:

1. Τα σφάλματα δεν είναι κανονικά.
2. Τα σφάλματα έχουν άνισες διασπορές
3. Περιορισμός στη συνάρτηση απόκρισης

Παρόλο που στα δύο πρώτα προβλήματα είναι δυνατό να τα παραλείψουμε και να χρησιμοποιήσουμε την γραμμική παλινδρόμηση, εφαρμόζοντας κάποιες άλλες τεχνικές, το τρίτο πρόβλημα μας το απαγορεύει ρητά, γιατί δεν μπορεί να αντιμετωπιστεί με διαφορετικό τρόπο.

1.13 Περίληψη

Ένας νέος τομέας που μπαίνει στην ζωή του ανθρώπου. Οι όγκοι των δεδομένων μέρα με την μέρα αυξάνονται με ταχύτερους ρυθμούς και στην εποχή αυτή τα δεδομένα έχουν γίνει και ψηφιακά θα πρέπει να δημιουργηθεί μια νέα γενιά από υπολογιστές που θα βοηθήσει στην εξόρυξη γνώσης ταχύτερα. Λόγω των τεράστιων βάσεων δεδομένων λοιπόν χρησιμοποιούνται διαφορετικά κριτήρια κατηγοριοποίησης των μεθόδων και των συστημάτων εξόρυξης δεδομένων ανάλογα με τις τεχνικές που θα χρησιμοποιηθούν, το είδος των βάσεων δεδομένων και το είδος της γνώσης που θα εξαχθεί. Μια αποθήκη δεδομένων από την οποία ο άνθρωπος ψάχνει να ανακαλύψει πρότυπα και πληροφορίες που παρουσιάζονται με διαφορετικές μορφές ώστε ταχύτατα να μπορεί να αντλεί πληροφορίες που θα είναι χρήσιμες για αυτόν.

Με την ανάπτυξης της τεχνολογίας κρίθηκε αναγκαίο η δημιουργία και χρήση βασικών εργαλείων για να μπορεί να γίνει συλλογή και αποθήκευση με βασικό στόχο την εξόρυξη γνώσης έτσι ώστε να εξάγονται χρήσιμα συμπεράσματα. Η ποιότητα της εξαγόμενης γνώσης είναι ένα ακόμα σημαντικό θέμα. Τα πρότυπα που παράγονται από την εξόρυξη θα πρέπει να είναι έγκυρα έτσι ώστε να μπορούν να φανούν χρήσιμα στην αξιολόγηση της ποιότητας των δεδομένων. Ωστόσο αυτή η διαδικασία εξαγωγής προτύπων γνώσης περικλείει και ένα ποσοστό αβεβαιότητας. Η διαχείριση της αβεβαιότητας προσπαθεί να μετριαστεί με χρήση κάποιων μέσων κατηγοριοποίησης.

Στην συνέχεια παρουσιάσαμε τον ορισμό της διαδικασίας ανακάλυψης γνώσης (KDD) δηλαδή την ανακάλυψη γνώσης και προτύπων με την χρήση αλγορίθμων καθώς και τον συσχετισμό της με την εξόρυξη γνώσης. Η ανακάλυψη γνώσης επιτυγχάνεται με την ανάλυση, ερμηνεία και αξιολόγηση των δεδομένων. Είναι ένα σύστημα με το οποίο η εξορυγμένη γνώση αναπαριστάτε. Αρχικά αναλύθηκαν τα βήματα της διαδικασίας καθώς και οι μέθοδοι που χρησιμοποιούνται για να γίνει με επιτυχία η εξόρυξη χρήσιμης γνώσης.

Οι μέθοδοι είναι η κατηγοριοποίηση και η συστηματοποίηση. Η κατηγοριοποίηση είναι μια σημαντική διαδικασία της εξόρυξης δεδομένων που έχει σαν σκοπό τη δημιουργία του μοντέλου που χρησιμοποιείται να κατηγοριοποιηθούν τα δεδομένα των οποίων η κατηγοριοποίηση δεν είναι γνωστή. Στην κατηγοριοποίηση ανήκουν οι Bayesian κατηγοριοποίηση, τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα. Η συσταδοποίηση είναι μια από τις σημαντικές διεργασίες στην εξόρυξη δεδομένων και αφορά την ανακάλυψη

συστάδων καθώς και τον προσδιορισμό κατανομών ή προτύπων. Οι Εφαρμογές της συσταδοποίησης είναι ο επιχειρηματικός κλάδος, η χωρική ανάλυση στοιχείων και ο παγκόσμιος ιστός. Επίσης τα είδη της συσταδοποίησης είναι η διαιρετική συσταδοποίηση, η ασαφής συσταδοποίηση, η μη ασαφής συσταδοποίηση, η συσταδοποίηση βασισμένη στα δίκτυα Kohonen, η Ιεραρχική συσταδοποίηση η συσταδοποίηση βασισμένη στην πυκνότητα και η συσταδοποίηση βασισμένη σε πλέγμα.

Κεφάλαιο 2

2.1 Κανόνες Συσχέτισης

Μια νέα - σύγχρονη μέθοδος για την εξαγωγή γνώσης της βάσης δεδομένων είναι οι κανόνες συσχέτισης. Δημιουργήθηκε στις αρχές του ′90 για τις ανάγκες των υπεραγορών ώστε να καταχωρηθούν οι συναλλαγές του κάθε πελάτη ηλεκτρονικά αναλύοντας το καλάθι αγοράς. Γι αυτόν άλλωστε το λόγο έχει δημιουργηθεί μια πλειάδα αλγορίθμων που παράγουν κανόνες συσχέτισης και πρότυπα. Με τους κανόνες συσχέτισης εξετάζονται πολλά καλάθια αγοράς πελατών συνδυάζοντας τα αντικείμενα μεταξύ τους με σχέσεις εξάρτησης. Οι κανόνες εφαρμόζονται στην προώθηση προϊόντων, την τοποθέτηση προϊόντων στα ράφια καταστημάτων και την διαχείριση αποθεμάτων. Για αυτό πρέπει να γίνεται πάντα σωστή χρήση και καλός έλεγχος των δεδομένων πριν την εξόρυξη.

Οι κανόνες συσχέτισης έχουν δημιουργηθεί για την βελτίωση των αποτελεσμάτων και με του ς αλγόριθμους που μπορεί να ανακαλυφθούν βοηθούν για την δημιουργία πιο συμπαγών αλλά και χρήσιμων μοντέλων έτσι ώστε η διαδικασία να είναι πιο χρήσιμη και σε μελλοντικές χρήσεις.

Προώθηση προϊόντων: Ο τρόπος που προωθούν τα προϊόντα τα super market ή τις υπεραγορές δίνουν πληροφορίες πως θα επηρεάσει το ένα προϊόν με το άλλο. Για παράδειγμα ξηρούς καρπούς, ποτά. Τα ποτά βρίσκονται του κανόνα η επιχείρηση μπορεί να χρησιμοποιήσει να αυξήσει τις πωλήσεις τους. Οι ξηροί καρποί που βρίσκονται στο πρώτο μέρος δίνουν την πληροφορία τι θα συμβεί αν δεν πωλούνται από την επιχείρηση.

Τοποθέτηση προϊόντων στα ράφια καταστημάτων: Οι κανόνες συσχέτισης δείχνουν τις τάσεις αγορών των πελατών. Οι τάσεις αυτές χρησιμοποιούνται από τις υπεραγορές ώστε να τοποθετούνται τα προϊόντα στα ράφια με τρόπο να διευκολύνονται οι αγορές των πελατών ακόμα και να πραγματοποιηθεί η παρακίνηση των αγορών τους. Για παράδειγμα τα super market τοποθετούν τα πατατάκια δίπλα στα αναψυκτικά όπως και οι ξηροί καρποί με τα ποτά.

Διαχείριση αποθεμάτων: Η επιχείρηση πρέπει να παρακολουθεί τα αποθέματα της για να μην έχει έλλειψη σε κάποιο προϊόν όπου έχει ζήτηση και δεν ικανοποιεί τις ανάγκες του καταναλωτή.

2.2 Ορισμός προβλήματος

Ο ορισμός του προβλήματος της εξαγωγής κανόνων συσχέτισης περιγράφεται ως εξής. Έστω $I=\{i_1,i_2,i_3,\ldots,i_n\}$ ένα σύνολο από διακριτά κατηγορήματα που αποκαλούνται *items (αντικείμενα*). Έστω ακόμα *D* ένα σύνολο από *δοσοληψίες (transactions*), όπου κάθε δοσοληψία(transaction) *T* είναι ένα σύνολο από αντικείμενα, το οποίο καλείται itemset, και για το οποίο ισχύει $T \subseteq I$.Κάθε δοσοληψία χαρακτηρίζεται από ένα μοναδικό αναγνωριστικό που καλείται *TID*. Στο εξής θα λέμε ότι μια δοσοληψία *T* περιέχει το X , ένα σύνολο από κάποια αντικείμενα του *I*, εάν ισχύει $X \subseteq T$.

Ένας Κανόνας Συσχέτισης(Association Rule) είναι μια συσχέτιση της μορφής $X \rightarrow Y$,όπου $X \subseteq I$, $Y \subseteq I$ και $X \cap Y = \varnothing$. Το πρώτο μέλος του κανόνα ονομάζεται υπόθεση και το δεύτερο συμπέρασμα. Ωστόσο ο κανόνας $X \rightarrow Y$ ισχύει στο D που είναι το σύνολο των δοσοληψιών με εμπιστοσύνη (confidence) c, εάν το c% των δοσοληψιών στο D που περιέχουν το X περιέχουν και το Y.Η υποστήριξη (support) σημαίνει ότι ο κανόνας $X \rightarrow Y$ έχει υποστήριξη s, εάν το s% των δοσοληψιών στο D περιέχουν το $X \cup Y$. Αυτές ήταν οι δυο βασικές μετρικές στον συσχετισμό κανόνων.

Εν κατακλείδι αυτός ο ορισμός μπορεί να γενικευθεί και για ένα σύνολο από items (itemset). Έτσι λοιπόν θα μπορούσαμε να πούμε ότι ένα item X έχει υποστήριξη s δηλαδή sup(X)=s,εάν το s% των δοσοληψιών στο D περιέχουν το X. Έστω τώρα ότι ένα itemset X έχει μήκος k θα το εκφράζουμε ως k-itemset όταν εκείνο αποτελείται από k πλήθος items, συνοψίζοντας |X|=k.

Από τα παραπάνω προκύπτει ότι ο κανόνας $X \rightarrow Y$ έχει υποστήριξη s ,όταν $sup(X \Rightarrow Y)=sup(X \cup Y)$και εμπιστοσύνη c,όταν $conf(X \Rightarrow Y)=sup(X \cup Y)/sup(X)$.

Όσα αναφέρθηκαν νωρίτερα θα γίνουν καλύτερα αντιληπτά στο παράδειγμα που παρατίθεται πιο κάτω. Θα εξετάσουμε έναν κανόνα συσχέτισης ανάμεσα στα items: Bread,Butter,Milk. Θα έχουμε λοιπόν $\{ \text{Bread, Butter}\} \rightarrow$ Milk. Παρατηρούμαι στο πινάκα

2.1 ότι τα itemset { Bread, Butter, Milk} εμφανίζονται σε 3 δοσοληψίες, ενώ το itemset{ Bread, Butter} σε 2. Άρα με βάση των ορισμών που διατυπώθηκαν πιο πάνω ο κανόνας συσχέτισης διαμορφώνεται ως εξής, { Bread, Butter} Milk ,έχει υποστήριξη s=(3/5)=0,6 και εμπιστοσύνη c=(2/3)=0.67.

ΠΙΝΑΚΑΣ

TID	items
1	Butter, Beer, Bread
2	Bread, Eggs, Soda,Butter
3	Coke, Butter, Bread, Milk
4	Bread, Milk, Butter
5	Butter, Bread, Milk,Coke

2.3 Πρόβλημα Εξαγωγής Κανόνων Συσχέτισης

Το πρόβλημα Εξαγωγής Κανόνων Συσχέτισης αφορά την αναζήτηση και εύρεση όλων εκείνων των κανόνων συσχέτισης που θα πρέπει να ικανοποιούν κάποια κατώτατα όρια σχετικά με την υποστήριξη (support) και την εμπιστοσύνη(confidence). Ωστόσο, η υποστήριξη (support) ενός κανόνα θα πρέπει να είναι μεγαλύτερη από μια τιμή που ορίζουμε ως όριο και την ονομάζουμε ελάχιστη υποστήριξη (minsup),ακόμα και η εμπιστοσύνη πρέπει να είναι μεγαλύτερη από το όριο που ονομάζεται ελάχιστη εμπιστοσύνη (minconf). Ο αριθμός των κανόνων που θα προκύψουν καθορίζεται από αυτούς τους δυο παράγοντες και πρέπει να επιλέγονται με τον τύπο του πίνακα δοσοληψιών. Τα σύνολα στοιχείων (itemset) που έχουν μεγαλύτερη υποστήριξη από τη τιμή που έχουμε ορίσει ως όριο minsup ονομάζονται *συχνά(frequent) ή μεγάλα (large).*

Τα στάδια για την Εξαγωγή Κανόνων Συσχέτισης είναι τα ακόλουθα:

- Το πρώτο βήμα πραγματοποιεί η εύρεση όλων των συχνών (frequent) συνόλων στοιχείων (itemset)
- Ενώ στο δεύτερο βήμα εκτελείται η εξαγωγή των κανόνων που έχουν μεγαλύτερη τιμή από την ελάχιστη εμπιστοσύνη (minconf).

Κάνοντας εφαρμογή στα παραπάνω στάδια προκύπτει το αν ο κανόνας είναι χρήσιμος ή όχι. Ο αλγόριθμος που εφαρμόζεται για το δεύτερο στάδιο είναι ο εξής: για κάθε ένα από τα συχνά (frequent) σύνολα στοιχείων (itemset) l,βρες όλα τα μη κενά υποσύνολα του. Για κάθε τέτοιο υποσύνολο α, παρουσίασε το κανόνα $\alpha \rightarrow (l-\alpha)$, εάν ο λόγος $sup(l)/sup(\alpha)$, που αντιστοιχεί στην εμπιστοσύνη (confidence) του κανόνα είναι τουλάχιστον όσο η τιμή που ορίσαμε σαν όριο,minconf. Ωστόσο οι ερευνητές αντιμετωπίζουν αρκετές δυσκολίες στην υλοποίηση του πρώτου βήματος όπου υπάρχει πλειάδα αλγορίθμων για την επίλυση του με σημαντικότερο τον αλγόριθμο Apriori όπου θα αναλυθεί στη συνέχεια.

Οι κανόνες συσχέτισης μπορούν να χωριστούν σε αντιπροσωπευτικούς κανόνες και σε ποσοτικούς. Ο λόγος που γίνεται αυτό είναι ότι πολλές φορές ο αριθμός κανόνων που μπορεί να παραχθούν είναι πολύ μεγάλος γι'αυτό χρησιμοποιούνται κάποια κριτήρια σημαντικότητας. Από μια άλλη οπτική γωνία αυτό το πρόβλημα μπορεί να αντιμετωπιστεί με τη παραγωγή ενός ελάχιστου αριθμού συνόλου κανόνων που ονομάζονται Αντιπροσωπευτικοί Κανόνες Συσχέτισης (Representative Association Rules).

2.4 Αλγόριθμος Apriori ή περιγραφικά αρχή της προς τα κάτω κλειστότητας.

Ο αλγόριθμος Apriori είναι ένας κλασσικός αλγόριθμος που μας βοηθάει στην παραγωγή των κανόνων συσχέτισης και που χρησιμοποιείται στον τομέα της εξόρυξης δεδομένων για την εκμάθηση κανόνων συσχέτισης. Υπάρχουν πολλοί αλγόριθμοι με τους οποίους μπορούμε να επεξεργαστούμε τα μεγάλα σύνολα δεδομένων και να παράγουμε κανόνες συσχέτισης, όπως είναι οι αλγόριθμοι Partition, PredictiveApriori, Winepi και Minepi που έχουν σχεδιαστεί για την εύρεση κανόνων συσχέτισης σε δεδομένα που δεν έχουν συναλλαγές, δύο

νέοι και πολύ διαδεδομένοι αλγόριθμοι είναι ο Apriori και AprioriTid αλγόριθμοι που μας βοηθάει στην εξαγωγή των κανόνων συσχέτισης και οι οποίος εμφανίστηκαν το 1994 από τους Agrawal και Srikant. Ο αλγόριθμος Apriori έχει πάρει το όνομα του από την προγενέστερη γνώση (prior knowledge) των χαρακτηριστικών των συχνών συνόλων αντικειμένων, που χρησιμοποιεί. Ο Apriori γενικά υιοθετεί την τεχνική αναζήτηση, level-wise, η οποία είναι μια επαναλαμβανόμενη τεχνική που χρησιμοποιεί τα k-itemsets για να κτίσει τα (k+1)-itemsets.

Η διαδικασία εξαγωγής γνώσεων από μεγάλα σύνολα δεδομένων που συλλέγονται στις επιχειρήσεις κρύβουν ένα μεγάλο πλούτο πληροφοριών και η διαδικασία αυτή επαναλαμβάνεται αρκετές φορές και πολλοί είναι οι αλγόριθμοι που δεν μπορούν να ανταπεξέλθουν με μεγάλη ταχύτητα πράγμα που αποτελεί μεγάλο μειονέκτημα και πρόβλημα. Εδώ έρχονται οι αλγόριθμοι Apriori και AprioriTid με τους οποίους καταφέρνουμε να καταπολεμήσουμε αυτά τα προβλήματα και εύκολα αλλά και γρήγορα να περάσουμε πολλές φορές τις βάσεις δεδομένων.

Η ιδιότητα Apriori αναφέρει ότι*: όλα τα μη κενά υποσύνολα των συχνών συνόλων αντικειμένων πρέπει να είναι επίσης συχνά.*

Η διαδικασία έχει ως εξής: Αρχικά ψάχνουμε να βρούμε όλα τα frequent itemsets, και μετράμε τα στοιχεία που έχουν υποστήριξη μεγαλύτερη από ή ίση με ελάχιστη υποστήριξη Επειδή πραγματοποιούνται πολλά περάσματα της διαδικασίας αυτής, για κάθε κ πέρασμα τα μεγάλα στοιχειοσύνολα στο προηγούμενο πέρασμα ομαδοποιούνται σε σετ των k αντικειμένων και παίρνουν την υποψήφια μορφή. Η στήριξη για τα διάφορα υποψήφια στοιχειοσύνολα μετριέται και εάν η υποστήριξη βρεθεί να είναι μεγαλύτερη από την ελάχιστη στήριξη εξάγουμε τους κανόνες συσχέτισης. Στο δεύτερο μέρος αυτής της διαδικασίας είναι που χρησιμοποιούμε διάφορους αλγόριθμους με πιο διαδεδομένο τον Apriori με τον οποίον μετράμε κατά πόσο υποστηρίζονται τα στοιχειοσύνολα για να πάρουμε χρήσιμες πληροφορίες που θα μας βοηθήσουν στην λήψη αποφάσεων. Η διαδικασία σταματάει όταν το στοιχειοσύνολο είναι ένα κενό σύνολο.

Ο αλγόριθμος Apriori είναι ένας αλγόριθμος που βασίζεται στους κανόνες απόφασης, πρέπει να παράγει κανόνες οι οποίοι έχουν υψηλές ικανότητες πρόβλεψης και ταυτόχρονα υψηλή αξιοπιστία. Η προσέγγιση του Apriori είναι «κάτω προς τα πάνω», και τα αποτελέσματα είναι σύνολα κανόνων που μας λένε πόσο συχνά τα στοιχεία που περιέχονται στα σύνολα δεδομένων που αναλύσαμε, εμφανίζονται.

Με τον όρο Itemsets εννοούμε τα σύνολα-στοιχειοσύνολα που υποστηρίζουν μεγάλα αντικείμενα και εμφανίζονται σαν σετ αντικειμένων που εμφανίζονται μαζί από τις συναλλαγές όπως π.χ. σε μαγαζιά. frequent itemsets είναι τα σύνολα των στοιχείων με την ελάχιστη στήριξη, συμβολίζεται με Li, όπου i στοιχειοσύνολο. Έτσι χρησιμοποιούμε τα συχνά στοιχειοσύνολα-frequent itemsets, για τη δημιουργία κανόνων συσχέτισης.

Η ιδιότητα Apriori χρησιμοποιείται για την παραγωγή του Lk δύο βήματα:

Βήμα : 1 Διαδικασία Ένωσης (join):
Για να βρεθεί το σύνολο Lk , παράγετε ένα σύνολο από υποψήφια σύνολα με k χαρακτηριστικά (k-itemsets) από την ένωση του συνόλου Lk-1 με τον εαυτό του. Το σύνολο με τα υποψήφια σύνολα χαρακτηριστικών καλείται Ck. Εάν το li είναι μέλος του Lk-1, τότε το li[j] αναφέρεται στο χαρακτηριστικό j του συνόλου χαρακτηριστικών li. Ο Apriori θεωρεί ότι τα χαρακτηριστικά στα σύνολα είναι ταξινομημένα σε αλφαβητική σειρά. Για κάποιο σύνολο χαρακτηριστικών li με (k-1) χαρακτηριστικά, τα χαρακτηριστικά είναι ταξινομημένα σε li[1] < li[2] < li[3] <...< li[k-1]. Όταν η ένωση Lk-1 L_{k-1} εκτελείται, τα μέλη του Lk-1 μπορούν να ενωθούν εάν τα πρώτα (k-2) χαρακτηριστικά είναι τα ίδια.

Βήμα : 2 Διαδικασία Κλαδέματος (prune):
Κάποια από τα σύνολα χαρακτηριστικών που ανήκουν στο Ck, μπορεί να είναι συχνά εμφανιζόμενα κι άλλα όχι, όμως όλα τα συχνά εμφανιζόμενα σύνολα k χαρακτηριστικών (k-itemsets) συμπεριλαμβάνονται στο Ck. Θα πρέπει να γίνει μία αναζήτηση στη βάση δεδομένων για να μετρηθεί ο αριθμός όπου κάθε υποψήφιο σύνολο στο Ck, εμφανίζεται στη βάση δεδομένων. Όλα τα σύνολα αντικειμένων που περιλαμβάνονται στο Ck, και εμφανίζονται στη βάση δεδομένων όχι λιγότερό αριθμό από την ελάχιστη υποστήριξη, τότε αυτό το σύνολο χαρακτηριστικών προστίθεται στο Lk. Αυτό γίνεται, όπως αναφέρει η Apriori ιδιότητα, οποιοδήποτε (k-1)-itemset σύνολο χαρακτηριστικών δεν είναι συχνό τότε δεν μπορεί να είναι υποσύνολο κάποιου k-itemset σύνολο χαρακτηριστικών. Έτσι επειδή το σύνολο Ck, μπορεί να γίνει αρκετά μεγάλο, τα σύνολα χαρακτηριστικών που δεν είναι συχνά αφαιρούνται.

Η μορφή ψευδοκώδικα του Αλγόριθμου Apriori είναι η παρακάτω:

(1) $L_t = \{large\ 1 - itemsets\};$

(2) **For** $(k = 2; L_{k-1} \neq \varnothing; K++)$

(3) **do begin**

(4) $C_k = Apriori - gen(L_{k-1});$

(5) **forall** transactions $t \in D$

(6) **do begin**

(7) $C_t = subset(C_k, t);$

(8) **forall** candidates $c \in C_t$

(9) ***do***

(10) $c.count++;$

(11) **end**

(12) $L_k = \{c \in C_t \mid c.count \geq min\ sup\}$

(13) **End**

(14) **Return** $U_k L_k$

Γενικά ο αλγόριθμος Apriori χρησιμοποιώντας την επαναλαμβανόμενη τεχνική level-wise, ψάχνει να βρει τα πιο συχνά itemsets που είναι και κατάλληλα.

Σαν είσοδο:

- D, βάση δεδομένων με δοσοληψίες
- min-sup, ο ελάχιστος αριθμός υποστήριξης.

Και σαν έξοδο:

- L, σύνολο με όλα τα συχνά σύνολα χαρακτηριστικών που ανήκουν στο D

Ανάλυση των εντολών

Στην εντολή 1: Μετράμε πόσο συχνά εμφανίζεται κάθε item και αυτά με το μικρότερο minsup τα αφαιρούμε.

Στην εντολή 2: Έχουμε δύο βήματα, στο πρώτο παράγουμε τα υποψήφια k-itemsets Ck από τα συχνά (k-1)-itemsets Lk από το προηγούμενο πέρασμα. Με k εννοούμε το πέρασμα.

Στην εντολή 3: Έχουμε την δημιουργία των υποψηφίων χρησιμοποιώντας την συνάρτηση Apriori_gen. Γίνεται, δηλαδή η διαδικασία όπου το σύνολο υποψηφίων Ck παράγεται από την ένωση του Lk-1 με τον εαυτό του. Η διαδικασία apriori_gen παράγει τα υποψήφια σύνολα χαρακτηριστικών και μετά χρησιμοποιεί την ιδιότητα Apriori για να αφαιρέσει αυτά που δεν είναι συχνά.

Στην εντολή 4: Είναι το δεύτερο βήμα στο οποίο υπολογίζουμε το support count των υποψήφιων itemsets.

Στην εντολή 5: Βρίσκουμε τα υποψήφια που περιέχονται στο t και έχουμε αύξηση κατά 1 μονάδα, εντολή 7.

Στη εντολή 9: Οπού είναι και το τέλος της διαδικασίας του περάσματος, υπολογίζουμε το Lk και τα itemsets του Ck που εμφανίζονται λιγότερο συχνά τα απορρίπτουμε.

Στην εντολή 11: Όλα τα συχνά σύνολα χαρακτηριστικά itemsets που έχουμε βρει από τον αλγόριθμο ενώνονται στο L. Έτσι μετά τη διαδικασία για εξαγωγή κανόνων συσχέτισης ο αλγόριθμος μπορεί να χρησιμοποιήσει το σύνολο L.

Κανόνες συσχέτισης και στοιχειοσύνολα

Ένας κανόνας συσχέτισης έχει την μορφή A -> B , όπου A και B είναι στοιχειοσύνολα, τα οποία συνδέονται με την εμπιστοσύνη μέτρο που είναι ο λόγος του την υποστήριξη του στοιχειοσυνόλου A ⌣B για την υποστήριξη του στοιχειοσυνόλου A . Η εμπιστοσύνη αυτή μας δείχνει την πιθανότητα ότι, όταν αγοράζεται το ένα θα πρέπει και το B επίσης να αγοραστεί.

Για την εύρεση συχνών στοιχειοσυνόλων χρησιμοποιούμε την στρατηγική Apriori:

Η αρχή του αλγόριθμου Apriori έχει ως εξής: «Αν ένα στοιχειοσύνολο είναι συχνό, τότε όλα τα υποσύνολα του είναι συχνά ή ισοδύναμα αν ένα στοιχειοσύνολο είναι μη συχνό, όλα τα υπερσύνολα του είναι μη συχνά »

Έστω k = 1, k: μήκος στοιχειοσυνόλου

1. Παράγουμε υποψήφια (k+1)-στοιχειοσύνολα. Για να αποφύγουμε τη δημιουργία του ίδιου στοιχειοσυνόλου, κρατάμε κάθε στοιχειοσύνολο (λεξικογραφικά) ταξινομημένο

2. Στο επόμενο βήμα ψαλιδίζουμε τα υποψήφια στοιχειοσύνολα που περιέχουν μη συχνά στοιχειοσύνολα μεγέθους k. Κάνουμε απλούς ελέγχους έτσι ώστε να ελέγχουμε αν τα υποσύνολα τους είναι συχνά και έτσι να αποφύγουμε να υπολογίσουμε την υποστήριξή του κάθε υποψήφιου (k+1)- στοιχειοσυνόλου διασχίζοντας τη βάση των δοσοληψιών.

3. Τα υποψήφια στοιχειοσύνολα που δεν είναι συχνά, απλά τα σβήνουμε.

Ένα στοιχειοσύνολο λέμε ότι είναι κλειστό (closed) αν κανένα από τα άμεσα υπερσύνολα του δεν έχει την ίδια υποστήριξη με αυτό, δηλαδή έχει μικρότερη υποστήριξη.

Ενώ ένα στοιχειοσύνολο είναι κλειστό συχνό στοιχειοσύνολο αν είναι κλειστό και συχνό, δηλαδή η υποστήριξη του είναι μεγαλύτερη ή ίση με minsup.

Ο Apriori αποδίδει καλύτερα από ό,τι AprioriTid στα αρχικά περάσματα, αλλά στην συνέχεια την σκυτάλη την παίρνει ο αλγόριθμος AprioriTid ο οποίος έχει καλύτερη απόδοση από ό, τι Apriori.

Ο αλγόριθμος Apriori χρησιμοποιεί «Οριζόντια Διάρθρωση Δεδομένων» για την αναπαράσταση της Βάσης Δεδομένων των Δοσοληψιών.

Ενώ χρησιμοποιούμε «Κάθετη Διάρθρωση Δεδομένων» για βρούμε κάθε στοιχείο σε ποιες δοσοληψίες εμφανίζεται. Η υποστήριξη υπολογίζεται παίρνοντας τις τομές των TID-λιστών.

Παράδειγμα εργασίας για Apriori:

Βάση Δεδομένων: TID ΣΗΜΕΙΑ 100 XZ 200 MXZ 300 MXYZ 400 MN Υποστήριξη = 50% του συνολικού αριθμού των συναλλαγών = 50% 4 = 2. Επανάληψη-1: C1 (BAR) τρεις φορές την ημέρα ΣΕΤ στοιχειοσυνόλων 100 {X}, {Z} 200 {M}, {X}, {Z} 300 { M}, {X}, {Y}, {Z} 400 {M}, {N} L1 στοιχειοσύνολο ΥΠΟΣΤΗΡΙΞΗ {X} 3 {M} 3 {Z} 3 Όπως φαίνεται στους

παραπάνω πίνακες, το πρώτο βήμα είναι να δημιουργήσει Γ1 (BAR) το οποίο περιέχει όλες τις 1-στοιχειοσυνόλων από τη βάση δεδομένων και διατηρούνται με την Tid της παραγωγής των συναλλαγών. Το επόμενο βήμα είναι να δημιουργήσει όλες τις μεγάλες 1-στοιχειοσυνόλων από C1 (BAR). Έτσι, στο τέλος του επανάληψη 1 παίρνουμε {X}, {M} και {Z}, όπως οι μεγάλες 1-στοιχειοσύνολα. επανάληψη-2: C2 στοιχειοσύνολο ΥΠΟΣΤΗΡΙΞΗ {MX} 2 {} XZ 3 } {MZ 2 C2 η επανάληψη -2 το πρώτο βήμα είναι να μάθετε C2 που είναι το σύνολο των υποψηφίων μεγάλο 2-στοιχειοσύνολα. Για αυτό καλούμε apriori-gen λειτουργία που επεξεργάζεται L1, προκειμένου να πάρει C2. Στο C2 βρίσκουμε την υποστήριξη όλων των υποψηφίων μεγάλο 2-στοιχειοσύνολα. Στο C2 (BAR) έχουμε τώρα όλα αυτά τα μεγάλα 2-στοιχειοσύνολα που σχετίζονται με τη συναλλαγή τους Tid του. Θα παρατηρήσετε ότι Tid 400 δεν έχει καμία υποψήφιος μεγάλο 2-στοιχειοσύνολα. Από C2 (BAR) θα επεξεργάζονται L2 η οποία θα έχει μόνο μεγάλο 2-στοιχειοσύνολα. επανάληψη -3: C3 στοιχειοσύνολο ΥΠΟΣΤΗΡΙΞΗ {} MXZ 2 C3 (BAR) τρεις φορές την ημέρα ΣΗΜΕΙΑ 100 ---------- 200 {MXZ } 300 {} MXZ 400 ---------- L3 στοιχειοσύνολο ΥΠΟΣΤΗΡΙΞΗ {} XMZ 2 Στο -3 επανάληψη, και πάλι το πρώτο βήμα είναι να μάθετε C3 που είναι το σύνολο των υποψηφίων μεγάλο 3-στοιχειοσύνολα από το L2. Για αυτό καλούμε apriori-gen λειτουργία που επεξεργάζεται L2, προκειμένου να πάρει C3. Το C3 συναντάμε τη στήριξη όλων των υποψηφίων μεγάλο 3-στοιχειοσύνολα. Το C3 (BAR) έχουμε τώρα όλα αυτά τα μεγάλα 3-στοιχειοσύνολα που σχετίζονται με τη συναλλαγή τους Tid του. Παρατηρούμε ότι μόνο Tid 200 και 300 έχουν πήρε μεγάλη υποψήφιος 3-στοιχειοσύνολα. Από C3 (BAR) θα επεξεργάζονται L3 που θα έχουν μόνο μεγάλη 3-στοιχειοσύνολα. Έτσι, στο τέλος της επανάληψης-3 έχουμε μείνει μόνο με ένα μεγάλο στοιχειοσύνολο {} XMZ και γι 'αυτό δεν μπορεί να έχει καμία περαιτέρω υποψήφιος μεγάλο 4-στοιχειοσύνολο και έτσι ο αλγόριθμος τερματίζεται.

2.5 Συναρτήσεις που περιέχονται στο αλγόριθμο Apriori

2.5.1 Συνάρτηση Apriori-gen.

Η συνάρτηση Apriori-gen, όπως έχει αναφερθεί και στον ψευδοκώδικα παραπάνω, παράγει τα υποψήφια k-itemsets C_k από τα συχνά (k-1)-itemsets L_k που έχουμε βρει από προηγούμενα περάσματα. Έχει σαν είσοδο το σύνολο Lk και σαν έξοδο έχει το υπερσύνολο του Lk δηλαδή το σύνολο C_k. Η συνάρτηση περιλαμβάνει δύο φάσεις. Η πρώτη φάση αφορά την λεγόμενη "ένωση" , το join-step και η άλλη το "ξεκαθάρισμα" prune-step .

Περιγραφή της ένωσης, join-step: $C_k = \{X \smile Y \mid X,Y \varepsilon L_{k-1}, \mid X\ Y \mid = k-2\}$

Περιγραφή του ξεκαθαρίσματος, prune-step: : $C_k = \{X \varepsilon C_k, \mid X \text{ contains members of } L_{k-1}\}$

- Ένωση: στο βήμα αυτό έχουμε την ένωση δύο (k-1) items τα οποία items έχουν κοινά ακριβώς (k-2) items. Το itemset που θα δημιουργηθεί μετά την ένωση θα έχει σύνολο k items δηλαδή θα περιλαμβάνει τα (k-2) κοινά items συν και το μη κοινό item από τα δύο (k-1)- itemsets. Για να διευκολυνθούμε μπορούμε να εκμεταλλευτούμε την διάταξη που έχουμε δώσει στα items με την γλώσσα SQL, η οποία έχει την εξής δομή:

```
insert into Ck ' '
        select   X[1], X[2], X[3], . . . ,X[k-1], Y[k-1]
            from   Lk-1  X, Lk-1 Y
            where  X[1]=Y[1] , . . . , X[k-2] =Y[k-2] , X[k-1]<Y[k-1]
```

- Ξεκαθάρισμα: σε αυτό το βήμα έχουμε και την αρχή του αλγόριθμου Apriori, αφαιρώντας εκείνα τα itemsets που έχουν το λιγότερο ένα (k-1) υποσύνολο που δεν ανήκει στο σύνολο L_k.

2.5.2 Συνάρτηση subset.

Ο ρόλος της συνάρτησης subset είναι να υπολογίσει και να βρει το υποσύνολο του C_k , το οποίο αποτελείται από όλα τα itemsets που περιέχονται σε κάθε transaction ξεχωριστά. Τα υποψήφια που βρίσκονται σε μια δοσοληψία t. Σημαντικό είναι και ο τρόπος με τον οποίο πρέπει να γίνεται η αποθήκευση των itemsets για να μην δημιουργηθεί πρόβλημα στην καθυστέρηση όλης της διαδικασίας. Γι' αυτό τον λόγο η αποθήκευση γίνεται σε ένα δέντρο κερματισμού (hash-tree) , το οποίο αποτελείται από κόμβους και μπορεί να είναι ένας κόμβος φύλλου, μια δηλαδή λίστα στην οποία αποθηκεύονται τα itemsets, είτε ένας εσωτερικός κόμβος , ένας πίνακας κατακερματισμού οπού περιέχονται πληροφορίες που αφορούν τα itemsets που ψάχνουμε.

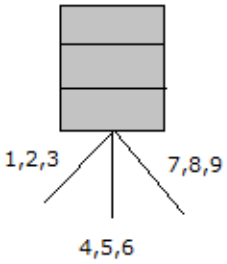
Σε έναν πίνακα, κουβά (bucket) εννοούμε αυτό που δίνει πληροφορίες από τον έναν κόμβο στον άλλον και ρίζα είναι η αρχή του δέντρου , εκεί οπού ξεκινάει το itemsets που βάζουμε στο δέντρο μέχρι να καταλήξει σε κάποιο φύλλο στο οποίο έχουμε αποθηκεύσει και άλλα itemsets. Αν υποθέσουμε ότι η ρίζα έχει βάθος 1 και το βάθος του εσωτερικού κόμβου είναι d, τότε το βάθος στο οποίο βλέπουμε είναι d+1.

Δημιουργώντας ένα δέντρο hash-tree αποθηκεύονται k-itemsets του συνόλου C_k το οποίο θα έχει βάθος το πολύ k+1, δηλαδή το μονοπάτι θα φτάνει σε ένα φύλλο και θα έχει μέγιστο k hash-tree.

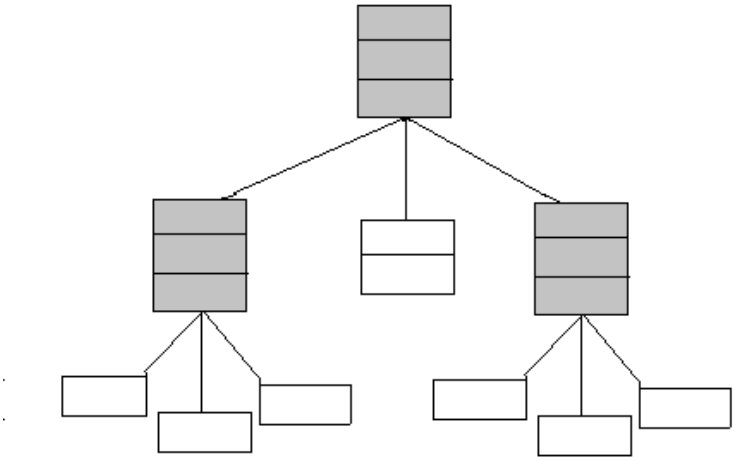
Η διαδικασία είναι αναδρομική και είναι η εξής:

Επιλέγοντας την συνάρτηση του καταμερισμού ,επιλέγουμε το υπόλοιπο της διαίρεσης με το π.χ. 3. Έτσι εμφανίζονται 3 κόμβοι σε κάθε εσωτερικό κόμβο. Η συνάρτηση του καταμερισμού γίνεται για όλα τα items της δοσοληψίας και συγκεκριμένα μέχρι και το τρίτο από το τέλος item, έτσι ώστε να βρούμε τα 3-itemsets.

Η μορφή ενός Hash Function είναι η παρακάτω:



Η μορφή του Candidate Hash Tree:



Η διαδικασία σταματάει όταν βρεθούμε σε ένα φύλλο γιατί θα έχουμε αποθηκεύσει εκείνα τα itemsets στην υπό-αναζήτηση στο σύνολο C_k που είναι αποθηκευμένα στο φύλλο και που ήδη βρίσκονται στην t. Αν είμαστε σε κάποιο εσωτερικό κόμβο επαναλαμβάνουμε την διαδικασία του καταμερισμού, στην ρίζα του δέντρου για κάθε item που βρίσκεται μετά το i μέχρι να φτάσουμε σε κάποιο φύλλο οπού θα σταματήσει η διαδικασία.

2.5.3 Apriori TID.

Ο Apriori TID στηρίζεται στη γενική ιδέα του Apriori με τη διαφορά ότι ο πίνακας δοσοληψιών D διαβάζεται μόνο μια φορά στην αρχή. Οι πληροφορίες που περιλαμβάνει στον πίνακα C_k -. Η κάθε εγγραφή του πίνακα της μορφής <TID,{X_k}> όπου X_k είναι το υποψήφιο k-itemset που περιλαμβάνει στη δοσοληψία με αναγνωριστικό TID. Ο αρχικός πίνακας δοσοληψιών είναι ο C_1 με τη μόνη διαφοροποίηση ότι κάθε item i αντικαθιστάται από 1-itemset {1}. Για τις τιμές του k μεγαλύτερες του 1 ο πίνακας C_k – προκύπτει από το k βήμα του αλγορίθμου και κάθε εγγραφή του περιέχει itemset από το εκάτοστε σύνολο C_k. Η χρησιμοποίηση του C_k – εμφανίζει καλύτερα αποτελέσματα για μεγαλύτερες τιμές του k. Δίνονται καλύτερα αποτελέσματα λόγω ότι δεν αναπαρίστανται δοσοληψίες που δεν έχουν συχνά itemset. Η κάθε εγγραφή στον πίνακα C_k – μικραίνει λόγω ότι δημιουργούνται λιγότερα υποψήφια itemset όσο το k μεγαλώνει. Ο Apriori TID έχει καλύτερα αποτελέσματα στα τελευταία περάσματα για τις πρώτες επαναλήψεις χρησιμοποιείται ο Apriori.

Ψευδοκώδικας αλγορίθμου Apriori TID

1. L_i=(large 1-itemsets);
2. C_i=database D;
3. For (k=2; L_k <>0; k+1) do begin
4. C_k=Apriori-gen (L_{k-1});// δημιουργία υποψηφίων
5. C_k=0;
6. Forall entries t€ C_{k-1} do begin
7. // εύρεση υποψηφίων που περιέχονται στην εγγραφή t C_t =(c∈ C: (c-c[k]∈t.set_of_itesmet ^ (c-c[k-1]) ∈ t.set_of_itesmet
8. Forall candidates c ∈ C_t do
9. c.count++
10. if (C_t≠0) then C_k-+=<t.TID,C_t >;
11. End
12. L_k ={c∈C_k : c.count>= minsup)
13. End
14. Return ∪$_k$$L_k$;

BFS και απευθείας μέτρηση των υποψηφίων

Η Breadth-First Search , BFS είναι η κατά πλάτος αναζήτηση που χρησιμοποιείται για την αναπαράσταση ενός δέντρου συνόλων itemset. Ο αλγόριθμος Dynamic Itemset Counting ανήκει στην κατηγορία αύτη και χαρακτηριστικό του είναι όταν ένα Itemset ξεπεράσει το όριο minsup τότε ο αλγόριθμος παράγει υποψήφιους που προέρχονται από αυτό χωρίς να μετρήσει μέχρι τέλους τα Itemset. Για αυτό τον λόγο χρησιμοποιείται ένα δέντρο με προθέματα όπου στον κάθε κόμβο γίνεται η αποθήκευση ενός ακριβώς Itemset.

BFS και τομή συνόλων από TID

Ο αλγόριθμος Partition υπολογίζει την υποστήριξη των Itemset που γίνεται αποθήκευση για κάθε Itemset. Το μειονέκτημα των αλγορίθμων είναι ότι πιάνουν πολύ χώρο οι λίστες των TIDs. Γι' αυτό ο αλγόριθμος Partition διασπά τον αρχικό πίνακα σε υποπίνακες ώστε οι λίστες των TIDs να χωρούν στην κύρια μνήμη του συστήματος.

DFS και απευθείας μέτρηση των υποψηφίων

Ο αλγόριθμος EP-growth ανήκει στην κατηγορία όπου γίνεται μέτρηση της υποστήριξης, διάβασμα του πίνακα δοσοληψιών, των υποψήφιων εκείνων μόνο που ανήκουν σε ένα μόνο κόμβο του δέντρου κλάσεων Ε. Στο πρώτο του βήμα ο πίνακας δοσοληψιών αντικαθιστάται από το δέντρο Frequent Pattern tree που παράγει ο ίδιος ο αλγόριθμος.

DFS και τομή συνόλων από TID

Ο αλγόριθμος Elcat χρησιμοποιείται στην κατά βάθος αναζήτηση. Στο δέντρο κλάσεων τα σύνολα από TIDs κρατούνται μόνο για τα itemsets που βρίσκονται στο μονοπάτι από τη ρίζα μέχρι τον κόμβο που εξετάζουμε.

2.6 Διαφορά Apriori με Apriori TID

Η κύρια διαφορά μεταξύ των δύο αλγόριθμων Apriori και AprioriTID είναι ότι με τον αλγόριθμο Apriori σε κάθε επανάληψη λαμβάνετε υπόψη μόνο οι λίστες που είναι μεγάλες ενώ ο αλγόριθμος Apriori TID η βάση δεδομένων χρησιμοποιείται στην αρχή, μετά την πρώτη επανάληψη δεν χρησιμοποιείται για υπολογισμό της εμπιστοσύνης των υποψήφιων μεγάλων λιστών, αλλά γίνεται χρήση μίας κωδικοποίησης υποψηφίων μεγάλων λιστών που είχε χρησιμοποιηθεί στην προηγούμενη επανάληψη. Συγκεκριμένα ο Apriori αποδίδει καλύτερα από ότι ο AprioriTID στα αρχικά περάσματα, γιατί δίνει πιο γρήγορα αποτελέσματα, αλλά στην συνέχεια την σκυτάλη την παίρνει ο αλγόριθμος AprioriTID ο οποίος έχει καλύτερη απόδοση από ό, τι Apriori γιατί δίνει γρηγορότερα αποτελέσματα στις μεταγενέστερες επαναλήψεις.

2.7 Μέτρα ενδιαφέροντος των κανόνων συσχέτισης

Τα μέτρα ενδιαφέροντος είναι ένας τρόπος ένδειξης της σημαντικότητας και της εμπιστοσύνης των κανόνων συσχέτισης. Ωστόσο οι κανόνες που παράγονται θα πρέπει να ικανοποιούν κάποια κριτήρια που έχει ορίσει ο χρηστής. Παρακάτω παρουσιάζονται κάποια από τα μέτρα τα οποία βοηθούν ενδεχομένως στη λήψη αποφάσεων καθώς βοηθούν τον λήπτη να αναγνωρίσει χρήσιμα πρότυπα γνώσης που εξάγουν οι κανόνες συσχέτισης .
Ένας κανόνας συσχέτισης ορίζεται ως LHS→RHS. Στη συνέχεια θα αναφερθούν κάποια από τα μέτρα ενδιαφέροντος.

2.7.1 Υποστήριξη

Η υποστήριξη(support) μπορεί να χρησιμοποιηθεί και ως ένδειξη σε ένα σύνολο στοιχείων με ποια συχνότητα εμφανίζεται ένας κανόνας και αυτό έχει ως αποτέλεσμα το πόσο σημαντικός είναι δηλαδή το LHS και το RHS του κανόνα. Η υποστήριξη ορίζεται ως

$$\text{Support}=n(\text{LHS}\cap\text{RHS})/N \quad \text{ή} \quad \text{Support}=p\,(\text{LHS}\cap\text{RHS})$$

Όπου N είναι ο συνολικός αριθμός των περιπτώσεων που εξετάζονται και n(LHS) φανερώνει τον αριθμό των περιπτώσεων που ικανοποιούν το αριστερό μέλος.

2.8 Αντιπροσωπευτικοί κανόνες συσχέτισης - Representative association Rules.

Ο αριθμός των κανόνων που δημιουργούνται είναι αρκετά μεγάλος τις πιο πολλές φορές αν δεν γίνει εφαρμογή κάποιων κριτηρίων για τη σημαντικότητα των κανόνων. Η προσέγγιση των Αντιπροσωπευτικών Κανόνων Συσχέτισης έχει σαν στόχο την αντιμετώπιση του προβλήματος της παραγωγής πολλών κανόνων.

Οι αντιπροσωπευτικοί κανόνες συσχέτισης με ορισμένη την ελάχιστη υποστήριξη s και ελάχιστη εμπιστοσύνη c συμβολίζεται ως RR(s,c) και ορίζεται ως εξής :

$$RR(s,c)=\{w\in AR(s,c) \mid \neg\exists\ w' \in AR(s,c), w'\neq w \text{ και } w \in C(w') \}.$$

Ο ορισμός που διατυπώθηκε υποδηλώνει ότι κανένας αντιπροσωπευτικός κανόνας δε μπορεί να ανήκει στη κάλυψη ενός άλλου κανόνα.

Ακόμα, στους αντιπροσωπευτικούς κανόνες συσχέτισης RR(s,c) ισχύουν παρακάτω δυο ιδιότητες .

- Εάν $w \in RR(s,c)$ *τότε* $C(w) \subseteq AR(s,c)$.
- Οποιοσδήποτε κανόνας συσχέτισης μπορεί να παραχθεί από έναν αντιπροσωπευτικό κανόνα με τη χρήση του τελεστή κάλυψης δηλαδή, $\forall w \in AR(s,c)\ \exists w' \in RR(s,c) : w \in C(w)$.

2.8.1 Τελεστής Κάλυψης - Cover operator.

Οι κανόνες προκύπτουν από έναν κανόνα συσχέτισης όπου η εφαρμογή γίνεται στον αρχικό τελεστή κάλυψης.

$$X => Y, Y \neq \emptyset$$

$$C(X => Y) = \{X \cup Z => V \backslash Z, V\ Y\ and\ Z \cap V = \emptyset\ and\ V \neq \emptyset\}$$

$$C(X \Rightarrow Y) = \{X \cup Z \Rightarrow V \mid Z, V \subseteq Y\ and\ Z \cap V = \varnothing\ and\ V \neq \varnothing\}$$

Όλοι οι κανόνες συσχέτισης που ικανοποιούν τις απαιτήσεις που έχουν οριστεί για την ελάχιστη υποστήριξη s και ελάχιστη εμπιστοσύνη c θα καλούνται ως AR(s,c).

Ο κανόνας $C(X => Y)$ αποτελείται από ένα υποσύνολο των items που περιέχονται στον κανόνα $X => Y$. Το πρώτο μέρος ενός κανόνα που ανήκει στο σύνολο $C(X => Y)$ αποτελείται από τα items του X και πιθανώς κάποια από το Y. Το δεύτερο μέρος (consequent) ενός τέτοιου κανόνα είναι ένα μη υποσύνολο των υπόλοιπο items το Y.

Ο τελεστής κάλυψης εμφανίζει κάποιες ιδιότητες που θα αναφερθούν στη συνέχεια:

- Ιδιότητα 1

Έστω w ένας κανόνας συσχέτισης με υποστήριξη s και εμπιστοσύνη c. Κάθε κανόνας w΄ που θα ανήκει στην κάλυψη C(w) και έχει υποστήριξη όχι μικρότερη από s εμπιστοσύνη επίσης όχι μικρότερη από c. Αυτό συνεπάγεται ότι εάν ένας κανόνας w ανήκει στο AR(s,c),τότε κάθε κανόνας w΄ στο C(w) θα ανήκει στο AR(s,c).

- Ιδιότητα 2

Έστω ένας κανόνας συσχέτισης X⇒Y. Τότε το πλήθος των κανόνων που περιέχονται στη κάλυψη αυτού του κανόνα θα είναι $|C(w)|=3^n-2^n$, όπου n=|Y|.

- Ιδιότητα 3

  - Αν ένας κανόνας συσχέτισης w: (X⇒Y) είναι μικρότερος από έναν άλλο κανόνα συσχέτισης w′: (X′⇒Y′) τότε ισχύει w ∈ C(w′) αν και μόνο αν X∪Y ⊆X′∪Y′ και X⊇X′.

  - Εάν ένας κανόνας w είναι μεγαλύτερος από ένα κανόνα συσχέτισης w′ τότε w ∉ C(w′).
  - αν w: (X⇒Y) και w′: (X′⇒Y′) είναι δυο διαφορετικοί κανόνες συσχέτισης με το ίδιο μήκος τότε ισχύει ότι w C(w′) αν και μόνο αν X∪Y=X′∪Y′ και X⊃X′.

- Ιδιότητα 4

  Έστω δυο κανόνες συσχέτισης w: X⇒Y και w′: X′⇒Y′. τότε ο κανόνας w θα ανήκει στη κάλυψη του κανόνα w′, αν και μόνο αν X∪Y ⊆X′∪Y′ και X⊇X′.
  Με μαθηματική σχέση μπορεί να εκφραστεί ως w∈ C(w′)⇔ X∪Y⊆X′∪Y′ ∧ X⊇X′.

2.8.2 Παραγωγή αντιπροσωπευτικών κανόνων συσχέτισης

Ο αλγόριθμος FastGenAllReprentatives θεωρείται αλγόριθμος παραγωγής των αντιπροσωπευτικών κανόνων προϋποθέτει ότι έχουν βρεθεί συχνά itemsets και θέτει ως προϋπόθεση να είναι γνωστά όλα τα συχνά itemsets.

Οι δύο ιδιότητες που στηρίζεται η λειτουργία είναι:

- Έστω $\emptyset \neq X \subset Z \subseteq I \quad και \ r \ κανόνας \ της \ μορφής \ r: (X \Rightarrow Z \backslash X) \in AR(s,c)$

Τότε ο κανόνας αυτός θα ανήκει στο RR(s,c) αν ισχύουν τα εξής:

- maxSup<=s maxSup/sup(X)<c όπου sup maxSup=max(\sup(Z′)‖ $Z \subset Z'(I) \cup \{0\}$)
- $\exists X', \emptyset \neq X' \subset X \ τέτοιο \ ώστε \ (X' \Longrightarrow Z \backslash X') \in AR \ (s,c)$
  - Έστω $\emptyset \neq X \subset Z \subseteq I$. Αν sup(Z)= sup(Z′) τότε κανένας κανόνας της μορφής $(X \Rightarrow Z \backslash X) \in AR(s,c) με \emptyset \neq X \subset Z$ δεν ανήκει στο RR(s,c).

  Δεν μπορούν να παραχθούν αντιπροσωπευτικοί κανόνες αν για το συχνό itemset Z ισχύει maxSup=sup(Z).

Για να γίνει κατανοητός ο τρόπος παραγωγής αντιπροσωπευτικών κανόνων συσχέτισης παρουσιάζεται ο παρακάτω αλγόριθμος:

1. Procedure FastGenAllReprentatives (all frequent itemsets L);
2. Forall Z∈F do begin
3. k= apolito z; maxSup= max((sup(Z') apolito ZⵚZ'∈F_{k+1})∪(0));
4. if Z.sup ≠ maxsup then begin
5. A_1 =({Z[1], {Z[2],…{Z[K]})
6. For (i=1; (A_j.0) and (i<k); i++) do begin
7. Forall X∈A_i do begin
8. Find Y∈L_i such that Y=X;
9. XCount=Y.count;
10. //is X→Z∵X an association rule?
11. If (Z.count/ XCount. C) then begin
12. // aren't there any rep. rules longer that X=>Z∵X
13. If (maxsup/Xcount<c) then
14. Print (X, <⇔>>, Z ∵ X, <<with support: <<, Z.count, <<and confidence: <<, Z.count/XCount);
15. //antecedents of ass. Rules are not extended
16. A_i =A_i ∵{X};
17. Endif
18. Endfor
19. A_{i+1}= ApriorіGen(A)
20. Endfor
21. Endif
22. Endfor
23. endproce

Ο αλγόριθμος υπολογίζει τους αντιπροσωπευτικούς κανόνες συσχέτισης από κάθε itemset του L. Έστω ότι Z το υπό εξέταση itemset τότε k=Z. Το maxsup υπολογίζεται από ως η μέγιστη από τις τιμές εμπιστοσύνης των itemset στο L_{1+k} που είναι υπερσύνολα του Z. Αν δεν υπάρχει υπερσύνολο του Z στο L_{1+k} τότε το maxSup =0. Τα δύο πράγματα που μπορεί να ισχύουν είναι maxsup>s ή maxSup=0.

- maxSup=0 τότε κανένας αντιπροσωπευτικός κανόνας δεν μπορεί να εξεταστεί από το Z.
- maxsup>s τότε κατασκευάζονται όλα τα antecedents με μονά items.

Εάν το sup(Z) γίνει ίσο με το maxSup τότε δεν θα μπορέσει κανένας αντιπροσωπευτικός κανόνας να παραχθεί από το Z (εντολή 4). Αντίθετα θα κατασκευαστούν όλοι οι πρόγονοι(antecedents) του με μονά αντικείμενα (εντολή 5).
Στην εντολή 6 εξετάζεται κάθε itemset X από το σύνολο των i-itemsets A_j για την ύπαρξη αντιπροσωπευτικών κανόνων της μορφής X=>Z\X. Για γίνει έλεγχος της συνθήκης αν είναι κανόνας συσχέτισης η εμπιστοσύνη του sup(Z)/sup(X). Αντιπροσωπευτικοί θεωρούνται μόνο όταν τηρούνται και οι συνθήκες.

Βάση της αρχής Apriori εάν το Z είναι συχνό τότε και το X θα είναι συχνό (εντολή 10?) . Για να μπορέσει να ελέχθη αν το $X \Rightarrow Z \backslash X$ ανήκει στους κανόνες συσχέτισης θα πρέπει να οριστεί η εμπιστοσύνη του sup(Z)/sup(X). Το sup(Z) είναι γνωστό sup(Z)=Z.count, αντίθετα για να βρεθεί το

sup(X) θα πρέπει να βρεθούν από όλα τα συχνά itemsets εκείνο το Y που θα ισχύει Y=X έτσι ώστε να υπολογιστεί το sup(Y)=Y.count.

Για να είναι αντιπροσωπευτικός ένας κανόνας $X \Rightarrow Z \backslash X$ θα πρέπει να ισχύουν και τα δυο σκέλη της ιδιότητας 1. Άρα θα πρέπει να γίνει έλεγχος ,η συνθήκη (ii) ικανοποιείται αυτόματα ενώ η συνθήκη (i) για να ικανοποιηθεί θα πρέπει να εξασφαλιστεί ότι $maxSup \leq s$ *ή ότι* $maxSup/sup(X) < c$. Αν γίνει δεκτό $maxSup=0$ τότε η συνθήκη ικανοποιείται εάν όμως $maxSup > s$ τότε κρίνεται αναγκαία η εξέταση της εντολής 13. Στην εντολή 16 αφαιρείται από το A_i το πρώτο μέλος κάθε αντιπροσωπευτικού κανόνα που βρίσκεται.
Εφόσον έχουν βρεθεί όλοι οι κανόνες με μήκος k και i- antecedents από το Z, τα A_i παράγουν τα (i+1)-antecedents A_{i+1} κάνοντας χρήση της συνάρτησης AprioriGen του αλγόριθμου Apriori(εντολή 19). Τέλος στο A_{i+1} δεν περιέχεται κανένα itemset X τέτοιο ώστε να μπορεί να ισχύει ο κανόνας $X \Rightarrow Z \backslash X$ να είναι υπό την κάλυψη του κανόνα $X' \Rightarrow Z \backslash X'$ όπου $X' \subset X$ άρα μπορεί να ικανοποιείται και συνθήκη (ii) της ιδιότητας 1.

2.9 Ποσοτικοί κανόνες συσχέτισης – Quantitative Association Rules.

Για την παραγωγή κανόνων συσχέτισης από έναν πίνακα δοσοληψιών, όπου το σύνολο αποτελείται από αντικείμενα τα λεγόμενα items, έχουμε αναφέρει μερικούς τρόπους στις προηγούμενες σελίδες. Υπάρχουν όμως και οι πίνακες βάσης που παράγονται από τις βάσεις δεδομένων και έχουν άλλη μορφή από τους πίνακες δοσοληψιών που έχουμε συναντήσει νωρίτερα.

Ένας πίνακας βάσης αποτελείται από εγγραφές, τις γραμμές δηλαδή του πίνακα βάσεων δεδομένων και από τα γνωρίσματα – attributes τα οποία αποτελούν τον αριθμό από τις στήλες και είναι τα αντίστοιχα items που συναντάμε σε έναν πίνακα δοσοληψιών. Μια εγγραφή αποτελείται από ένα σύνολο γνωρισμάτων.

Όταν πραγματοποιείται μία εγγραφή γίνεται μια συσχέτιση παίρνει δηλαδή για ένα γνώρισμα μια τιμή, είτε «1» (αληθές) αν η συγκεκριμένη δοσοληψία έχει το γνώρισμα που υποδηλώνει το item , είτε διαφορετικά παίρνει τη τιμή «0» (ψευδές). Επομένως το πρόβλημα αυτό μπορεί να αποδοθεί και ως *Boolean Association Problem,* διότι σε έναν πίνακα βάσης δεδομένων δεν μπορούν να περιέχονται Boolean γνωρίσματα. Τα γνωρίσματα χωρίζονται σε δύο κατηγορίες, τα ποσοτικά όταν αναφερόμαστε σε τιμές, εισόδημα, ύψος, ηλικία κ.τ.λ. και σε κατηγορικά – μη αριθμητικά γνωρίσματα όπως μάρκα αυτοκινήτου, περιοχή, ομάδα αίματος, ταχυδρομικός κώδικας κ.τ.λ.. Έτσι έχουμε τον πίνακα με τα Boolean γνωρίσματα τα οποία είναι κατηγορικά

Στη συνέχεια καλούμαστε να λύσουμε αυτό το πρόβλημα που το αποκαλούν πρόβλημα εύρεσης Ποσοτικών Κανόνων Συσχέτισης (Quantitative Association Rules problem) που ασχολείται με την παραγωγή κανόνων συσχέτισης από πίνακες με ποσοτικά και κατηγορηματικά γνωρίσματα. Ένας κανόνας της μορφής αυτής είναι ο παρακάτω

{Ηλικία: 40..49} και {Παντρεμένος: Ναι}→{Αριθμός Αυτοκίνητων : 2}

2.9.1 Ίσο-βαθύς κατάτμηση.

Το πρόβλημα που παρουσιάζεται είναι με πιο τρόπο θα παράγουμε ποσοτικούς κανόνες συσχέτισης από κατηγορικά και ποσοτικά γνωρίσματα. Την λύση σε αυτό το πρόβλημα μας την δίνουν οι R.Agrawal και R.Strikant, δηλαδή επιχείρησαν να αντιστοιχίσουν πρόβλημα εύρεσης Ποσοτικών Κανόνων στο πρόβλημα εύρεσης Boolean Association Rules. Το πρόβλημα είναι σχετικά εύκολο αν τα ποσοτικά είτε τα κατηγορικά γνωρίσματα είχαν λίγες πιθανές τιμές γιατί σε αυτή την περίπτωση μπορούμε να αντιστοιχίσουμε πιο εύκολα. Θα υπάρχουν τόσα πεδία όσες είναι και οι τιμές που μπορεί να πάρει ένα γνώρισμα δηλαδή είτε την τιμή 1 είτε την τιμή 0. Σε άλλη περίπτωση όπου υπάρχουν πολλές τιμές για ένα γνώρισμα τότε χωρίζουμε το σύνολο τιμών σε διαστήματα και τα Boolean γνωρίσματα παίρνουν την μορφή <γνώρισμα 1, διάστημα 1>. Άρα το πρόβλημα μετατίθεται στην εύρεση της κατάλληλης κατάτμησης των γνωρισμάτων.

Η κατάτμηση αυτή είναι χωρισμένη σε δυο τμήματα:

- MinSup: Όπου αν τα διαστήματα είναι πολλά τότε υπάρχει η πιθανότητα η εμπιστοσύνη ενός διαστήματος να είναι μικρή. Για αυτό το λόγο πρέπει η κατάτμηση να παράγει μεγαλύτερα διαστήματα δηλαδή λιγότερα σε νούμερο.
- MinConf: Όταν μερικοί κανόνες έχουν ελάχιστη εμπιστοσύνη (minimum confidence) εξαιτίας του χωρισμού των τιμών σε διαστήματα και για αυτό το λόγο χάνεται χρήσιμη πληροφορία. Για αυτό το λόγο πρέπει η κατάτμηση να παράγει μικρότερα διαστήματα δηλαδή περισσότερα σε νούμερο.

Για την σωστή κατάτμηση χωρίς να χάνεται πολύτιμος χρόνος και χωρίς να γίνεται σύγκρουση ανάμεσα στις δύο παραπάνω λύσεις, βάζουμε ένα όριο στη διαδικασία ένωσης των διαστημάτων, δηλαδή όταν ένα διάστημα έχει υποστήριξη ίση με μια μέγιστη τιμή maxsup τότε δεν αυξάνουμε άλλο το διάστημα. Δημιουργούμε μικρά διαστήματα τα οποία θα ενώνονται στη συνέχεια φτιάχνοντας νέα μεγαλύτερα με μεγαλύτερη υποστήριξη. Εν συνεχεία καλούμαστε να ορίσουμε το πώς πρέπει να γίνει η τμηματοποίηση του κάθε γνωρίσματος. Σημαντικό ρόλο παίζει τόσο ο αριθμός των διαστημάτων που θα φτιάξουμε όσο και το μέγεθος που θα έχουν, ώστε να έχουμε όσο τον δυνατόν λιγότερη χαμένη πληροφορία.

Έστω λοιπόν ότι R είναι το σύνολο των κανόνων που βγαίνουν από όλα τα πιθανά διαστήματα από τις τιμές των ποσοτικών γνωρισμάτων και R' το σύνολο από τα αρχικά διαστήματα στα οποία κατατμήσαμε τα ποσοτικά γνωρίσματα. Ένας κανόνας που μπαίνει ανάμεσα στο σύνολο R και R' είναι ότι τα δύο αυτά σύνολα πρέπει να είναι όσο πιο κοντά γίνεται, έτσι ώστε να έχουμε πολύ μικρή απώλεια πληροφορίας. Το σύνολο R' αποτελεί γενίκευση του R, οι τιμές του δηλαδή συσχετίζονται με τις τιμές του R. Το μέτρο αυτό λέγεται και *μέτρο Μερικής Ολοκλήρωσης (Partial Completeness)* και είναι ο λόγος K των τιμών των συνόλων?, δηλαδή K=sup(r') / sup(r) όπου r ϵ R και r' ϵ R' είναι η μικρότερη γενίκευση του r.

Έτσι έχουμε την ισο-βαθύς κατάτμηση στην οποία όλα τα διαστήματα έχουν την ίδια υποστήριξη και είναι N= [2*n / (m* (K-1))], όπου n ο αριθμός των ποσοτικών γνωρισμάτων, m η ελάχιστη υποστήριξη και K ο καλύτερος βαθμός μερικής ολοκλήρωσης.

Αφού ολοκληρωθεί η κατάτμηση βρίσκουμε τα συχνά ή μεγάλα itemsets και στην συνέχεια έχουμε την παραγωγή των κανόνων συσχέτισης όπως παράγονται οι κανόνες συσχέτισης βάση της λογικής που χρησιμοποίει ο αλγόριθμος Apriori μόνο που εδώ έχουμε την πράξη της ένωσης ώστε να προκύψουν τα νέα διαστήματα.

2.9.2 Κανόνες με βάση την απόσταση.

Με την προηγούμενη διαδικασία δεν μετράμε την πυκνότητα που περιέχεται σε ένα διάστημα. Για αυτό το λόγο οι R.J. Miller και Y. Yang πρότειναν να παραχθούν κανόνες που θα γίνονται με κατάτμηση με βάση την απόσταση. Η κατάτμηση αυτή παρουσιάζεται πιο λογική καθώς συνυπολογίζει τις αποστάσεις μεταξύ των διαφορετικών τιμών.

Οι στόχοι που ικανοποιεί είναι οι εξής:

- ✓ Πρώτον θα υπάρχει ένα μέτρο που θα μετρά την ποιότητα του διαστήματος, δηλαδή την απόσταση των σημείων,
- ✓ δεύτερον ο κανόνας συσχέτισης ανάμεσα στα items στα ποσοτικά γνωρίσματα θα περιγράφεται από το $C_1 \rightarrow C_2$ όπου τα items του C_1 είναι πολύ κοντά με τα items του ικανοποιούνται του C_2 ,

- ✓ τρίτον τα μέτρα σημαντικότητας ενός κανόνα πρέπει να φανερώνουν την απόσταση που έχουν τα σημεία μεταξύ τους.

Οι κανόνες που θα παραχθούν θα είναι κανόνες μεταξύ συστάδων (clusters). Οι συστάδες είναι ομάδες σημείων όπου μεταξύ αυτών υπάρχουν μικρές αποστάσεις και αποτελούνται από σημεία της ίδιας ομάδας ενώ αντίθετα σημεία διαφορετικών ομάδων εμφανίζονται σε μεγάλες αποστάσεις. Βασιζόμενοι στο δεύτερο στόχο πρέπει να βρεθεί το σύνολο εκείνο των συστάδων που θα ελαχιστοποιεί ένα μέτρο απόστασης.

Ένα μέτρο για την ποιότητα των συστάδων είναι η μέση ανά δυο απόσταση των σημείων η όποια καλείται διάμετρος συστάδας.
Ορίζοντας τη διάμετρο d στο X ενός συνόλου από εγγραφές $S=\{t_i : 1\leq i\leq N\}$ είναι η μέση απόσταση ανά δυο προβολών των εγγραφών X.

2.10 Περίληψη

Οι κανόνες συσχέτισης δημιουργήθηκαν για τις ανάγκες των υπεραγορών ώστε να καταχωρηθούν οι συναλλαγές του κάθε πελάτη ηλεκτρονικά αναλύοντας το καλάθι αγοράς. Με τους κανόνες συσχέτισης εξετάζονται πολλά καλάθια αγοράς πελατών συνδυάζοντας τα αντικείμενα μεταξύ τους με σχέσεις εξάρτησης. Οι κανόνες εφαρμόζονται στην προώθηση προϊόντων, την τοποθέτηση προϊόντων στα ράφια καταστημάτων και την διαχείριση αποθεμάτων. Το πρόβλημα Εξαγωγής Κανόνων Συσχέτισης αφορά την αναζήτηση και εύρεση όλων εκείνων των κανόνων συσχέτισης που θα πρέπει να ικανοποιούν κάποια κατώτατα όρια σχετικά με την υποστήριξη (support) και την εμπιστοσύνη (confidence). Ωστόσο, η υποστήριξη (support) ενός κανόνα θα πρέπει να είναι μεγαλύτερη από μια τιμή που ορίζουμε ως όριο και την ονομάζουμε ελάχιστη υποστήριξη (minsup),ακόμα και η εμπιστοσύνη πρέπει να είναι μεγαλύτερη από το όριο που ονομάζεται ελάχιστη εμπιστοσύνη (minconf). Ο αλγόριθμος Apriori είναι ένας κλασσικός αλγόριθμος που μας βοηθάει στην παραγωγή των κανόνων συσχέτισης. Υπάρχουν πολλοί αλγόριθμοι με τους οποίους μπορούμε να επεξεργαστούμε τα μεγάλα σύνολα δεδομένων και να παράγουμε κανόνες συσχέτισης, δύο νέοι και πολύ διαδεδομένοι αλγόριθμοι είναι ο Apriori και AprioriTid αλγόριθμοι που μας βοηθάει στην εξαγωγή των κανόνων συσχέτισης. Ο αλγόριθμος Apriori είναι ένας κλασσικός αλγόριθμος που μας βοηθάει στην παραγωγή των κανόνων συσχέτισης. Υπάρχουν πολλοί αλγόριθμοι με τους οποίους μπορούμε να επεξεργαστούμε τα μεγάλα σύνολα δεδομένων και να παράγουμε κανόνες συσχέτισης, δύο νέοι και πολύ διαδεδομένοι αλγόριθμοι είναι ο Apriori και AprioriTid αλγόριθμοι που μας βοηθάει στην εξαγωγή των κανόνων συσχέτισης.

Κεφάλαιο 3

Πρόγραμμα WEKA

3.1 Εισαγωγή weka

Το Weka (Wekato Environment for Knowledge Analysis) είναι μια συλλογή από αλγορίθμους μηχανικής μάθησης για τρεις κατηγορίες (συσχέτιση, κατηγοριοποίηση, συσταδοποίηση) είναι δηλαδή ένα software για εξόρυξη δεδομένων γραμμένο σε JAVA το οποίο περιέχει μεθόδους για:

- Προεπεξεργασία Δεδομένων.
- Ταξινόμηση.
- Συσταδοποίηση.
- Εύρεση Κανόνων Συσχέτισης.

Το WEKA περιέχει τα εξής εργαλεία: φίλτρα, κατηγοριοποιητές, ταξινομητές, συσχετιστές και επιλογείς χαρακτηριστικών που δημιουργούν γραφικά με δισδιάστατα γραφήματα Έτσι με βάση τις γραφικές αναπαραστάσεις που δημιουργούνται των επιλεγμένων δεδομένων δοκιμάζονται και προβλέπεται η απόδοση των εναλλακτικών μοντέλων που θα επιλεγούν να χρησιμοποιηθούν τελικά για την διαμόρφωση της «θαμμένης» γνώσης.

3.2 Εγκατάσταση

Το weka μπορείτε να το βρει κάποιος και να το κατεβάσει από τις παρακάτω διαθέσιμες ιστοσελίδες:

http://sourceforge.net/projects/weka/

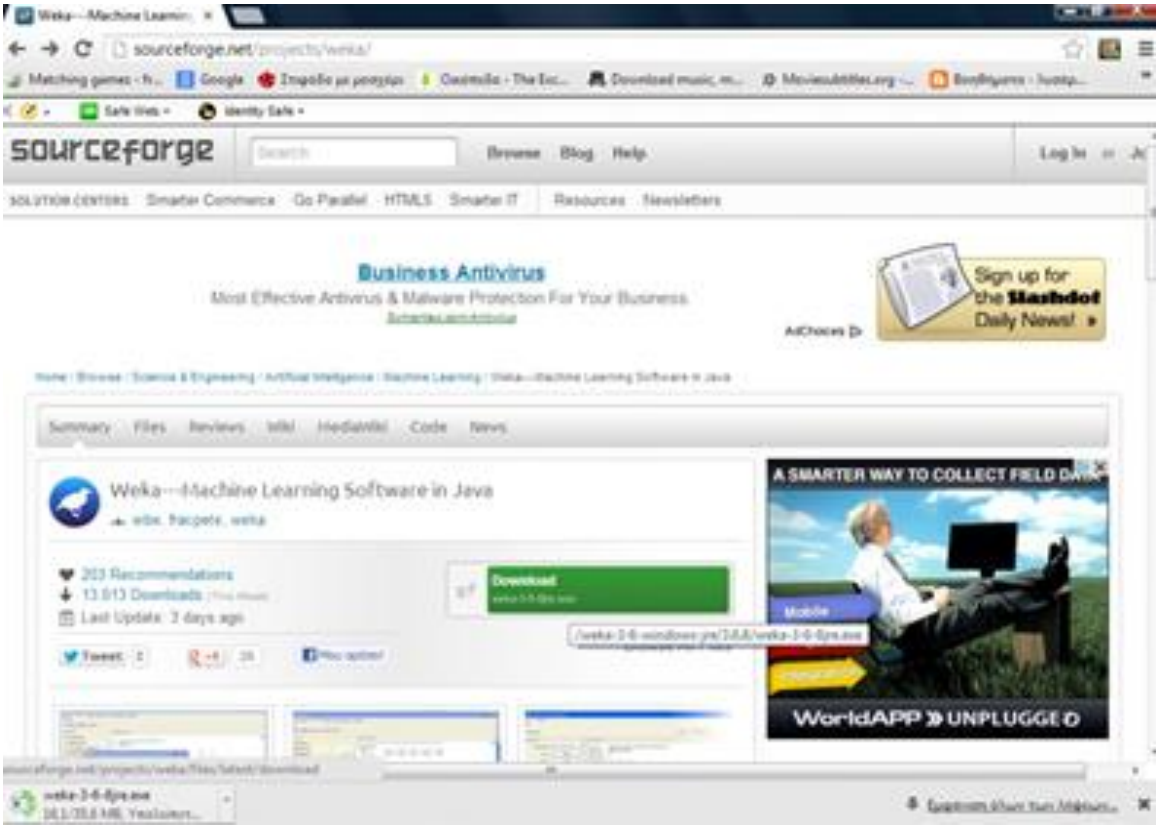
http://www.cs.waikato.ac.nz/ml/weka/

Ακόμα υπάρχει η δυνατότητα επιλογής της κατάλληλης version ανάλογα με το περιβάλλον των windows (XP,Vista,windows 7), σε περίπτωση που κάποια έκδοση της γλώσσας

προγραμματισμού java δεν είναι ήδη εγκατεστημένη, το εκτελέσιμο που θα αποθηκευτεί είναι η έκδοση (stable version) που περιλαμβάνει την java VM 1.6.

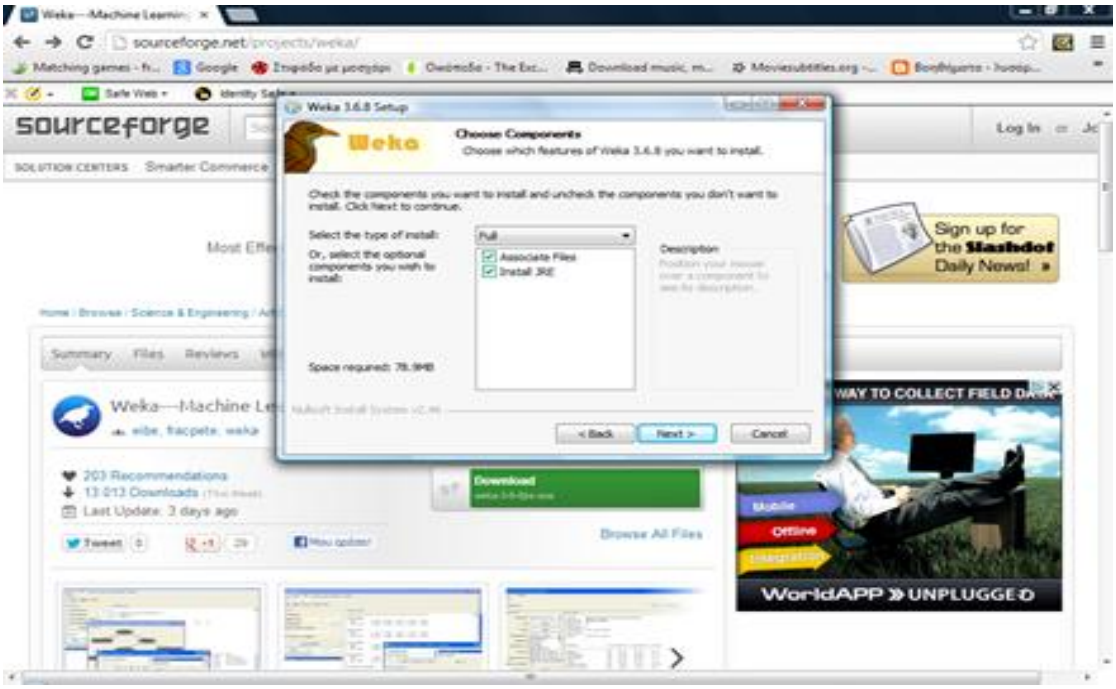
Πιο αναλυτικά για την εγκατάσταση του weka, το οποίο διανέμεται εύκολα και δωρεάν στο internet και είναι εύκολο στην εύρεση του θα αναφερθούμε εκτενέστερα παρακάτω:

Πρώτο βήμα : Πατάμε στο *download* για να κατέβει το πρόγραμμα και να αποθηκευθεί στο σκληρό δίσκο του υπολογιστή , όπως φαίνεται στην επόμενη εικόνα.

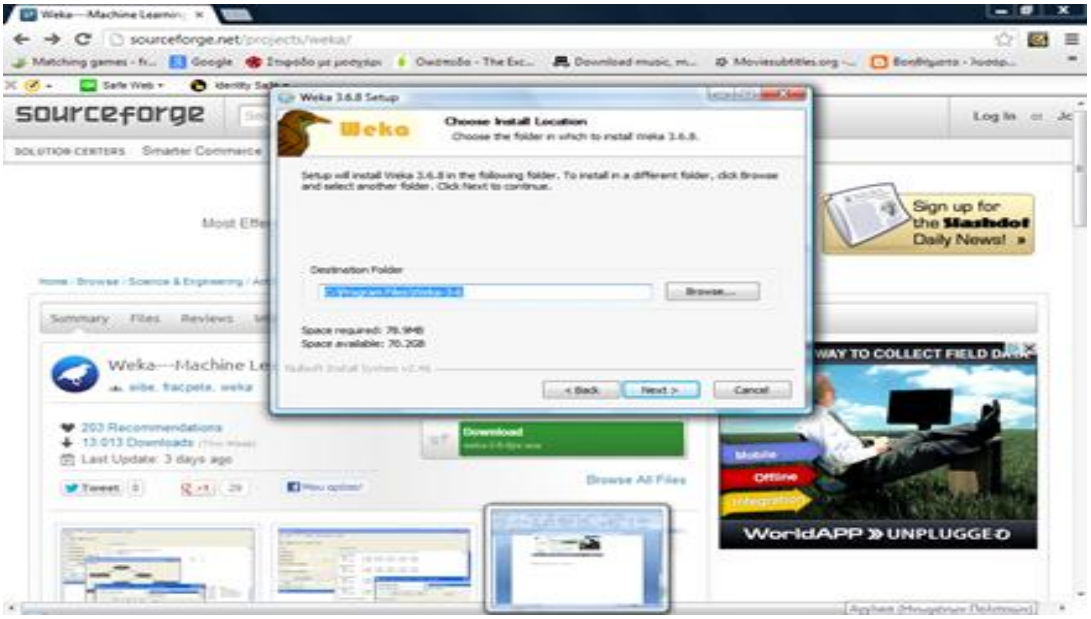


Δεύτερο βήμα : Αποδεχόμαστε τους όρους του προγράμματος, πατώντας το κουμπί *I agree*, όπως φαίνεται στην επόμενη εικόνα.

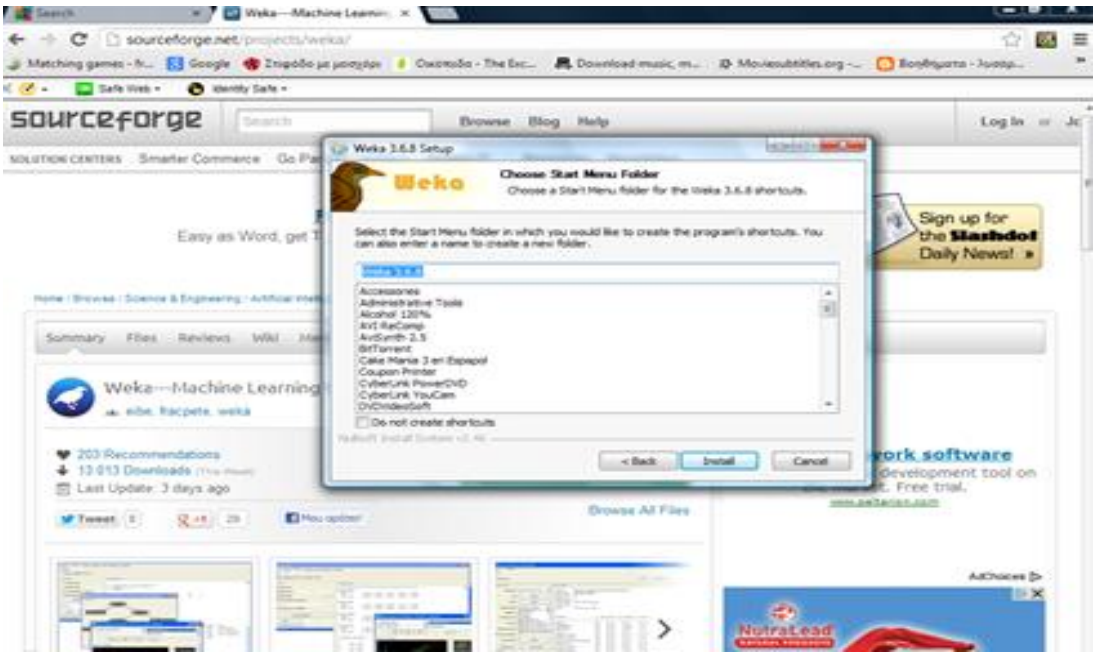
Τρίτο βήμα : Πατάμε το κουμπί *επόμενο* (next) για να συνεχίσει η εγκατάσταση του προγράμματος, όπως φαίνετε στην εικόνα.



Τέταρτο βήμα : Εν συνεχεία επιλέγεται η θέση που θα αποθηκευτεί το πρόγραμμα στο σκληρό δίσκο, πατάμε *επόμενο* (next) όπως φαίνετε στην παρακάτω εικόνα.

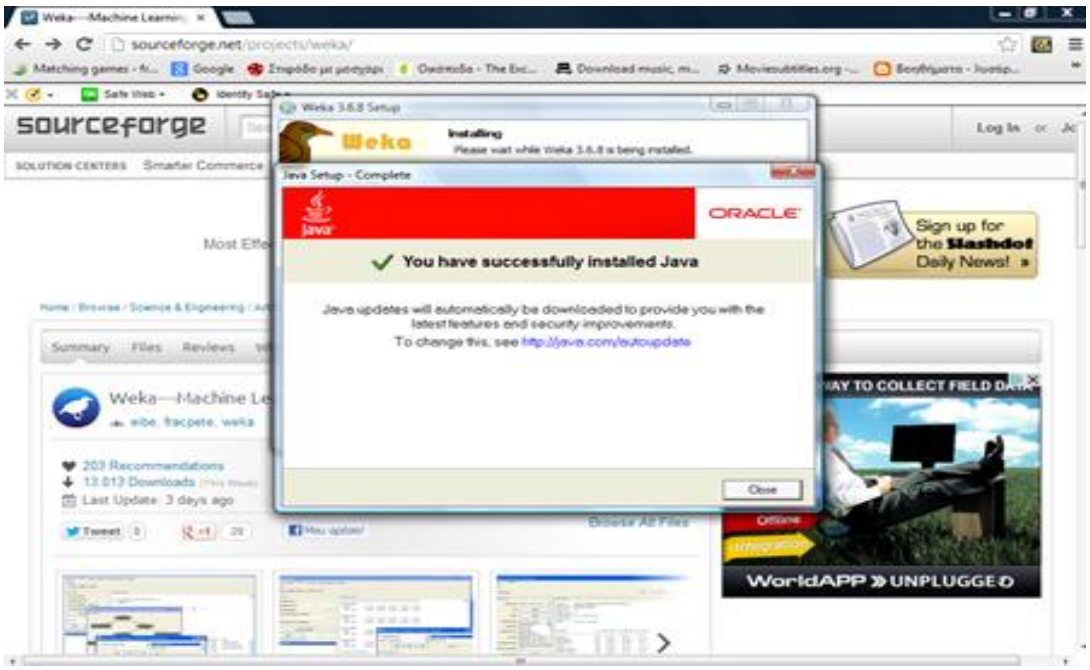


Πέμπτο βήμα : επιλέγουμε το κουμπί *εγκατάσταση* (install) για να προχωρήσει η εγκατάσταση του προγράμματος.

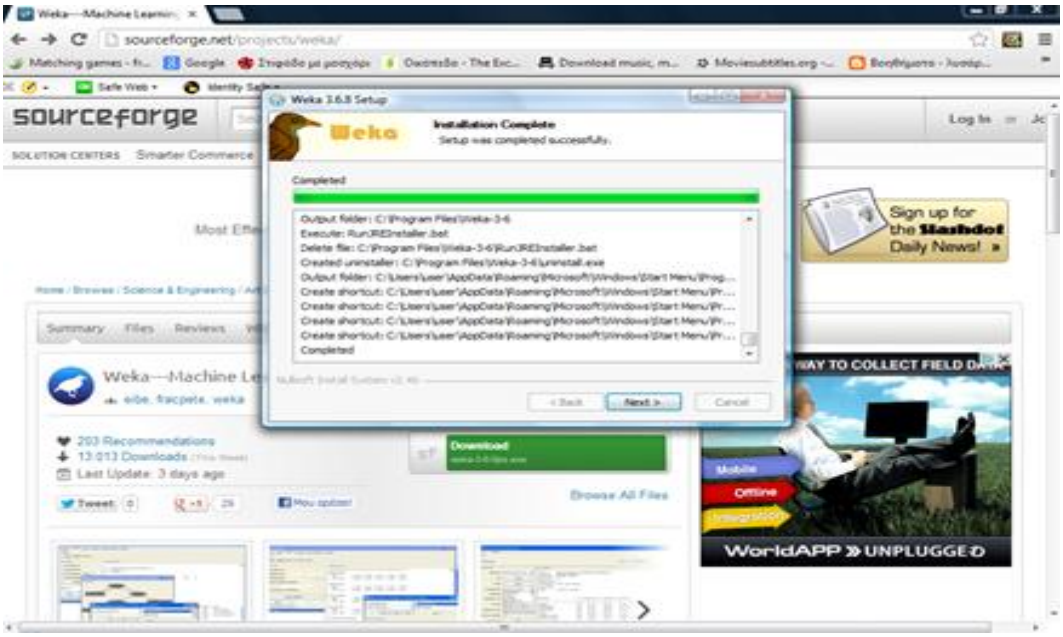


Έκτο βήμα : Στη συνέχεια αυτόματα εμφανίζεται η βάση δεδομένων που τρέχει το πρόγραμμα , επιλέγουμε *εγκατάσταση* (install) όπως φαίνεται παρακάτω.

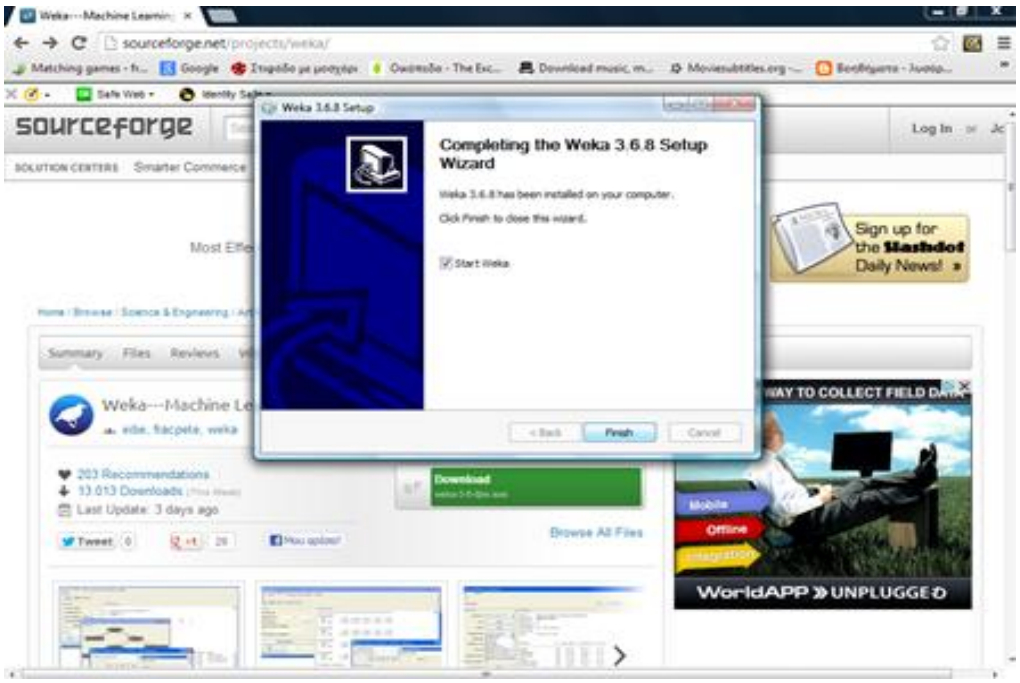
Έβδομο βήμα : Αφού εγκατασταθεί το πρόγραμμα επιτυχώς εμφανίζεται το παρακάτω παράθυρο και επιλέγουμε το κουμπί *κλείσιμο* (Close).



Όγδοο βήμα : Επιλέγουμε το κουμπί *επόμενο* (next) για να συνεχιστεί η εγκατάσταση του weka ,όπως φαίνετε στην εικόνα.



Ένατο βήμα : Στη συνέχεια το πρόγραμμα έχει εγκατασταθεί με επιτυχία και επιλέγουμε το κουμπί *τέλος* (finish) , όπως φαίνεται στην επόμενη εικόνα.



3.3 Δημιουργία νέου

Όλα τα αρχεία τα οποία δέχεται το πρόγραμμα Weka έχουν κατάληξη ARFF (Attribute-Relation File Format), είναι ένα αρχείο κειμένου χαρακτήρων ASCII (ASCII text file) που περιλαμβάνει ένα σύνολο από παραδείγματα (instances) τα οποία έχουν χαρακτηριστικά (attributes) που περιγράφουν τα παραπάνω σύνολα.

Αφού δηλωθεί το όνομα «attribute-name» ακολουθεί η επιλογή του τύπου του χαρακτηριστικού, το όρισμα δηλαδή <*datatype*>. Το weka έχει την δυνατότητα υποστήριξης των παρακάτω τύπων:

- Αριθμητικοί τύποι- Numeric attributes: αριθμητικά δεδομένα δηλαδή πραγματικοί και ακέραιοι αριθμοί, όπως ηλικία, θερμοκρασία, μισθοί κ.τ.λ.
- Ονομαστικά δεδομένα- Nominal attributes: Δεδομένα που ορίζουν την κάθε κατηγορία και τα οποία θα πρέπει να δηλώνονται μέσα σε αγκύλες, όπως για παράδειγμα: {<nominal-name1>, <nominal-name2>, <nominal-name3>, ...}
- Αλφαριθμητικά- (string): χρησιμοποιώντας αυτά τα χαρακτηριστικά υπάρχει δυνατότητα δημιουργίας δομών της μορφής text-mining applications.
- Ημερομηνίες: οι ημερομηνίες καθορίζονται με συγκεκριμένο format, όπως φαίνεται παρακάτω:
  "yyyy-MM-dd'T'HH:mm:ss" (ISO-8601)

Επόμενο βήμα είναι η δήλωση των δεδομένων που θα εισάγει ο χρήστης.

Κάθε γραμμή είναι και ένα διαφορετικό παράδειγμα ενώ οι τιμές των χαρακτηριστικών χωρίζονται μεταξύ τους με ένα κόμμα.

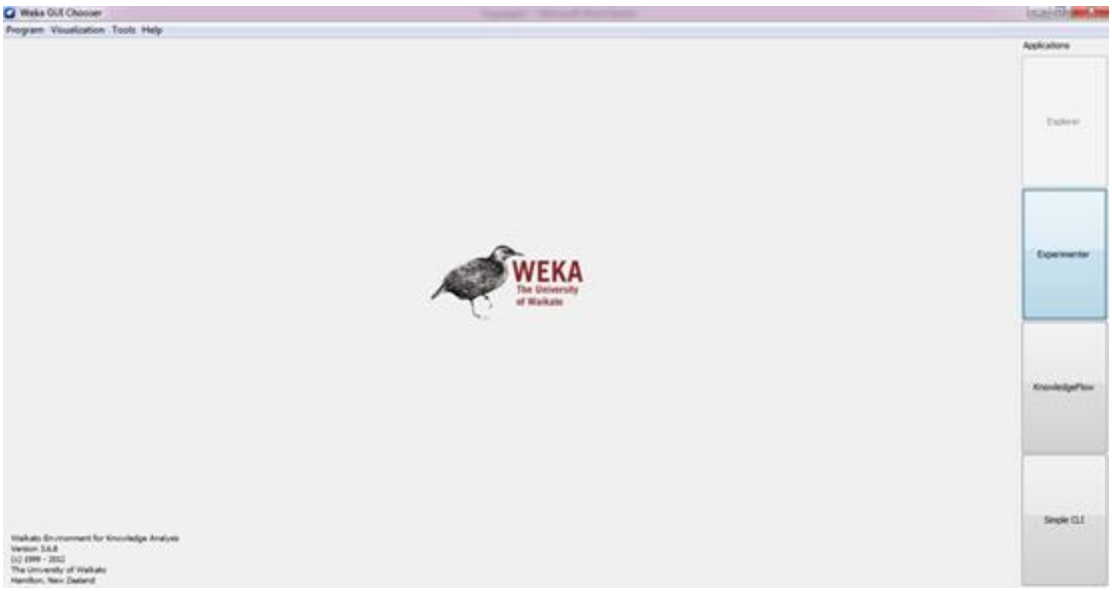
3.4 Η πρώτη επαφή με το weka

Ανοίγοντας το πρόγραμμα weka εμφανίζεται το παρακάτω παράθυρο, στο οποίο ο χρήστης καλείται να επιλέξει ένα από τα τέσσερα περιβάλλοντα εργασίας:

- Explorer
- Experimenter
- KnowledgeFlow
- Simple CLI

Τα οποία θα αναλύσουμε λεπτομερώς παρακάτω.

Ανοίγοντας το πρόγραμμα weka εμφανίζεται το παρακάτω παράθυρο (Εικόνα 4) .



«Εικόνα 4»

3.4.1 Explorer

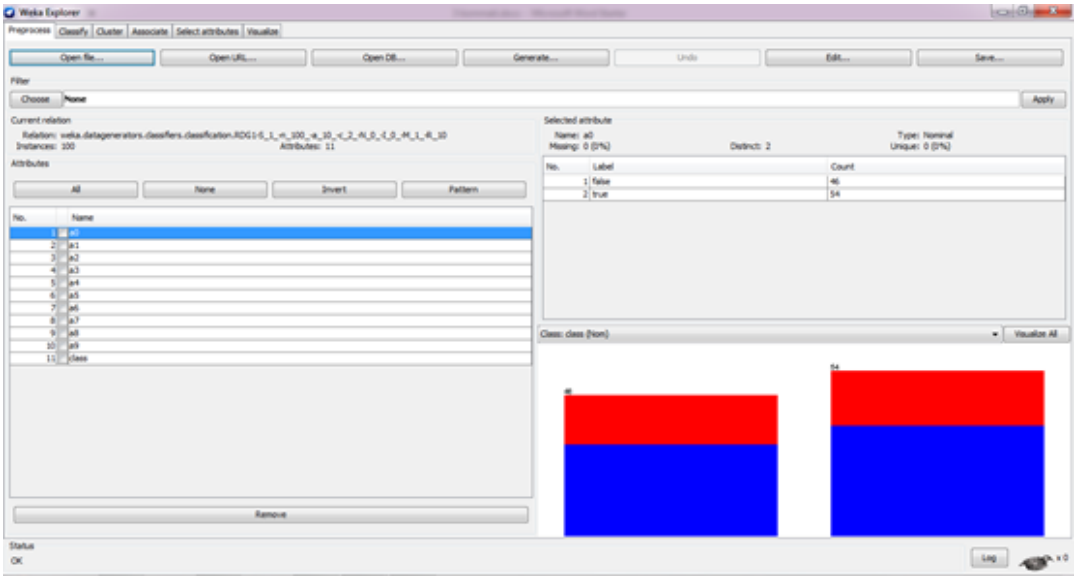
Ένας εύκολος και κατανοητός τρόπος με τον οποίο ο χρήστης μπορεί να χρησιμοποιήσει στο weka, αφού οι δυνατότητες που παρέχει είναι οργανωμένες σε λίστες και ο χρήστης απλά μαρκάροντας τις επιλογές που θέλει βγάζει και τα αντίστοιχα αποτελέσματα. Μέσω του μενού Application →Exporer→Open file υπάρχει δυνατότητα ο χρήστης να επιλέξει ένα σύνολο δεδομένων στο οποίο μπορεί να επιλέξει μία από τις έξι παρακάτω τεχνικές:

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

Τις οποίες θα αναλύσουμε ξεχωριστά στις επόμενες παραγράφους .

3.4.1.1 Preprocess

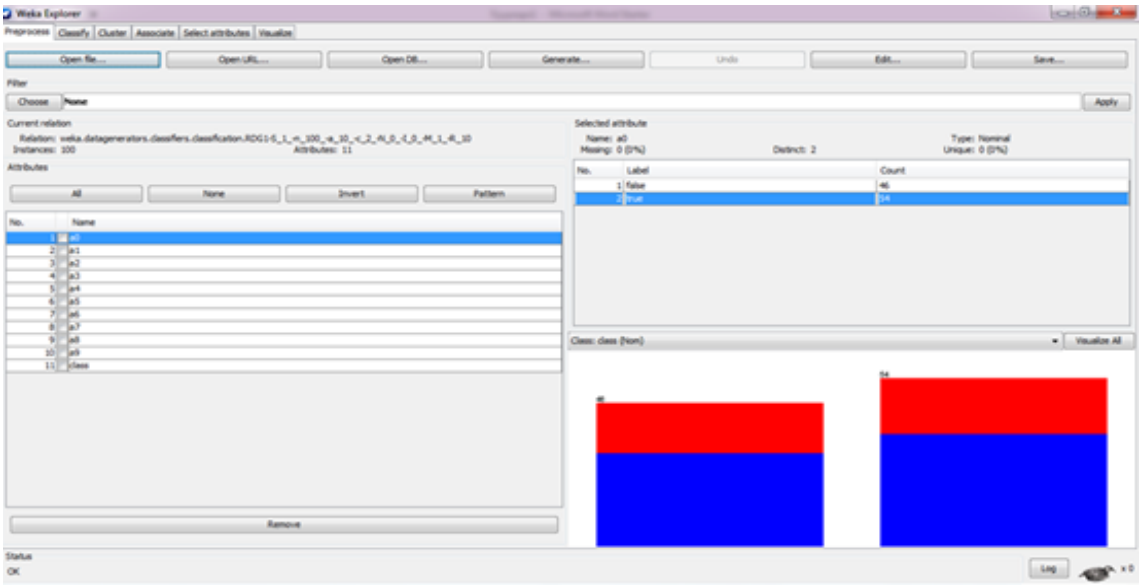
Μια τεχνική με την οποία μπορεί ο χρήστης να τροποποίηση τα δεδομένα με διάφορα εργαλεία και αλγόριθμους που περιέχονται στο Preprocess.



«Εικόνα 5»

Έχοντας επιλέξει ο χρήστης ένα σύνολο δεδομένων (αρχείο .arff), θα εμφανιστούν γραφικά τα δεδομένα για καθένα από τα γνωρίσματα ξεχωριστά όπως επίσης και στατιστικές πληροφορίες για αυτά. Εάν τα δεδομένα ανήκουν στην ίδια κλάση θα εμφανίζονται με το ίδιο χρώμα.

Αφού ο χρήστης επιλέξει ένα παράδειγμα από τα υπάρχοντα αρχεία και το τρέξει θα έχει σαν αποτέλεσμα την παρακάτω εικόνα:

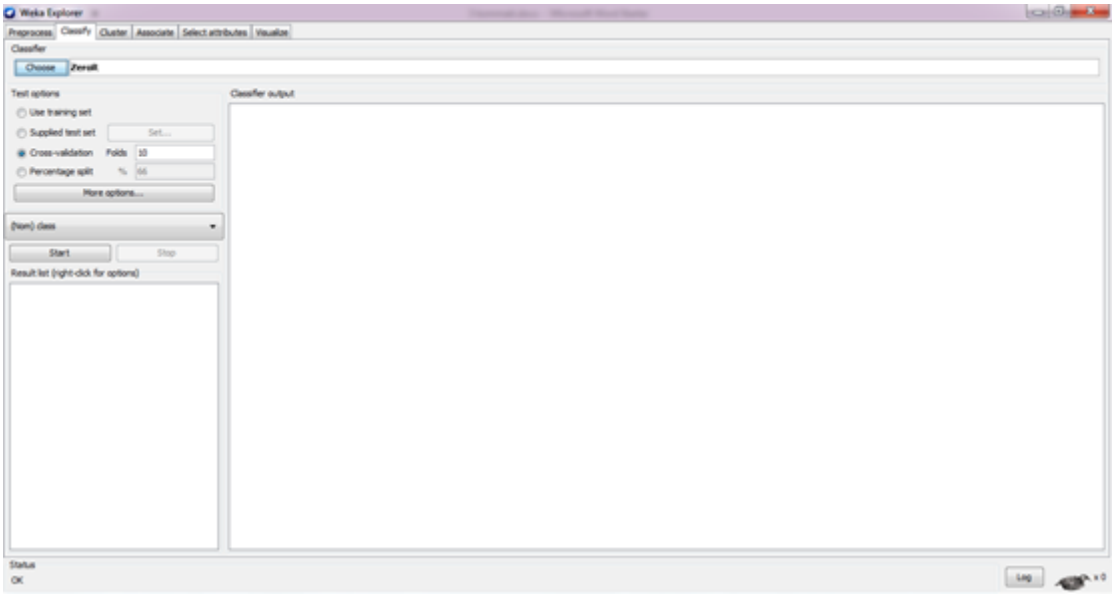


«Εικόνα 6»

3.4.1.2 Classify

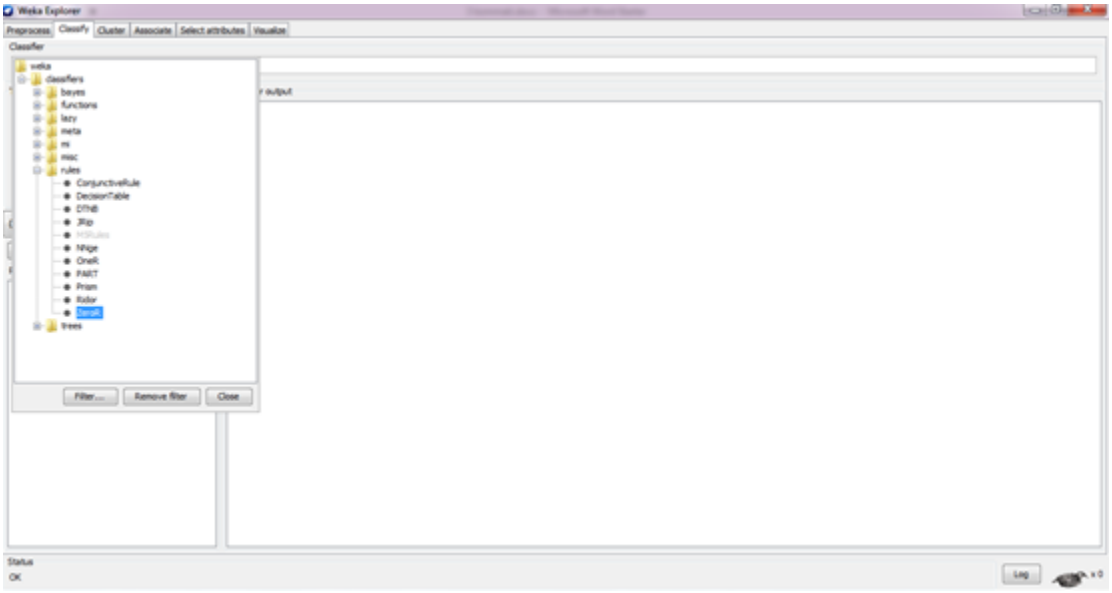
Μια τεχνική που χρησιμοποιείτε για ταξινόμηση με τους αντίστοιχους αλγόριθμους ταξινόμησης, ή αλγόριθμους για παλινδρόμηση.

Επιλέγοντας classify choose θα γίνει η επιλογή παραδείγματος από την βιβλιοθήκη του προγράμματος.



«Εικόνα 7»

Στο αριστερό μέρος όπως φαίνεται παρακάτω εμφανίζεται το dataset που περιλαμβάνει κάποια παραδείγματα που μπορεί να επιλέξει ο χρήστης.

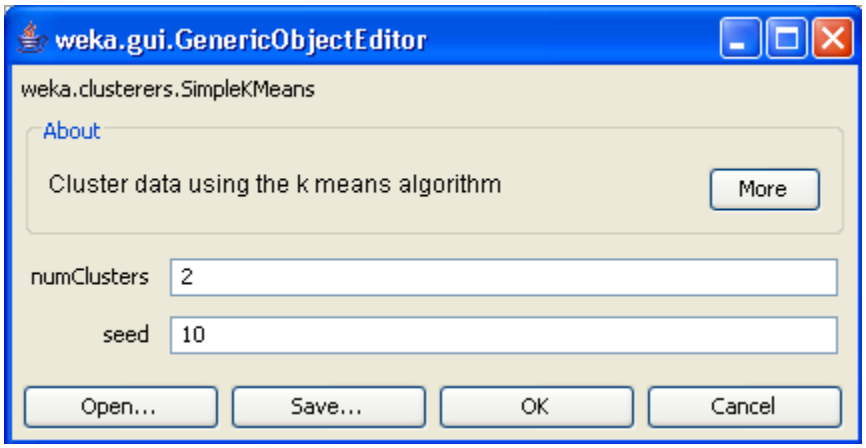


«Εικόνα 8»

3.4.1.3 Clustering- Εφαρμογή ομαδοποίησης

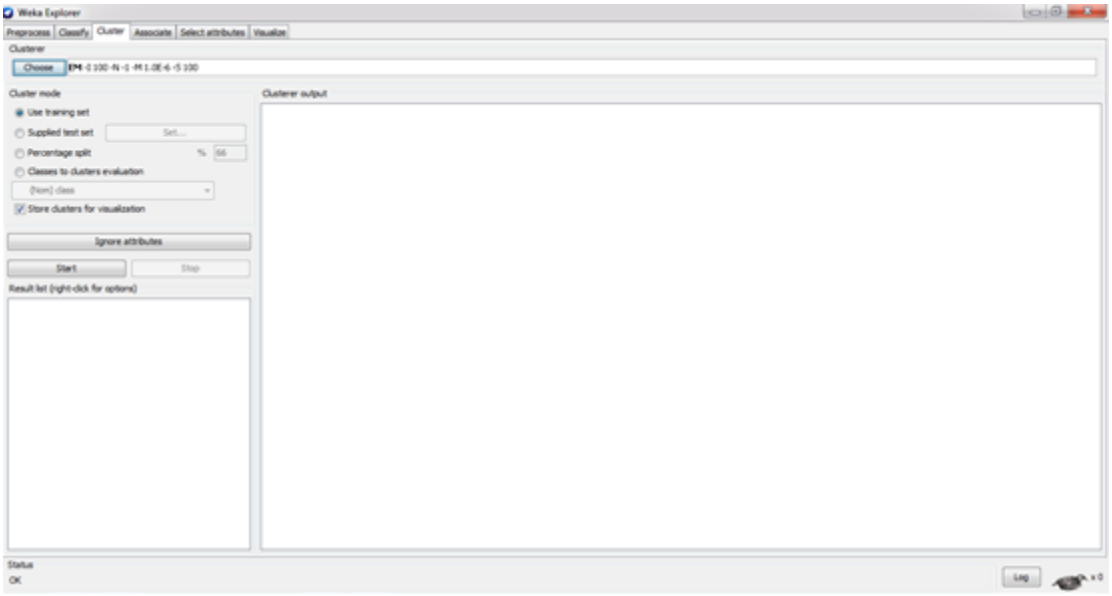
Στην εφαρμογή αυτή ο χρήστης έχει την επιλογή από διάφορους αλγόριθμους κατηγοριοποίησης, όπως είναι ο K- modes, K- means, SOM , RBF να επιλέξει τον αλγόριθμο που θέλει ώστε να τους χρησιμοποιήσει για νε βρει υποομάδες μέσα από τα σύνολα δεδομένων.

Στη περίπτωση αυτή επιλεγεί τον αλγόριθμο K-means, ενώ θα εμφανιστεί το παρακάτω παράθυρο έτσι ώστε ο χρήστης να καθορίσει τον αριθμό των συστάδων (K) .



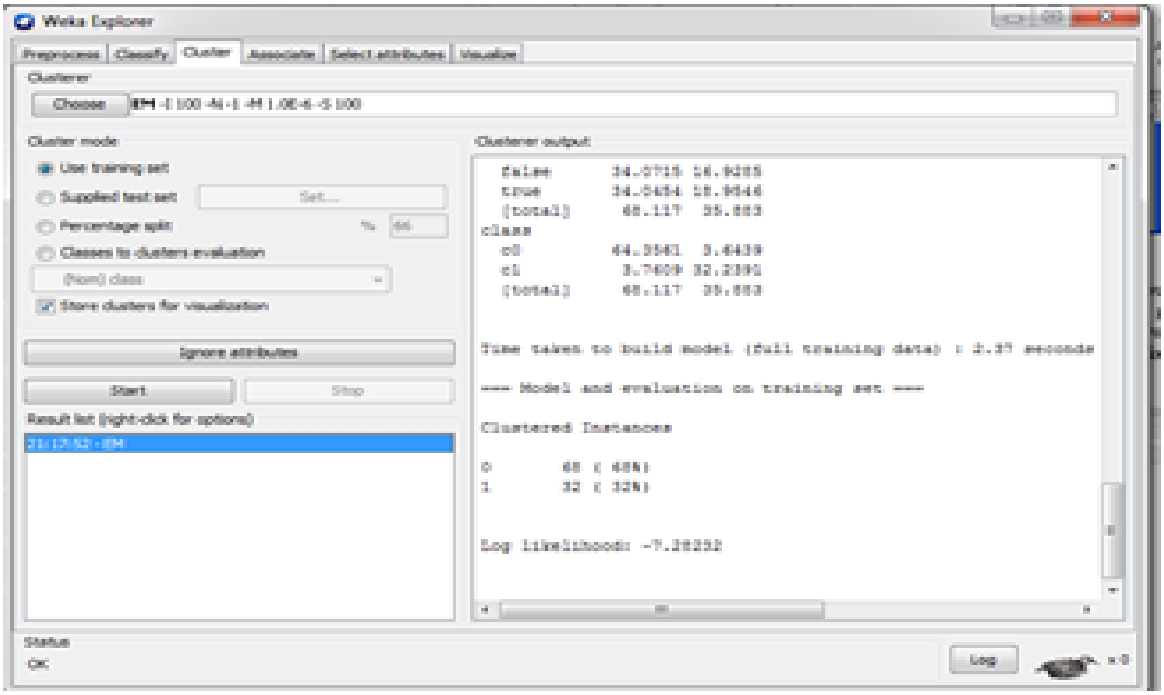
«Εικόνα 9»

Πιο αναλυτικά ο K- means είναι ένας αλγόριθμος που βασίζεται σε αριθμητικά δεδομένα, τα οποία χωρίζονται σε k-ομάδες με βάση όμοια χαρακτηριστικά. Η λειτουργία του αλγόριθμου, όπως και του προηγούμενου είναι επαναληπτική. Τα αποτελέσματα είναι σε μορφή πινάκων οπού κάθε ομάδα είναι και μια εγγραφή.



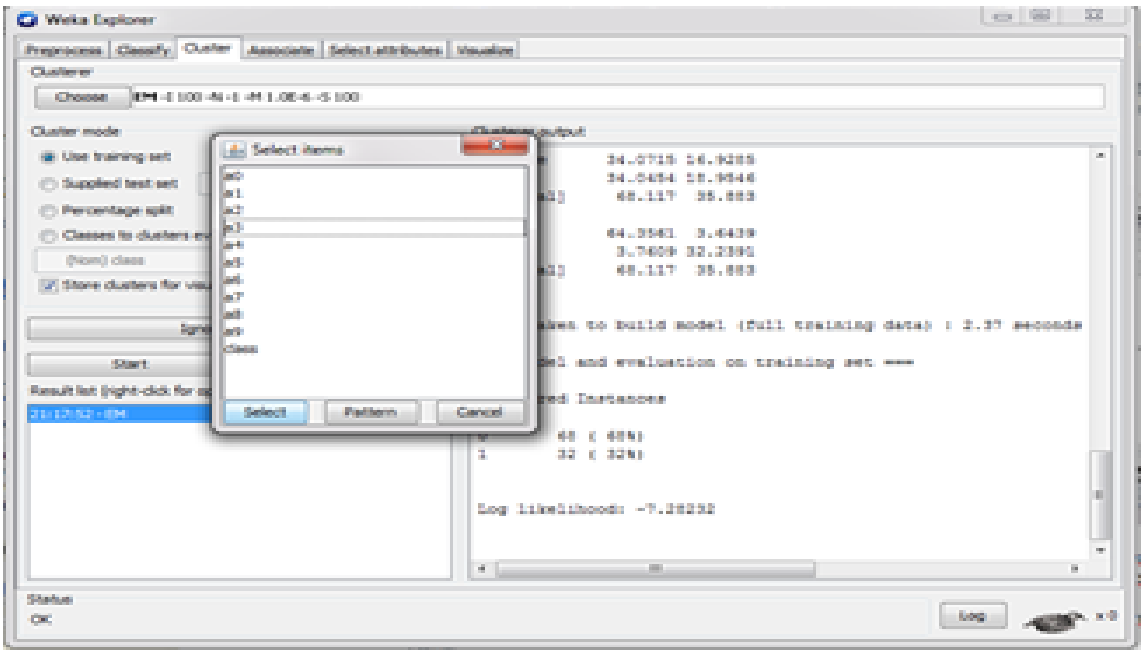
«Εικόνα 10»

Στη περίπτωση όπου έχουμε δεδομένα που ανήκουν σε μια κατηγορία (class) κατά την εκτέλεση του αλγορίθμου ο χρήστης θα πρέπει να έχει επιλέξει πρώτα το “ignore attributes” .



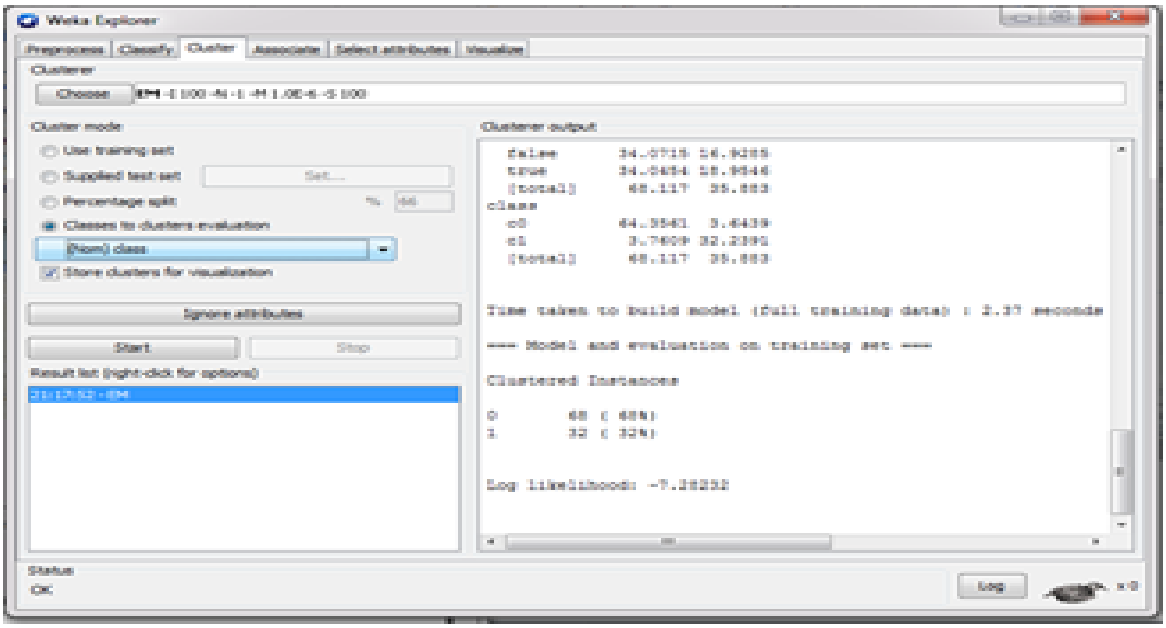
«Εικόνα 11»

Στο παράθυρο που θα ανοίξει θα πρέπει να επιλέξει class και στη συνεχεία να πατήσει το κουμπί select:



«Εικόνα 12»

Βέβαια επειδή στη διαδικασία συσταδοποίησης δεν είναι γνωστό εκ των πρότερων σε ποια κατηγορία ανήκει το κάθε παράδειγμα , όταν γίνεται διαθέσιμη όμως , όπως σε αυτή τη περίπτωση η συσχέτιση των συστάδων με τις κατηγορίες του προβλήματος μπορεί να παρέχει χρήσιμη πληροφόρηση στο χρήση για τη ποιότητα της διαδικασίας αυτής .

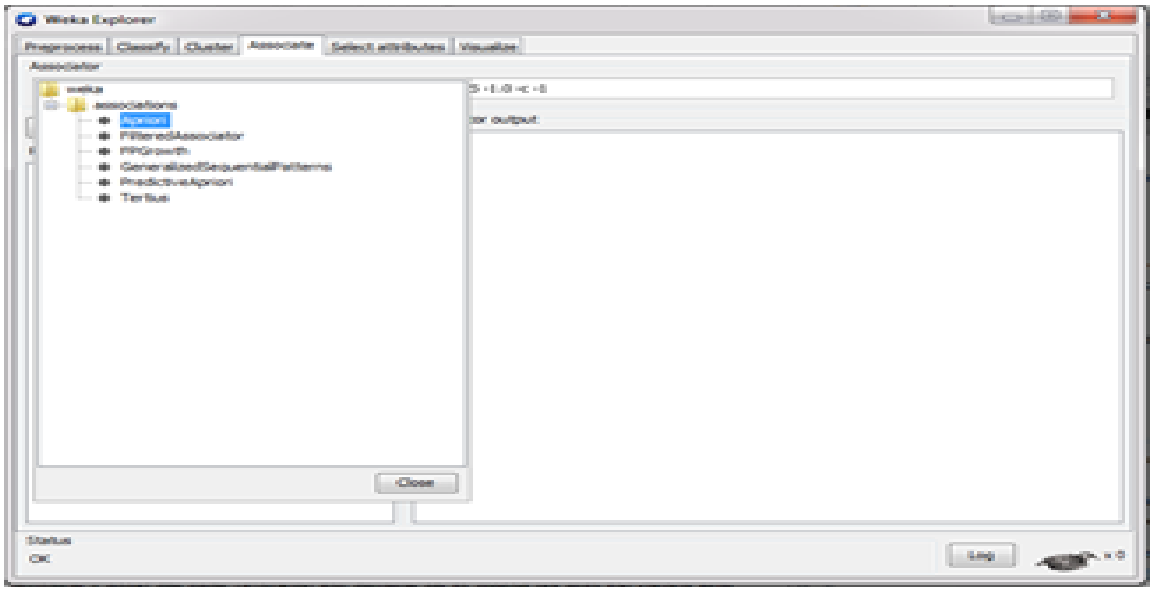


«Εικόνα 13»

3.4.1.4 Association- Εφαρμογή συσχέτισης

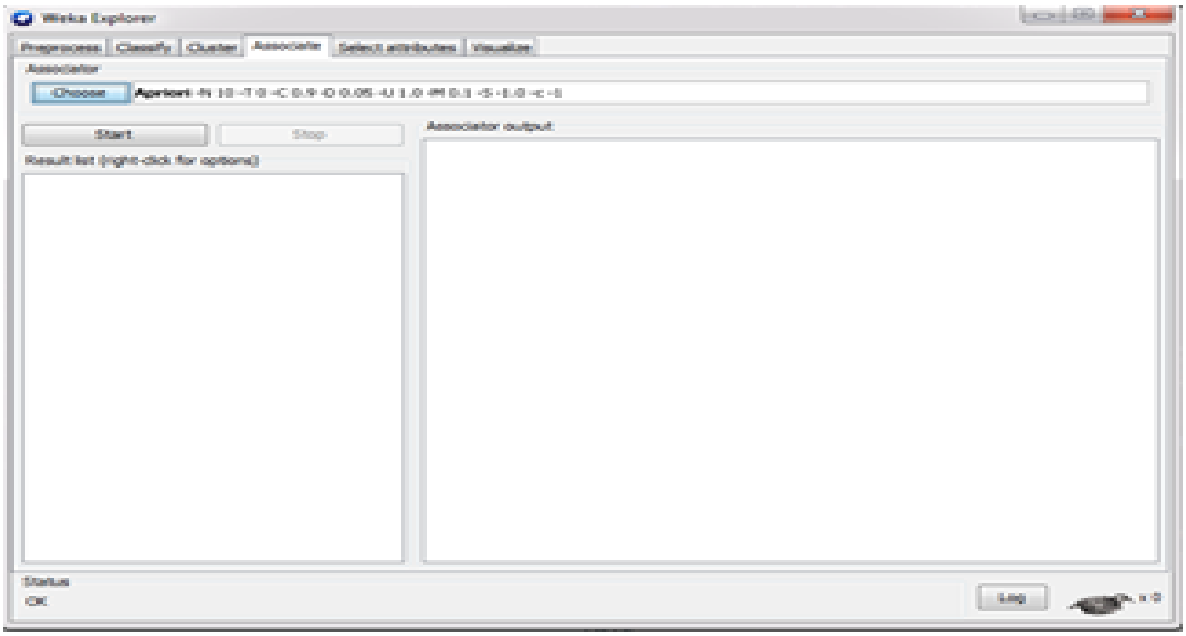
Η λειτουργία αυτή περιέχει αλγόριθμους για την εύρεση κανόνων συσχέτισης. Ένας γνωστός αλγόριθμος εξαγωγής κανόνων συσχέτισης από ένα σύνολο δεδομένων και εγγραφών όπως έχουμε δει παραπάνω σε προηγούμενα κεφάλαια είναι ο αλγόριθμος apriori.

Στην καρτέλα associate ο χρήστης πατώντας choose επιλέγει τον αλγόριθμο Apriori, όπως φαίνεται παρακάτω στη εικόνα:



«Εικόνα 14»

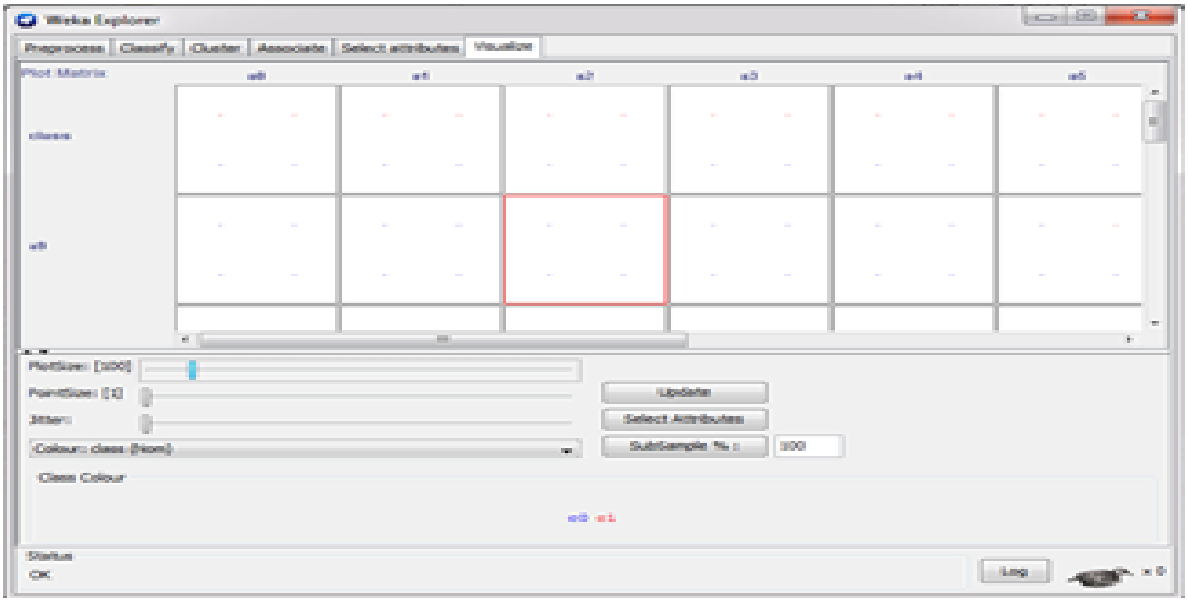
Στην συνέχεια αφού ο χρήστης διατηρήσει τις παραμέτρους default ίδιες πατάει το START για να εμφανιστούν τα αποτελέσματα:



«Εικόνα 15»

3.4.1.5 Visualize

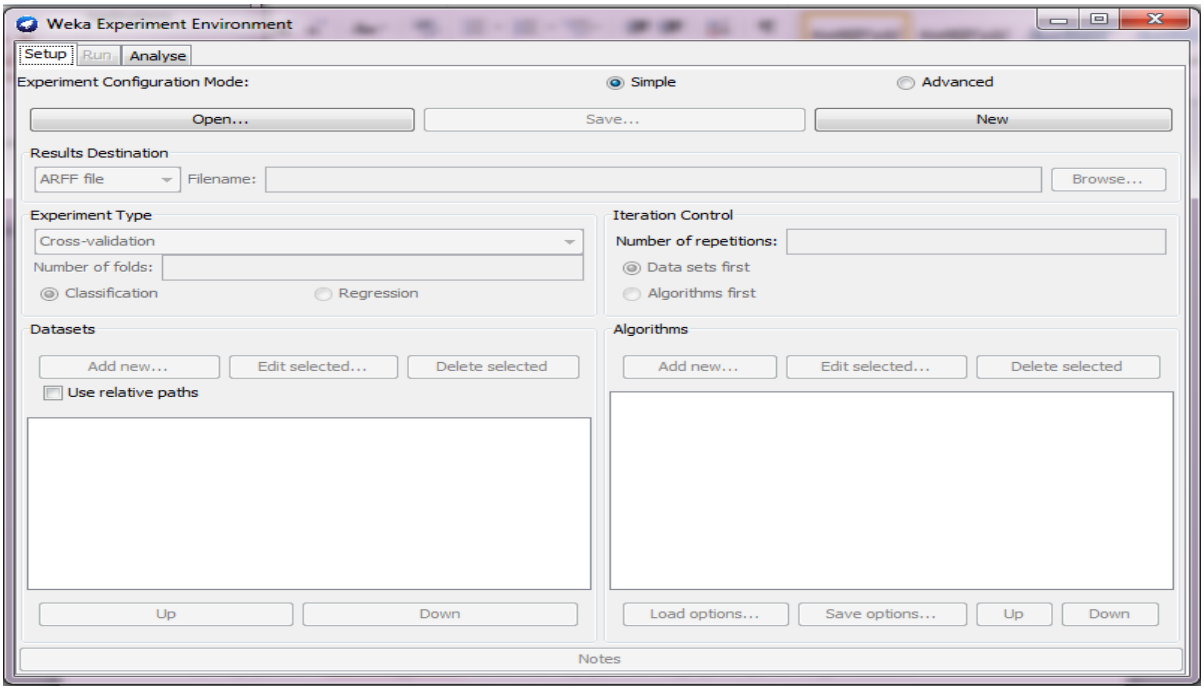
Η τελευταία λειτουργία που περιέχεται στο explorer και παρέχει εργαλεία για την οπτικοποίηση των δεδομένων και έχουμε ην παραγωγή δισδιάστατα γραφήματα.



«Εικόνα 16»

3.4.2 Experimenter

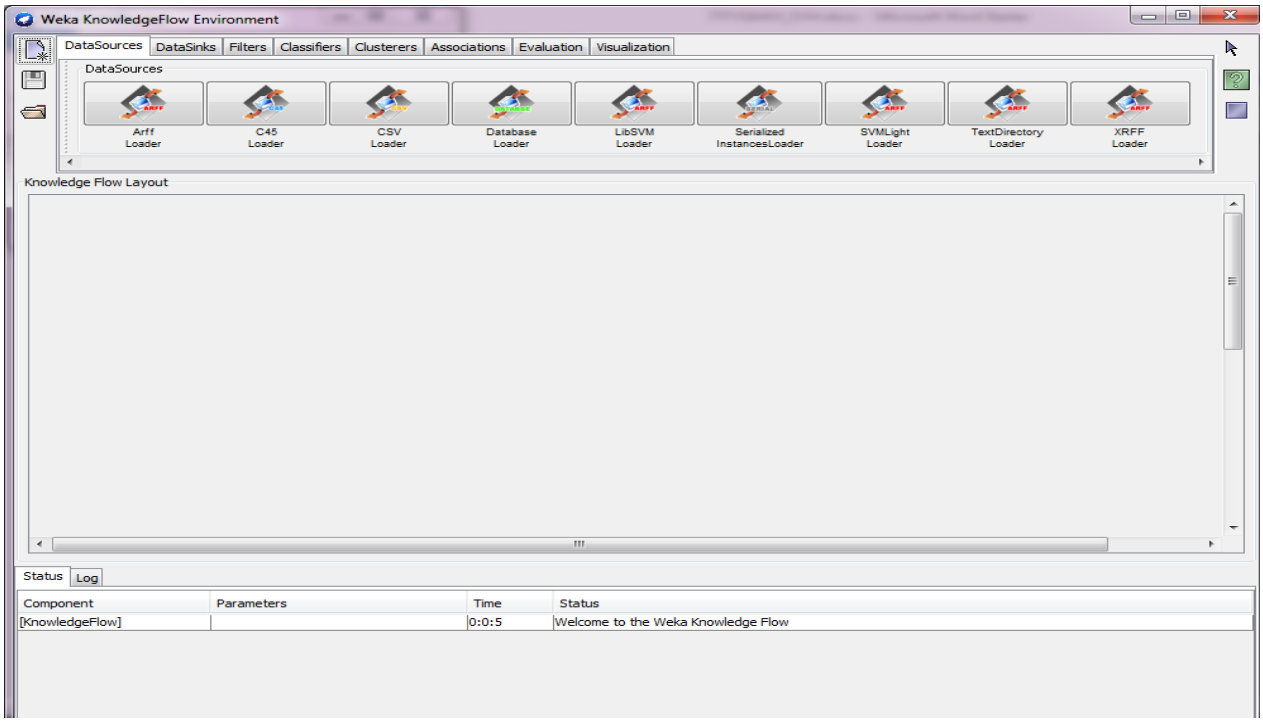
Το περιβάλλον Experimenter βοηθάει τον χρήστη να διαπιστώσει ποια είναι η απόδοση των μαθηματικών σχημάτων διάφορων σετ δεδομένων. Είναι αυτόματο και οι πληροφορίες που αποθηκεύονται μπορούν αν επιθυμεί ο χρήστης να τις ανατρέξει για περισσότερη μελέτη και εξόρυξη γνώσης. Υπάρχουν χρονικοί περιορισμοί σχετικά με το μέγεθος των δεδομένων και οι χρήστες πρέπει να είναι προχωρημένοι ώστε να μπορούν να μοιράσουν σωστά σε διαφορετικούς υπολογιστές ολόκληρο το φορτίο των σετ δεδομένων.



«Εικόνα 17»

3.4.3 KnowledgeFlow

Το περιβάλλον KnowledgeFlow είναι ένα περιβάλλον παρόμοιο με το Experimenter. Σε αυτό το περιβάλλον ο χρήστης μπορεί αν θέλει να δει πως «κινούνται» τα δεδομένα και οι πληροφορίες μέσα στο σύστημα. Έτσι μπορεί να έχει επίγνωση του τρόπου που γίνεται η ανάλυση και η επεξεργασία των δεδομένων και να φτιάξει ένα πίνακα στον οποίο θα μπορεί να συνθέτει ένα γράφημα. Το περιβάλλον KnowledgeFlow είναι πιο ευέλικτο, αφού ο χρήστης μπορεί να δει λεπτομερώς όλη την διαδικασία και όχι μόνο το αποτέλεσμα, όπως είναι στο περιβάλλον Explorer. Μια διαφορά με το experimenter περιβάλλον είναι ότι σε αυτό το περιβάλλον δεν υπάρχουν περιορισμοί σχετικά με το μέγεθος των αρχείων που εξετάζονται του κάθε υποδείγματος από τα σετ δεδομένων.



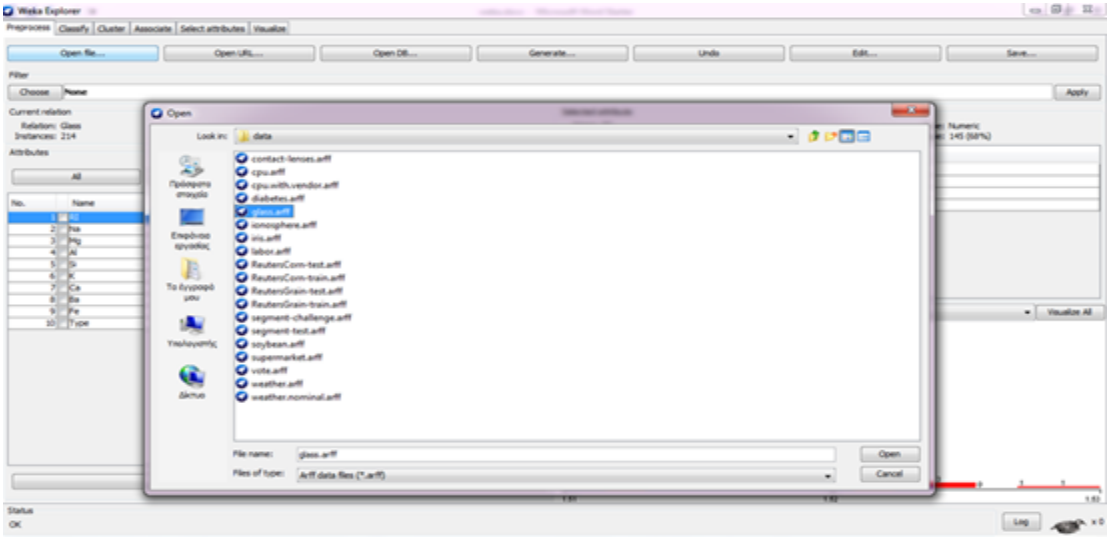
«Εικόνα 18»

3.4.4 Simple CLI

Το τελευταίο περιβάλλον εργασίας που μπορεί ο χρήστης να δει στο πρόγραμμα weka είναι το λεγόμενο περιβάλλον Command Line Interface. Είναι το πιο απλό περιβάλλον που υπάρχει στο weka και αυτό γιατί δεν αποτελείται από γραφήματα αλλά είναι ένας κενός χώρος με μια γραμμή στο κάτω μέρος για να εισάγει ο χρήστης εντολές. Ο χρήστης πρέπει να γνωρίζει εις βάθος τόσο το weka όσο και τις εντολές του.

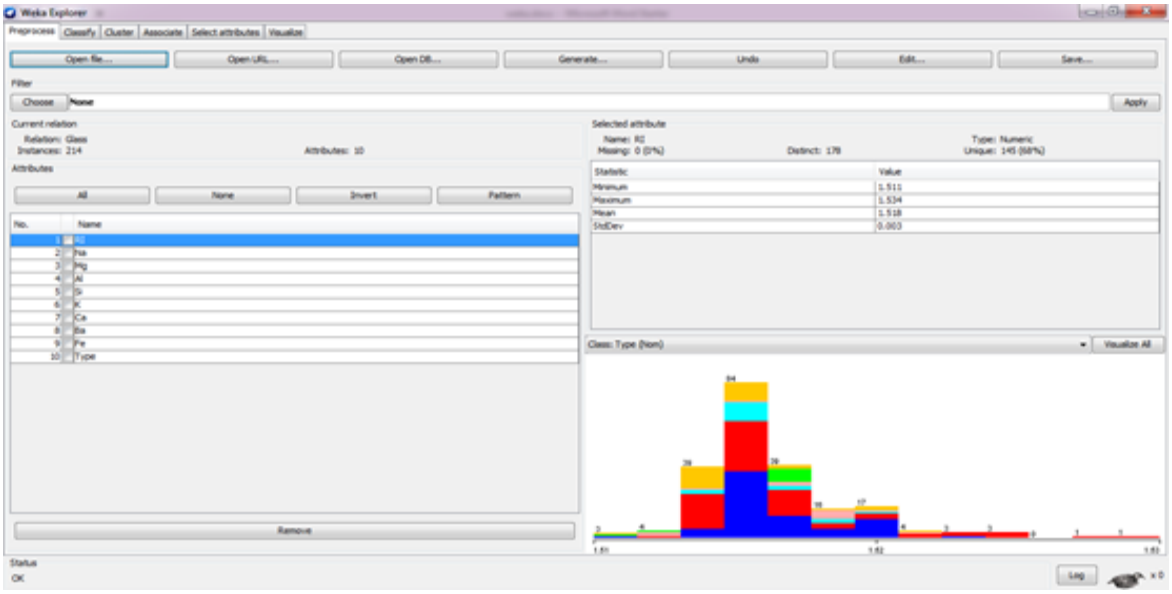
3.5 Εφαρμογή στο Weka

Το weka έχει ενσωματωμένα έτοιμες βάσεις δεδομένων και ο χρήστης πρέπει απλά να επιλέξει μια από την βάση που αποτελείται από 150 στοιχεία και 5 χαρακτηριστικά (attributes).



«Εικόνα 19»

Αφού γίνει η επιλογή των δεδομένων, μέσω του μενού Application→Explorer→Open File , εμφανίζονται τα παρακάτω στατιστικά δεδομένα, όπως φαίνεται στην εικόνα 19. Τα δεδομένα που βρίσκονται στην ίδια κλάση ταξινομούνται με το ίδιο χρώμα.

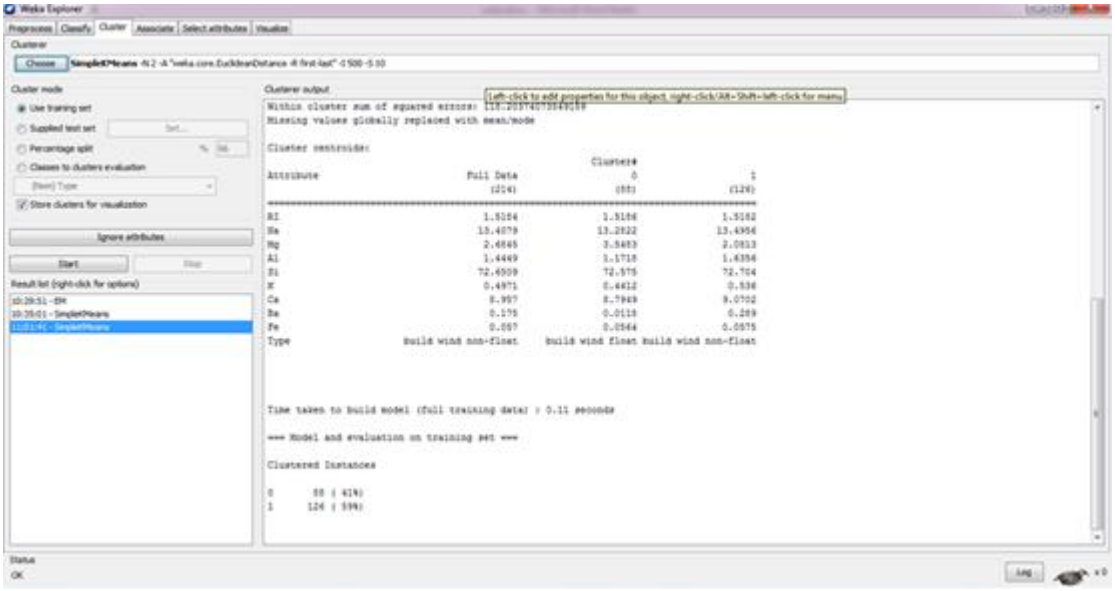


«Εικόνα 20»

Στην καρτέλα cluster έχουμε την ομαδοποίηση των δεδομένων και γίνεται η συσταδοποίηση των δεδομένων επιλέγοντας τον αλγόριθμο συσταδοποίησης “Simple K means” , παράγοντας τα εξής χαρακτηριστικά:

- Minimum: η ελάχιστη τιμή που εμφανίζεται στο σύνολο δεδομένων
- Maximum: οπού είναι η μέγιστη τιμή που έχει εντοπίσει το πρόγραμμα στο σύνολο δεδομένων.
- Mean: η μέση τιμή του συνόλου των δεδομένων.
- StdDev: η τυπική απόκλιση του συνόλου των δεδομένων.

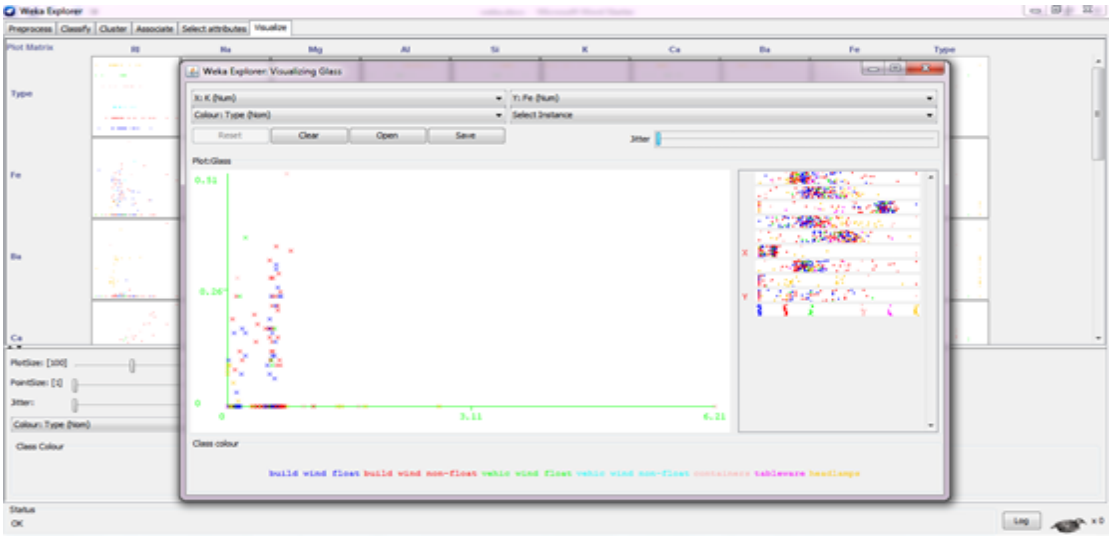
Αφήνοντας την επιλογή «use training» και πατώντας «start» θα ξεκινήσει η διαδικασία όπως φαίνεται στην παρακάτω εικόνα 21.



«Εικόνα 21»

Στη συνέχεια θα αναλυθεί η καρτέλα visualize όπου εμφανίζεται η γραφική αναπαράσταση των γνωρισμάτων (στη προκειμένη περίπτωση των “K” και “Fe’’)από τη λίστα των γνωρισμάτων. Από την καρτέλα visualize υπάρχει η δυνατότητα να εμφανιστεί η γραφική αναπαράσταση κάθε γνωρίσματος σε συνάρτηση με κάθε άλλο γνώρισμα, όπως φαίνεται στη

παρακάτω εικόνα όπου αναπαρίσταται γραφικά η συσχέτιση των γνωρισμάτων “K” και “Fe” (Εικόνα 22) .



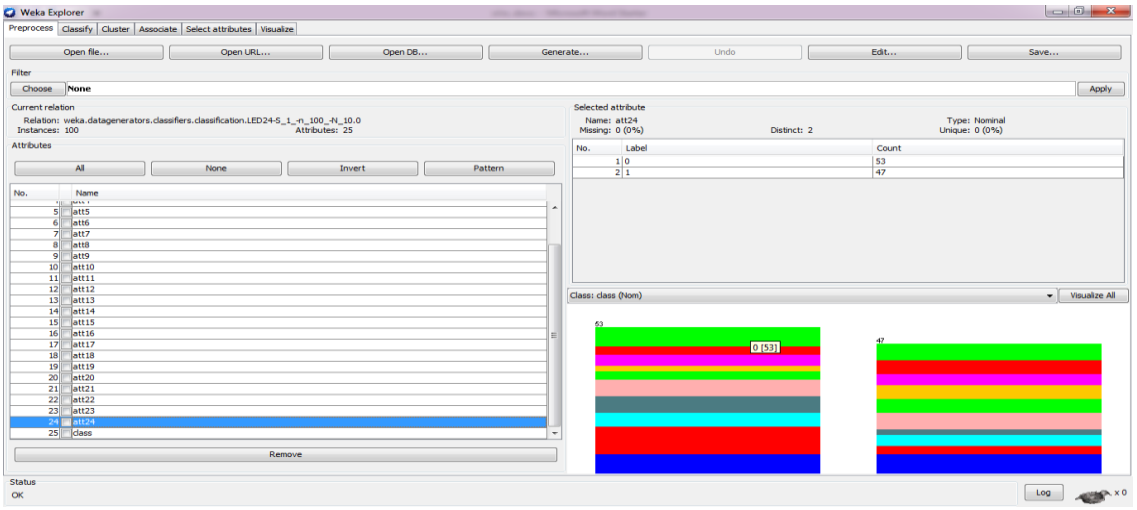
«Εικόνα 22»

Στην συνέχεια θα τρέξουμε κάποια παραδείγματα για να γίνει ακόμα πιο κατανοητό ο τρόπος που εφαρμόζεται το πρόγραμμα Weka.

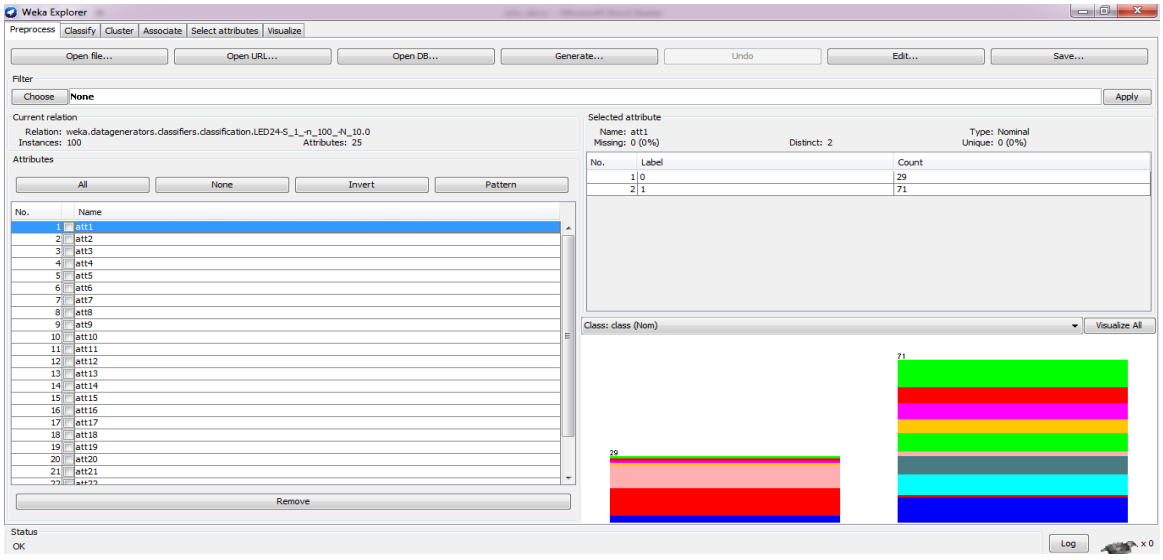
1^{ο} Παράδειγμα : Led24

Περιγραφή δεδομένων

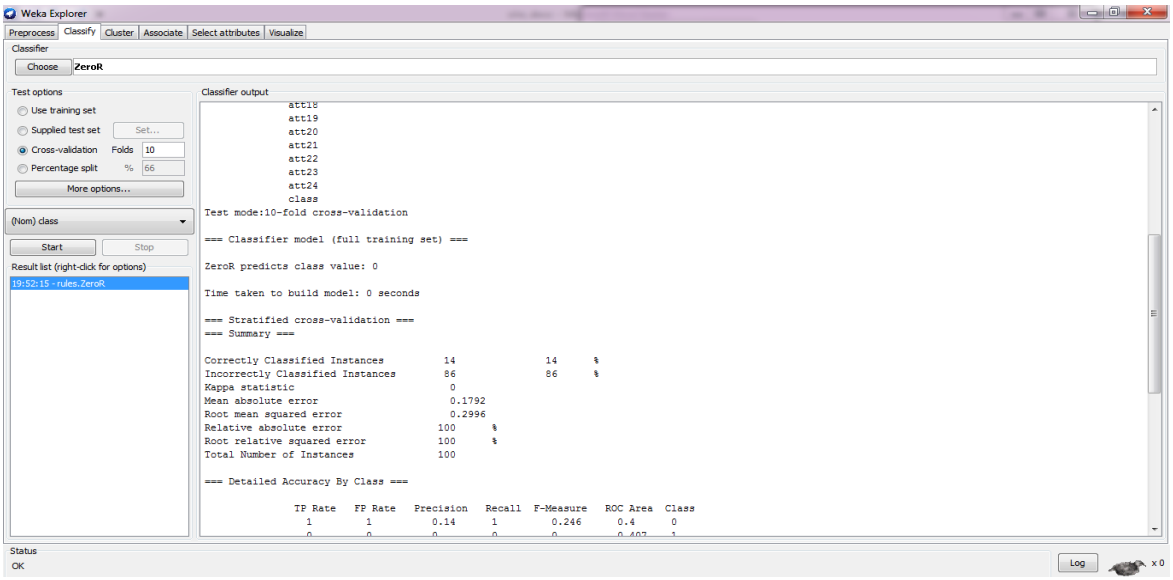
Επιλέγοντας από την επιλογή Generate → Choose διαλέγουμε το παράδειγμα που μας ενδιαφέρει και στη συνέχεια Generate. Στο παράθυρο Preprocess περιγράφονται τα δεδομένα του παραδείγματος γραφικά και αριθμητικά. Όπως προηγουμένως εισάγεται με την ίδια διαδικασία το παράδειγμα που θα εξεταστεί παρακάτω και είναι το «Led24». Το συγκεκριμένο παράδειγμα απαρτίζεται από 100 περιπτώσεις οι οποίες μπορούν να κατηγοριοποιηθούν σε 25 attributes (att1,att2,att3, att4,att5,att6,att7,.., att25).



Πιο αναλυτικά, στην επόμενη εικόνα φαίνεται ότι το attribute 1 χωρίζεται σε δυο label, 0 και 1 όπου 29 και 71 αντικείμενα αντίστοιχα. Το attribute 2 χωρίζεται σε δυο label, 0 και 1 όπου υπάρχουν 45 και 55 αντικείμενα αντίστοιχα στο κάθε label. Την ίδια εικόνα συναντά ο ερευνητής και στα υπόλοιπα attributes του ίδιου παραδείγματος.



Στην επόμενη εικόνα θα εξεταστούν κάποια στατιστικά σχετικά με τις περιπτώσεις του παραδείγματος δηλαδή, οι περιπτώσεις ταξινομήθηκαν σωστά είναι 14% ενώ οι περιπτώσεις ταξινομήθηκαν εσφαλμένα είναι 86%. Το γεγονός αυτό δεν το επιθυμητό καθώς γίνεται κατανοητό ότι το μεγαλύτερο ποσοστό των περιπτώσεων είναι ταξινομημένο λάθος. Ακόμα το απόλυτο μέσο σφάλμα είναι 0,17 ενώ το μέσο τετραγωνικό σφάλμα είναι 0,29.

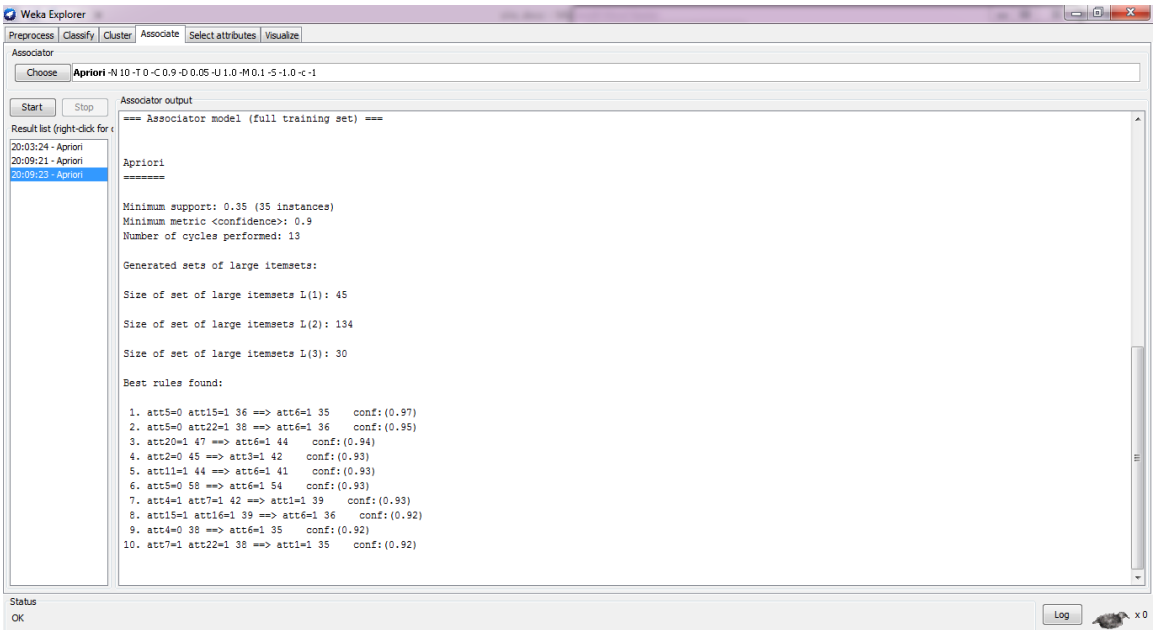


Διαδικασία επιλογής αλγορίθμου Apriori

Για να γίνει η επιλογή του αλγορίθμου που επιθυμούμε στη προκειμένη περίπτωση χρειαζόμαστε τον αλγόριθμο Apriori. Για την επιλογή του ακολουθούμε τη διαδρομή Choose→ Weka, associations→Apriori και στη συνέχεια πατάμε το κουμπί Start για να τρέξει ο αλγόριθμος για να προκύψει η συσχέτιση μεταξύ των αντικειμένων.

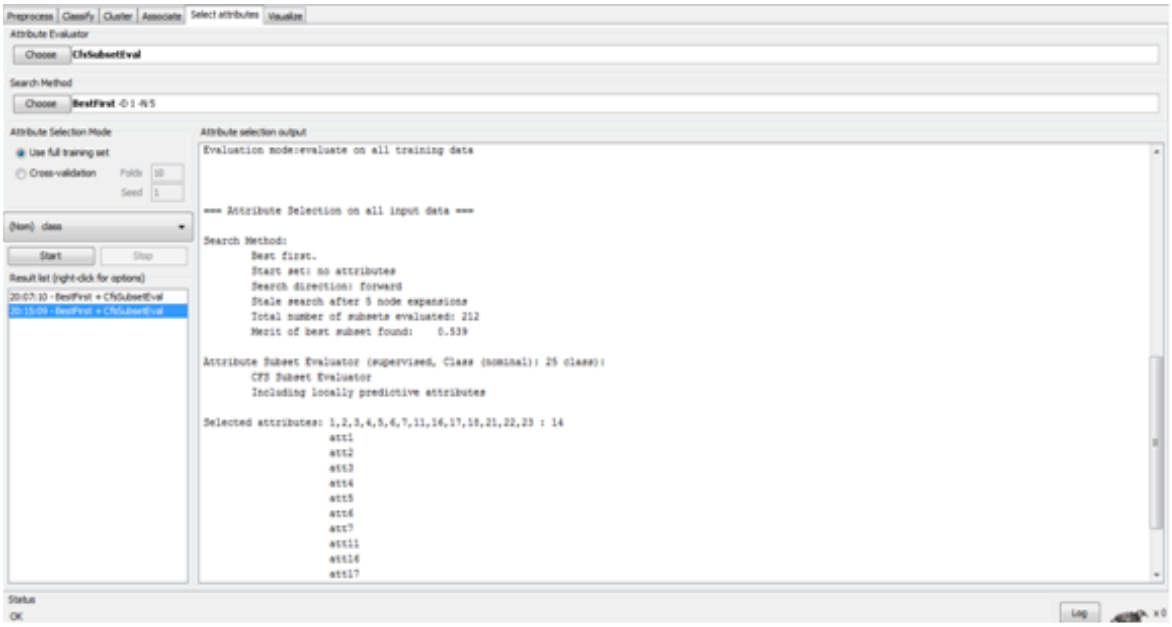
Οι συσχετίσεις μεταξύ των δεδομένων είναι 10. Οι καλύτεροι κανόνες συσχέτισης που βρέθηκαν από τον αλγόριθμο φαίνονται παρακάτω, τα αντικείμενα att1 και att6 εμφανίζονται συχνότερα στη συσχέτιση με τα υπόλοιπα αντικείμενα. Η συσχέτιση ωστόσο είναι θετική αφού κυμαίνεται από 0.92 έως 0.97.

1. Att5=0 att15=1 36 ==> att6=1 35
2. Att5=0 att22=1 38 ==> att6=1 36
3. Att20=1 47 == > att6=1 44
4. Att2=0 45 == > att3=1 42
5. att11=1 44 == > att6=1 54
6. att5=0 58 == > att6=1 54
7. att4=1 att7=1 42 == > att1=1 39
8. att15=1 att16 =1 39 == >att6=1 36
9. att4=0 38 == > att6=1 35
10. att7=1 att22=1 38 == > att1=1 35

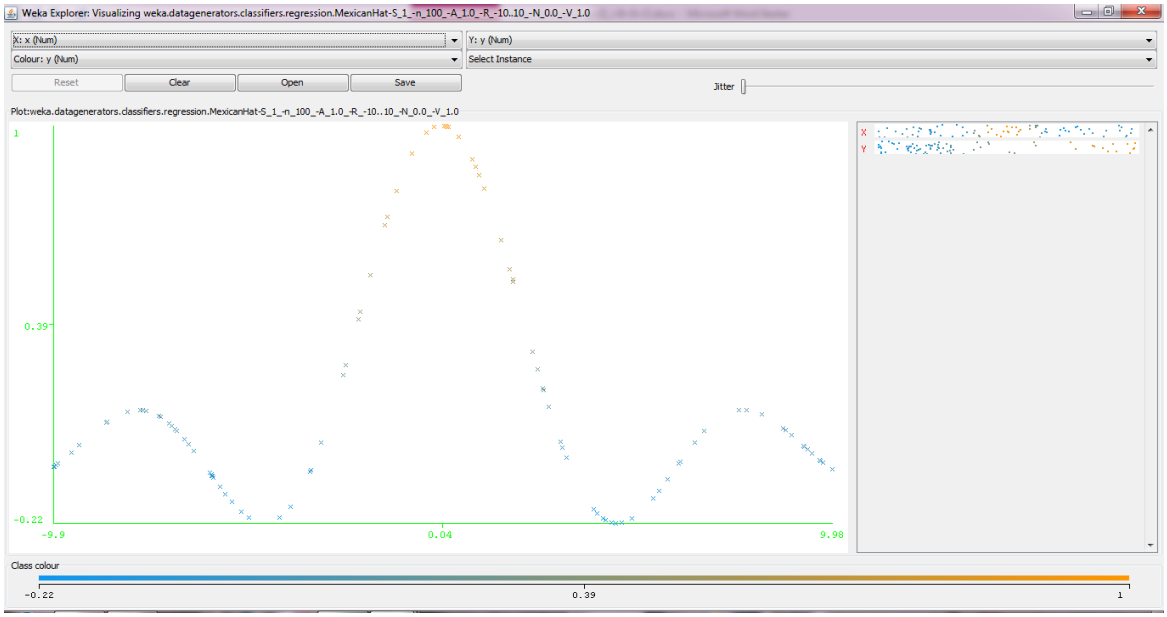


Αποτελέσματα

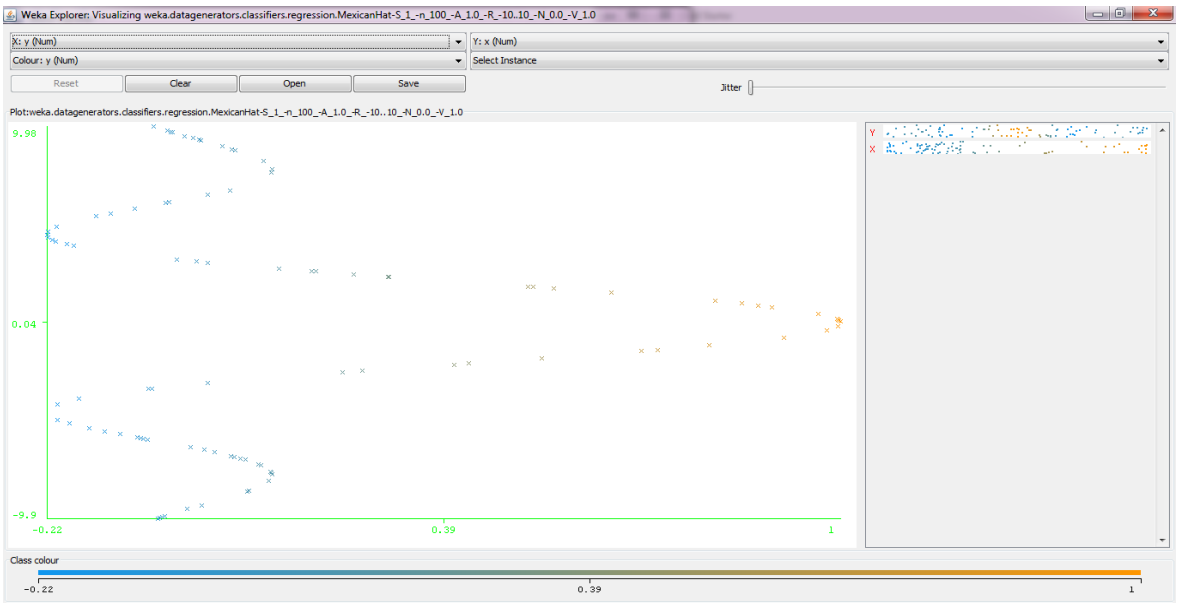
Στην επόμενη καρτέλα κάποια σημεία εκ των οποίων κεντρίζουν το ενδιαφέρον του ερευνητή είναι ο συνολικός αριθμός των υποσυνόλων που αξιολογήθηκαν που είναι 212.



Στην επόμενη εικόνα φαίνεται η θετική συσχέτιση μεταξύ των μεταβλητών x και y, αφού παρατηρούμε ότι όσο αυξάνονται-μειώνονται οι τιμές της x, τόσο επηρεάζονται και οι τιμές τις μεταβλητής y. Επίσης η μεταβλητή y έχει μια αυξητική τάση και παίρνει τη μεγαλύτερη της τιμή στο 1 ενώ στη συνέχεια μειώνονται οι τιμές της. Είναι στατιστικά σημαντική αφού η συσχέτιση είναι θετική.



Σε αυτή την εικόνα φαίνεται η αρνητική συσχέτιση καθώς η εξαρτημένη μεταβλητή είναι η x ενώ η ανεξάρτητη είναι η y. Σε αυτή τη περίπτωση η διακύμανση των τιμών της ανεξάρτητης μεταβλητής είναι από – 0,22 έως 1. Καθώς όσο αυξάνονται οι τιμές της ανεξάρτητης μεταβλητής y οι τιμές της εξαρτημένης μεταβλητής x μειώνονται.

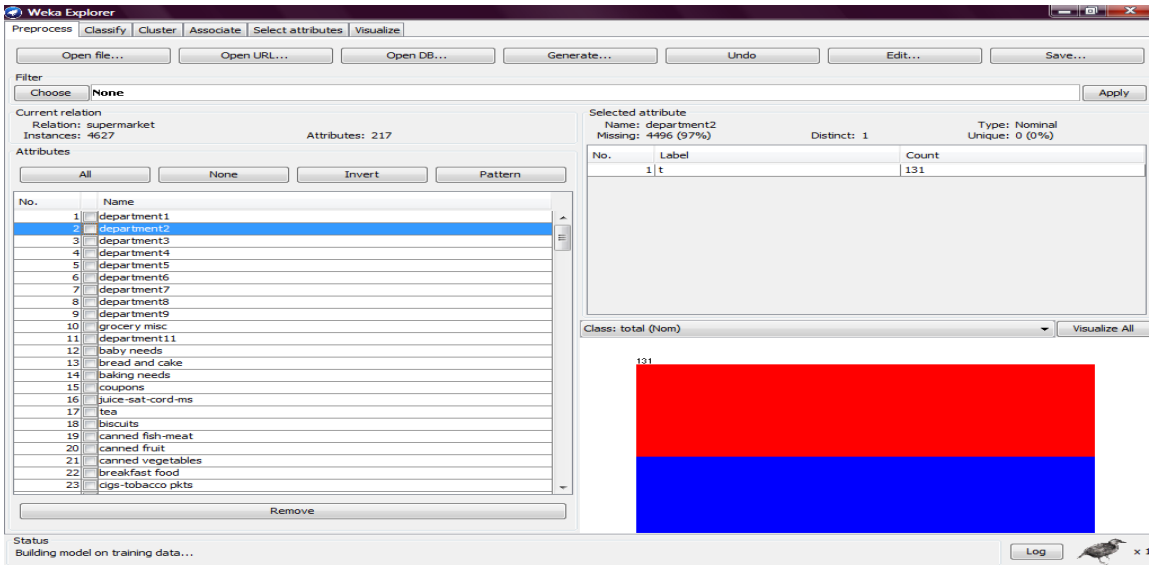


2ο Παράδειγμα: Supermarket

Περιγραφή δεδομένων

Επιλέγοντας από την επιλογή Generate → Choose διαλέγουμε το παράδειγμα που μας ενδιαφέρει και στη συνέχεια Generate. Στο παράθυρο Preprocess περιγράφονται τα δεδομένα του παραδείγματος γραφικά και αριθμητικά.

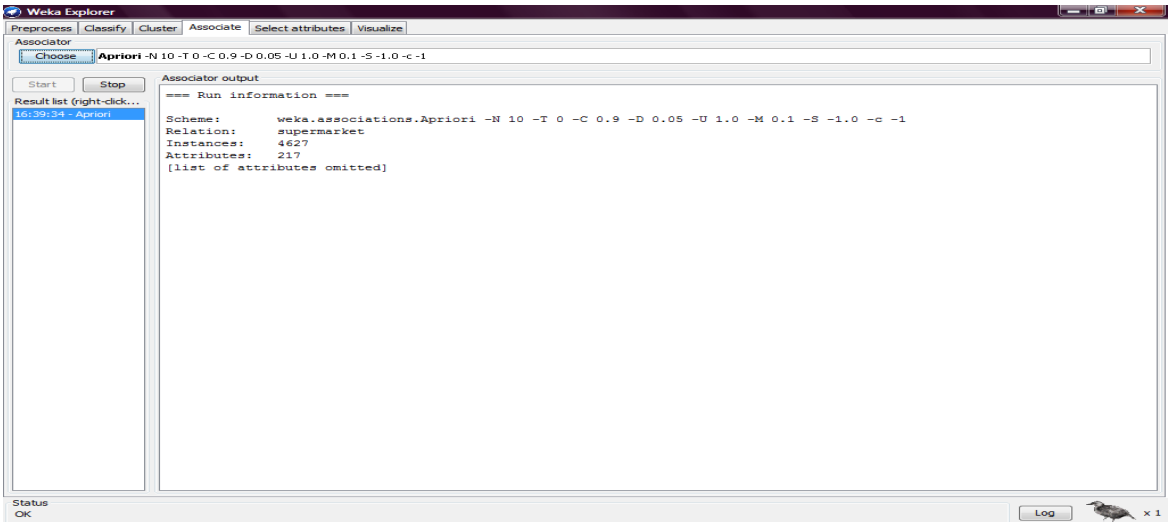
Το συγκεκριμένο παράδειγμα έχει 217 γνωρίσματα-attributes, τα οποία χωρίζονται σε δυο κλάσεις.



Διαδικασία επιλογής αλγόριθμου Apriori

Στην καρτέλα Associate δίνεται η δυνατότητα στο χρήστη να επιλέξει τον αλγόριθμο με τον οποίο θέλει να επεξεργαστεί τα δεδομένα και τον οποίο θα τρέξει το πρόγραμμα. Στη περίπτωση αυτή ο αλγόριθμος που θα χρησιμοποιήσουμε είναι ο Apriori , όπως φαίνεται και στην παρακάτω εικόνα.

Για να γίνει η επιλογή του αλγορίθμου που επιθυμούμε στη προκειμένη περίπτωση χρειαζόμαστε τον αλγόριθμο Apriori. Για την επιλογή του ακολουθούμε τη διαδρομή Choose→ Weka, associations→Apriori και στη συνέχεια πατάμε το κουμπί Start για να τρέξει ο αλγόριθμος για να προκύψει η συσχέτιση μεταξύ των αντικειμένων.



Αποτελέσματα

Οι συσχετίσεις ανάμεσα στα δεδομένα του supermarket είναι 10. Καταρχάς πρέπει να αναφέρουμε την κλίμακα εμπιστοσύνης με minimum=0,9. Παρατηρούμε στα αποτελέσματα παρακάτω πως το μέτρο εμπιστοσύνης κυμαίνεται από 0,91 μέχρι 0,92 επομένως έχουμε θετική συσχέτιση. Η συσχέτιση πραγματοποιείται ανάμεσα στο bread and cake που είναι σταθερή μεταβλητή

1)biscuits, frozen foods, fruit με bread and cake

2) baking needs, biscuits, fruit με bread and cake

3) baking needs, frozen foods, fruit με bread and cake

4) biscuits, fruit, vegetables με bread and cake

5)party snack foods,fruit με bread and cake

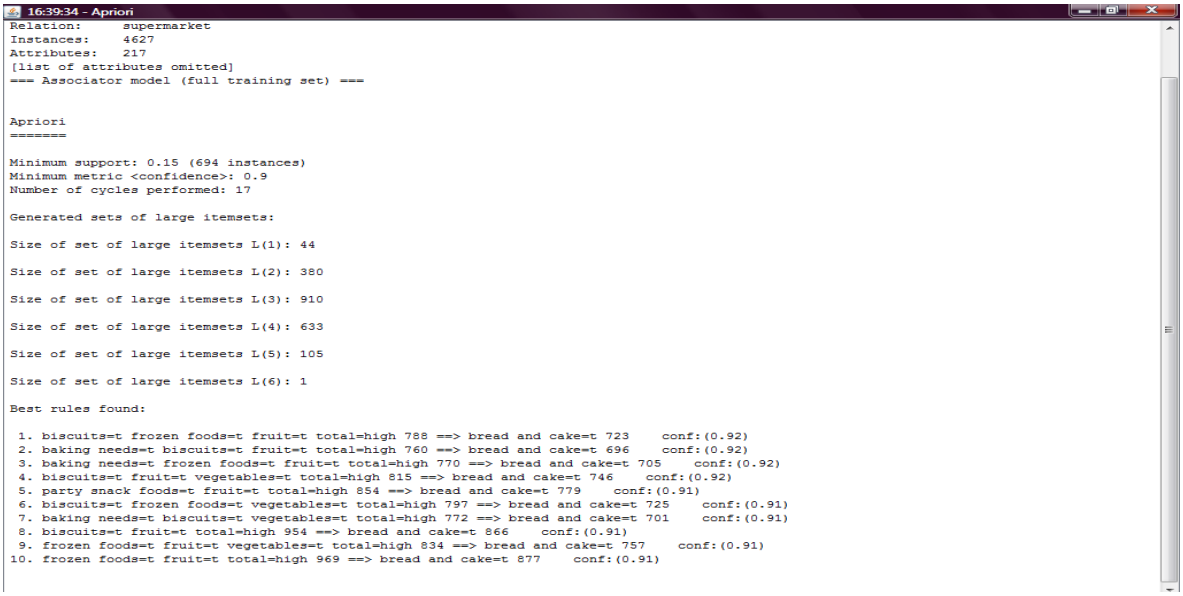
6) biscuits, frozen foods, vegetables με bread and cake

7) biscuits, baking needs, vegetables με bread and cake

8) biscuits, fruit με bread and cake

9) frozen foods ,vegetables με bread and cake

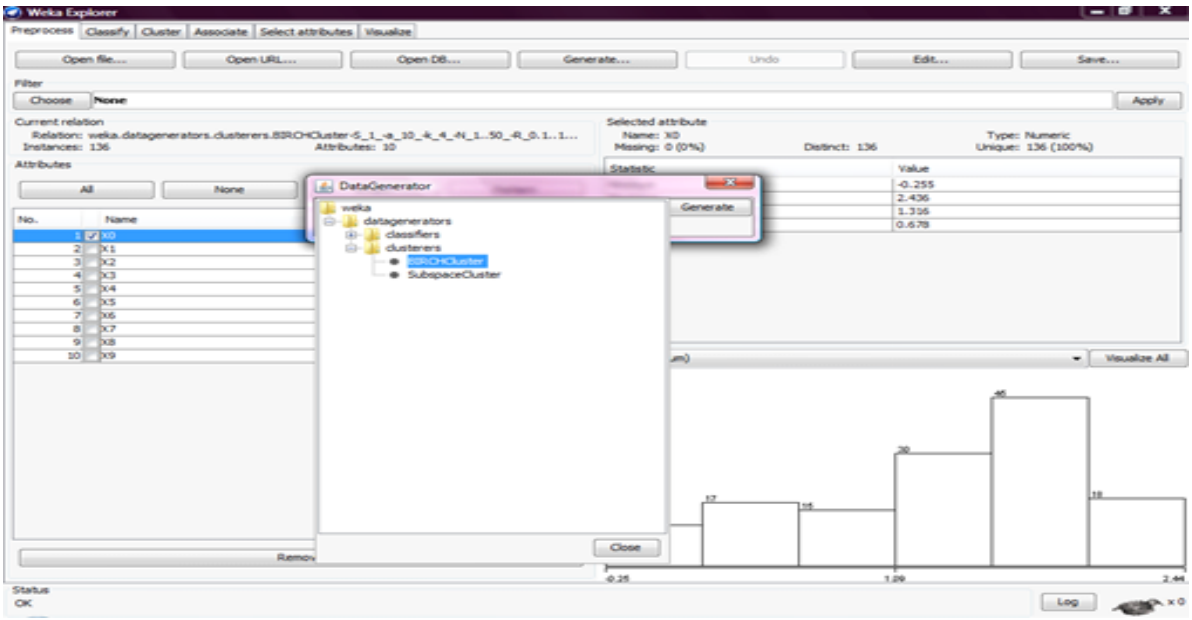
10) frozen foods, fruit με bread and cake



3ο Παράδειγμα: BIR CHCCUSTER

Περιγραφή δεδομένων

Επιλέγοντας τυχαία το παράδειγμα Big Chccuster παρατηρούμε ότι περιλαμβάνει 10 γνωρίσματα-attributes (x0, x1, x2. X3, x4, x5, x6, x7, x8, x9).



Σε πρώτη φάση έχουμε την ομαδοποίηση των δεδομένων που τρέχουμε. Στο παράθυρο Preprocess εμφανίζονται τα αποτελέσματα γραφικά και αριθμητικά όπως φαίνεται στην παρακάτω εικόνα για κάθε μεταβλητή ξεχωριστά.

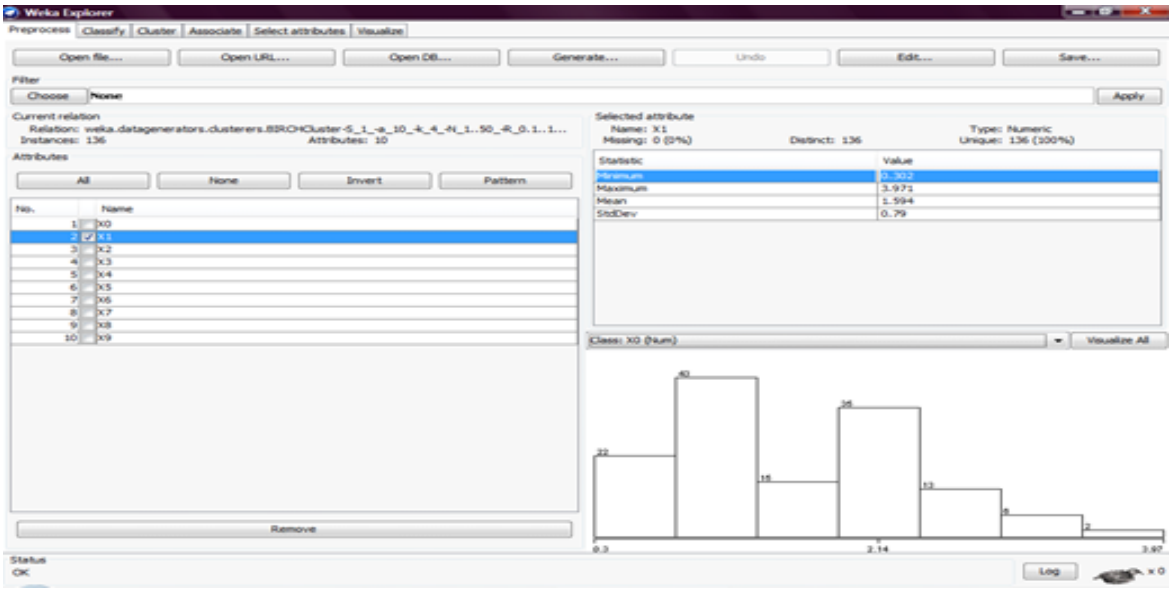
Για την μεταβλητή x0:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,255

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=2,436

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=1,316

StdDeav η τυπική απόκλιση των δεδομένων=0,678



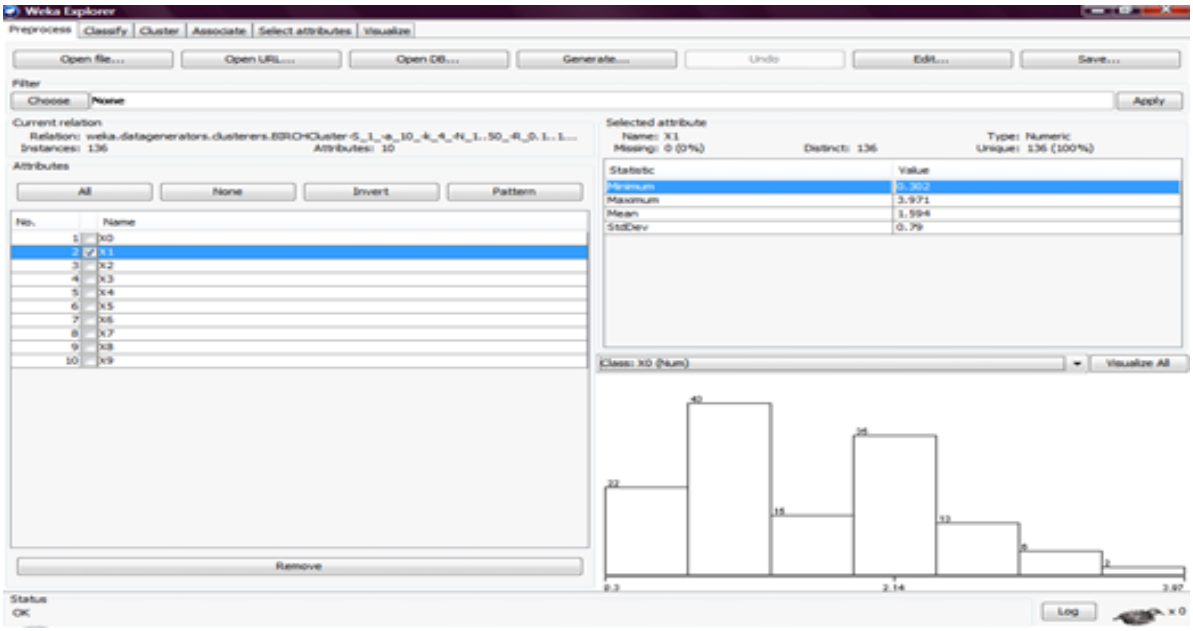
Για την μεταβλητή x1, όπως φαίνεται στην παρακάτω εικόνα:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,302

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=3,971

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=1,594

StdDeav η τυπική απόκλιση των δεδομένων=0,79



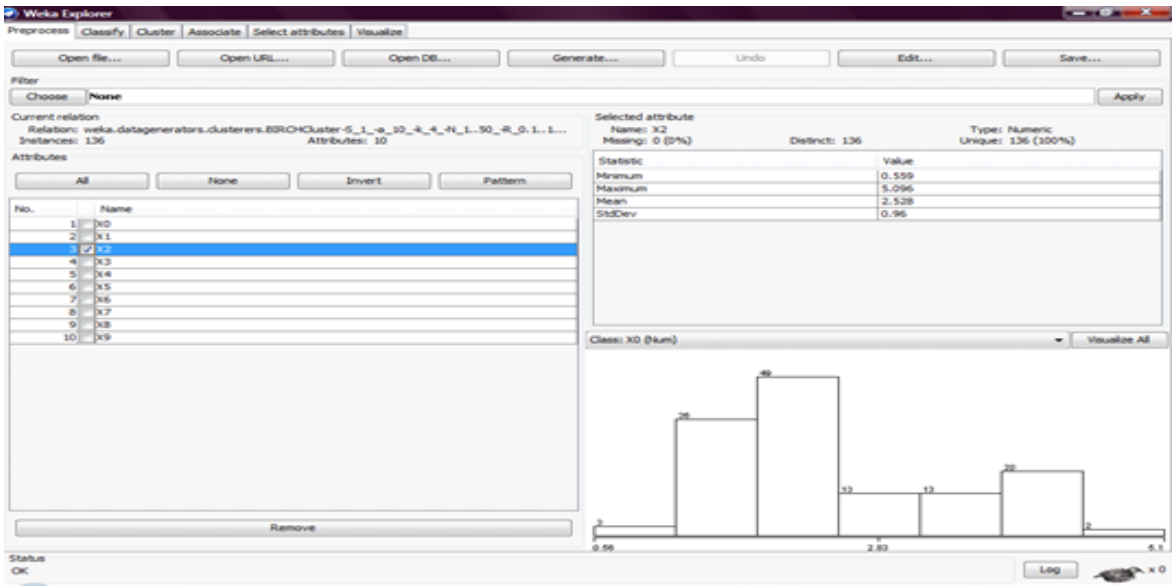
Για την μεταβλητή x2 όπως φαίνεται στην παρακάτω εικόνα:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,559

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=5,096

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=2,528

StdDeav η τυπική απόκλιση των δεδομένων=0,96



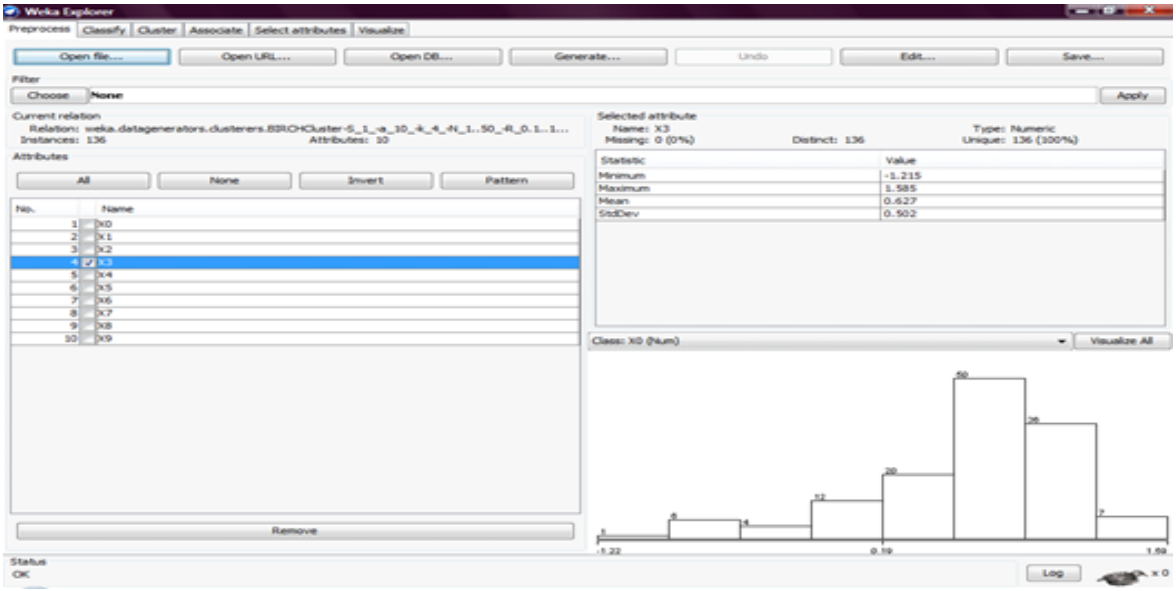
Για την μεταβλητή x3 όπως φαίνεται στην παρακάτω εικόνα:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-1,215

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=1,585

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=0,627

StdDeav η τυπική απόκλιση των δεδομένων=0,502



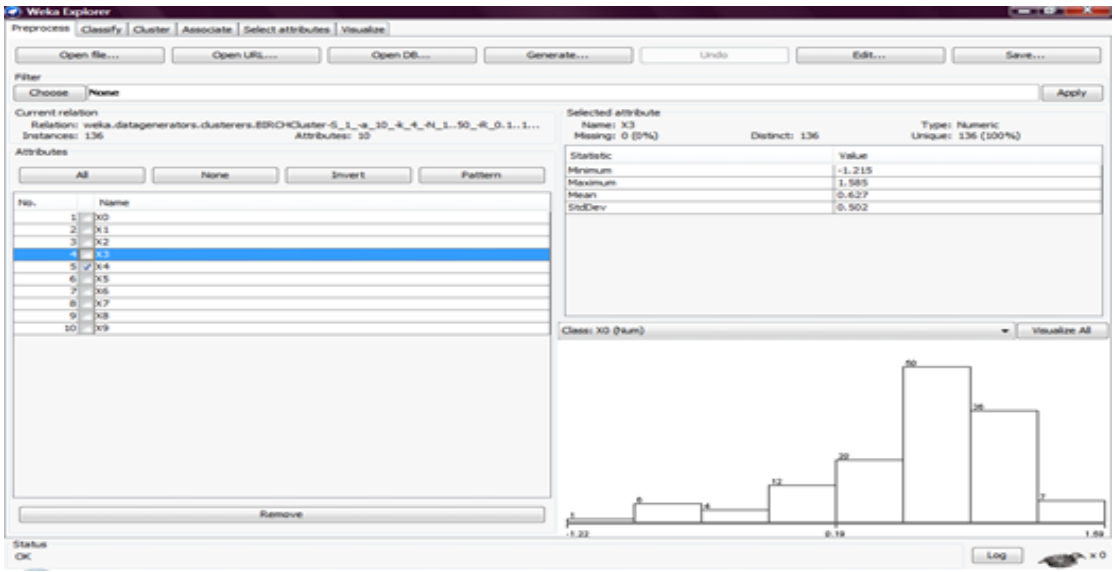
Για την μεταβλητή x4, όπως φαίνεται στην παρακάτω εικόνα:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,838

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=4,784

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=3,134

StdDeav η τυπική απόκλιση των δεδομένων=0,502



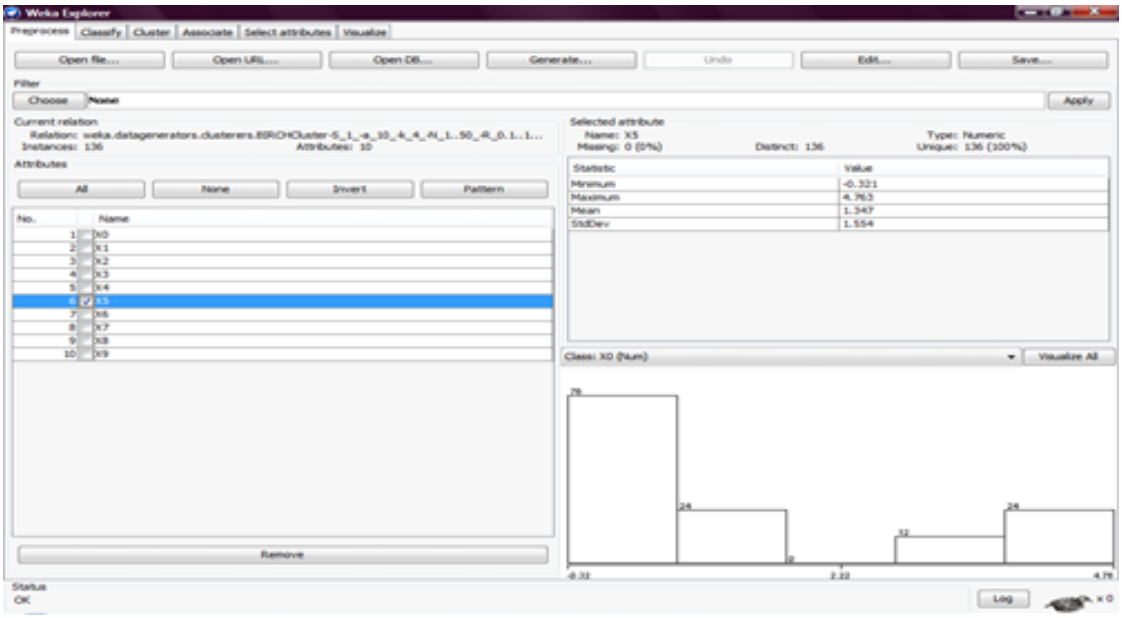
Για την μεταβλητή x5, όπως φαίνεται στο παρακάτω εικόνα :

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,321

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=4,63

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=1,347

StdDeav η τυπική απόκλιση των δεδομένων=1,554



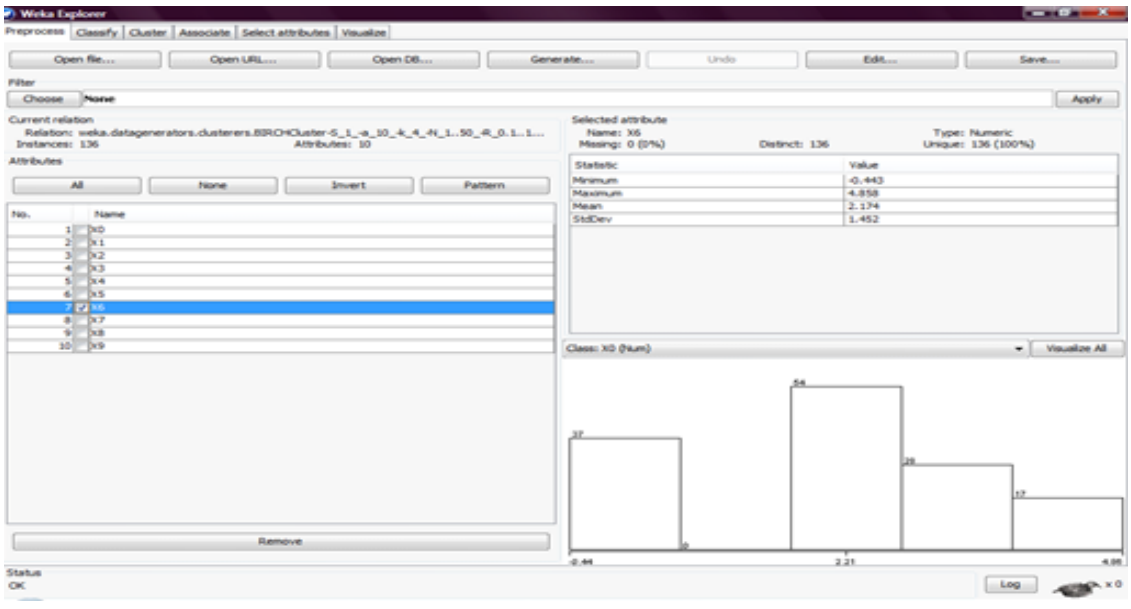
Για την μεταβλητή χ6 όπως φαίνεται στην παρακάτω εικόνα:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,443

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=4,858

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=2,174

StdDeav η τυπική απόκλιση των δεδομένων=1,452



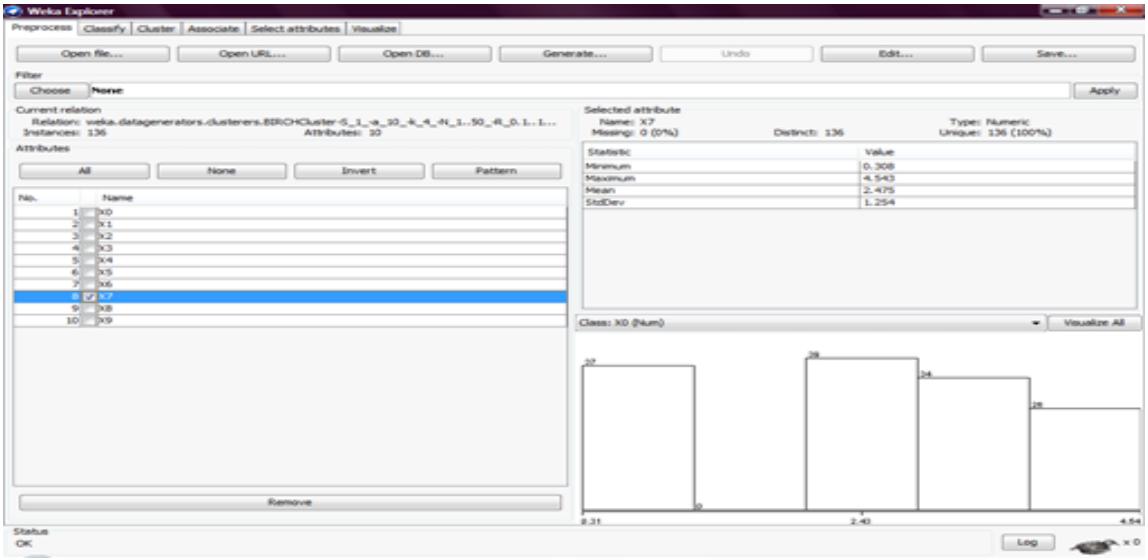
Για την μεταβλητή x7, όπως φαίνεται στο παρακάτω διάγραμμα:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,308

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=4,543

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=2,475

StdDeav η τυπική απόκλιση των δεδομένων=1,254



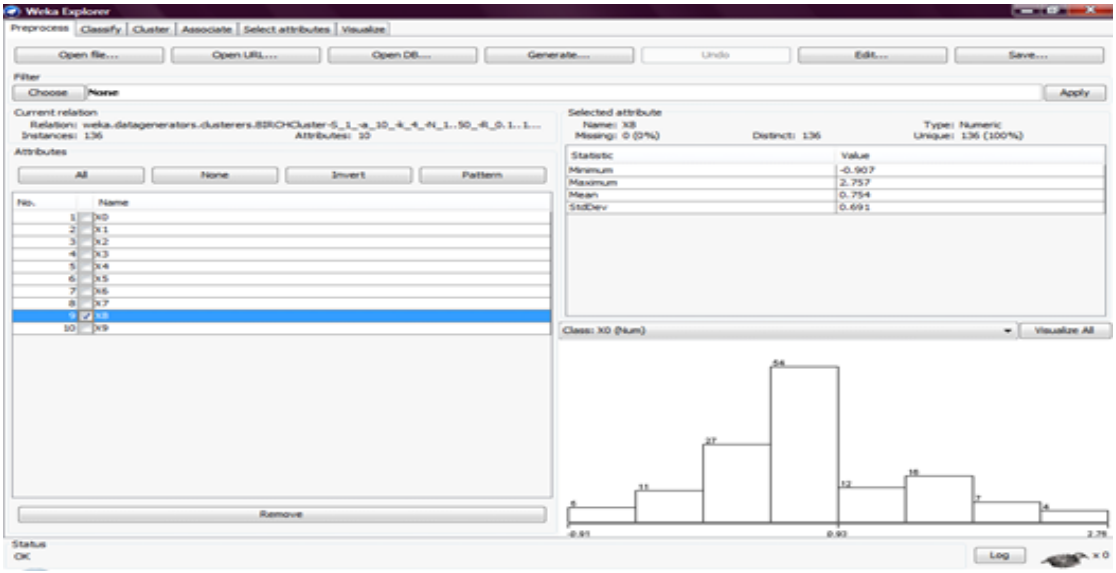
Για την μεταβλητή x8, όπως φαίνεται στην παρακάτω εικόνα:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,907

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=2,767

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=0,754

StdDeav η τυπική απόκλιση των δεδομένων=0,691



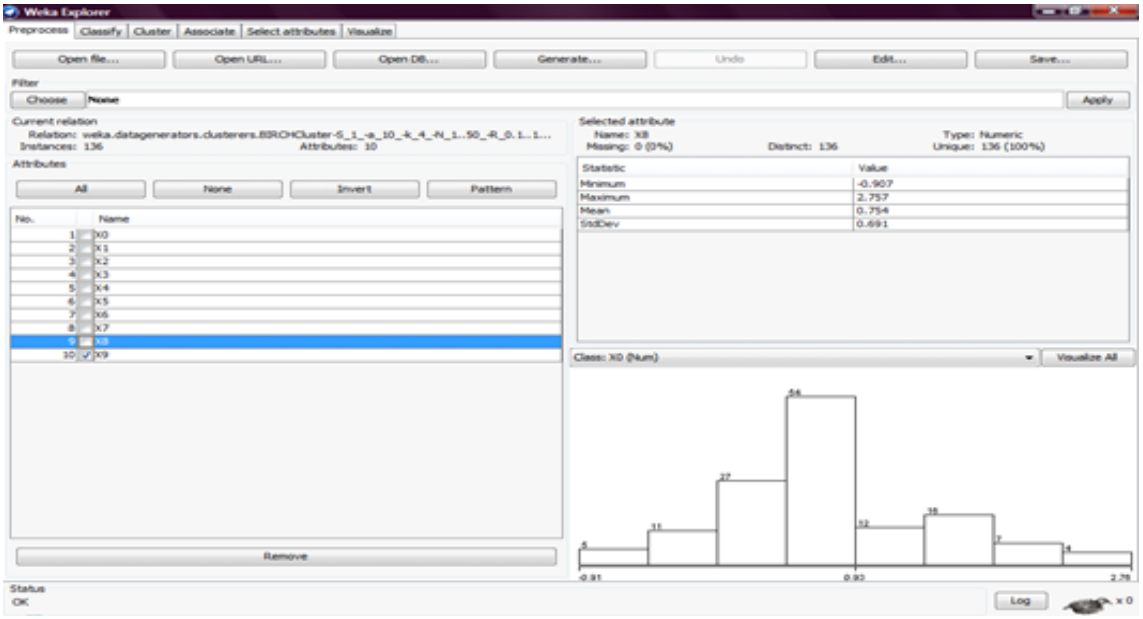
Για την μεταβλητή x9, όπως φαίνεται στο παρακάτω διάγραμμα:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων =-0,049

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=3,147

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=2,114

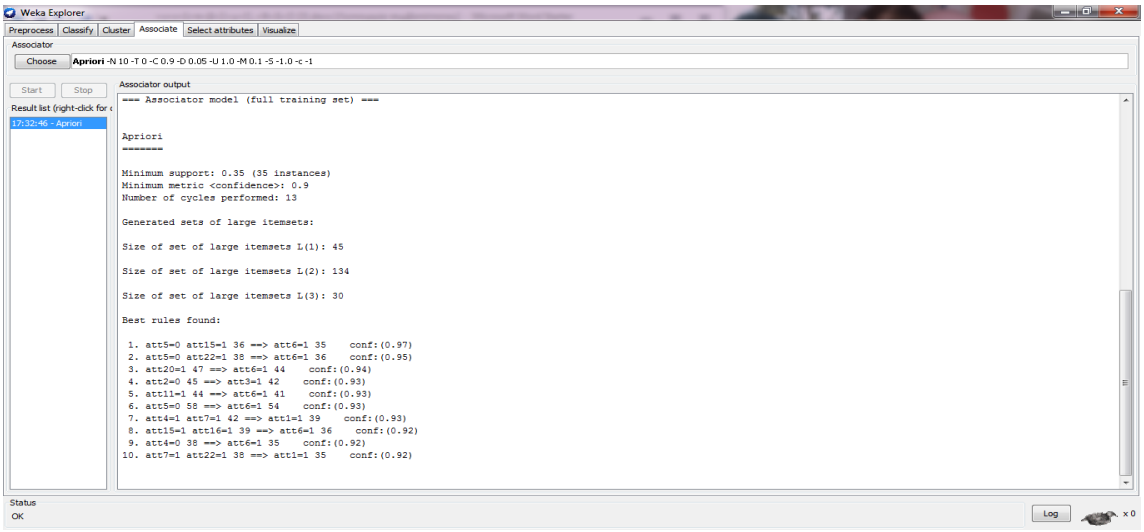
StdDeav η τυπική απόκλιση των δεδομένων=0,616



<u>Διαδικασία επιλογής αλγόριθμου Apriori</u>

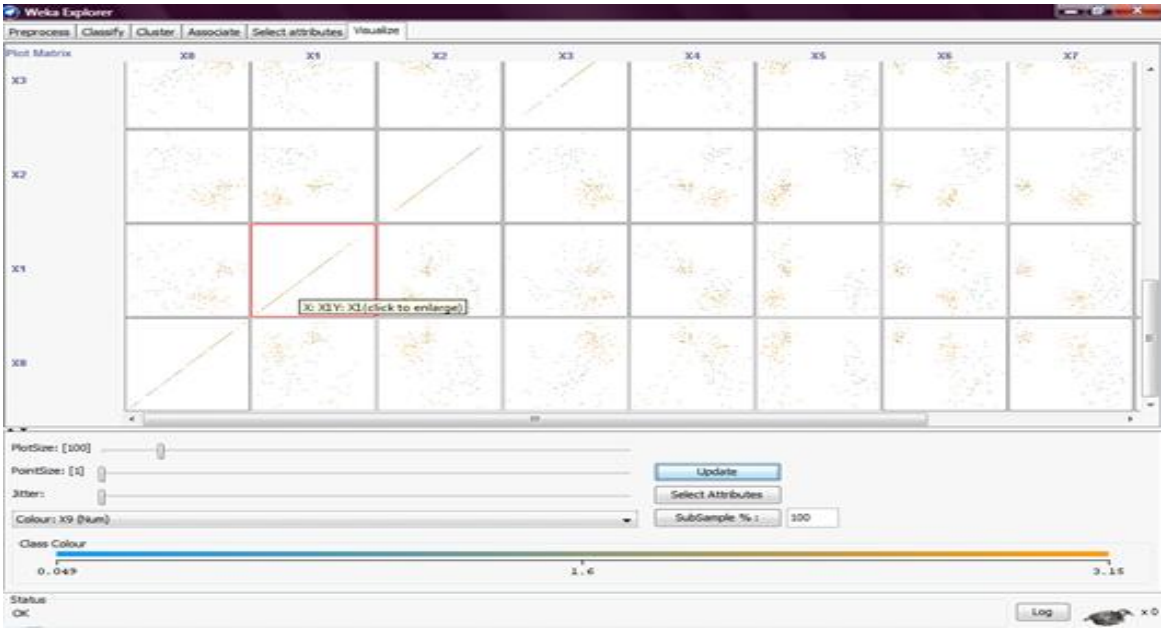
Στην συνέχεια στη καρτέλα Association θα αναλυθούν κάποια μέτρα όπως το ελάχιστο μέτρο εμπιστοσύνης που είναι 0,9, η ελάχιστη υποστήριξη που είναι 0,35 και ο αριθμός κύκλων που εκτελούνται όπου είναι 13.

Οι συσχετίσεις μεταξύ του συνόλου των δεδομένων είναι 10. Οι καλύτεροι κανόνες συσχέτισης που βρέθηκαν από τον αλγόριθμο φαίνονται πιο κάτω, τα αντικείμενα att1 και att6 εμφανίζονται συχνότερα στη συσχέτιση με τα υπόλοιπα αντικείμενα. Η συσχέτιση ωστόσο είναι θετική και στατιστικά σημαντική. αφού κυμαίνεται από 0.92 έως 0.97



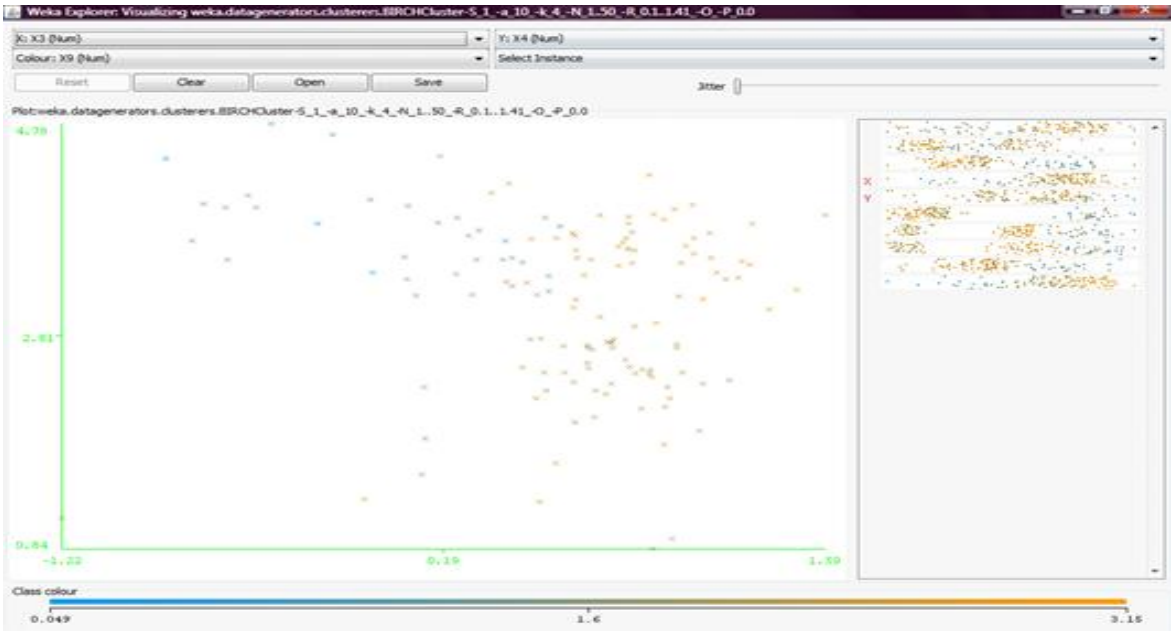
Αποτελέσματα

Στην καρτέλα Visualize έχουμε την οπτικοποίηση των δεδομένων σε δισδιάστατα γραφήματα.



Ενδεικτικά για ανεξάρτητη μεταβλητή X η X3 και εξαρτημένη μεταβλητή Y η X4 προκύπτει το παρακάτω διάγραμμα διασποράς και παρατηρούμε πως οι κλάσεις κινούνται από το 0,84 μέχρι

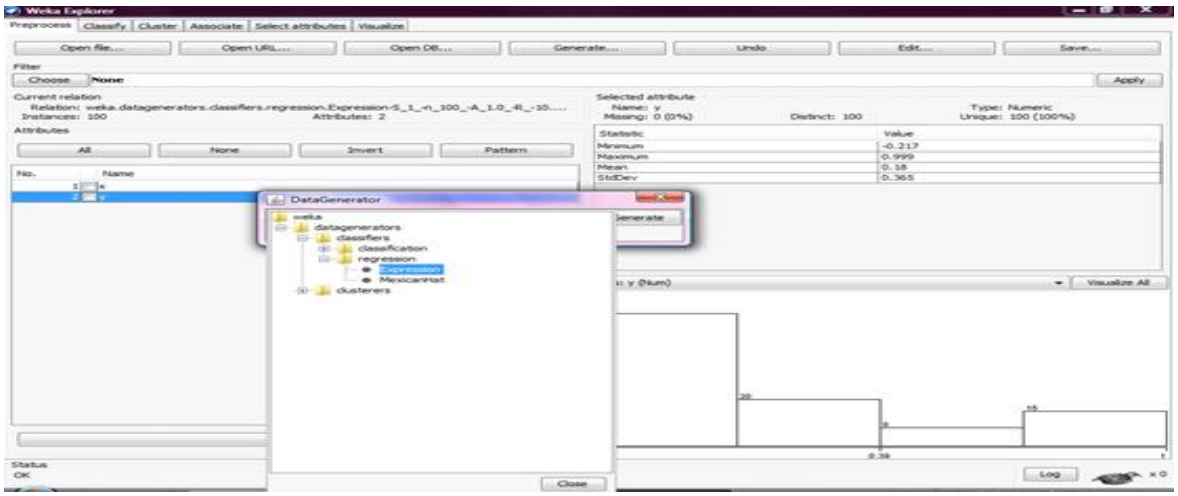
4,78. Όπως φαίνεται και στο διάγραμμα η διασπορά είναι μεγάλη και δεν υπάρχει καθόλου συσχέτιση μεταξύ των μεταβλητών.



4^{ο} Παράδειγμα: EXPRESSION

Περιγραφή δεδομένων

Στη συνέχεια θα εξετάσουμε τη συσχέτιση του παραδείγματος Expression το οποίο περιλαμβάνει 2 αντικείμενα (x και y). Επιλέγοντας από την επιλογή Generate → Choose διαλέγουμε το παράδειγμα που μας ενδιαφέρει και στη συνέχεια Generate.



Αποτελέσματα

Στο παράθυρο Preprocess εμφανίζονται τα αποτελέσματα γραφικά και αριθμητικά όπως φαίνεται στην παρακάτω εικόνα, πιο αναλυτικά:

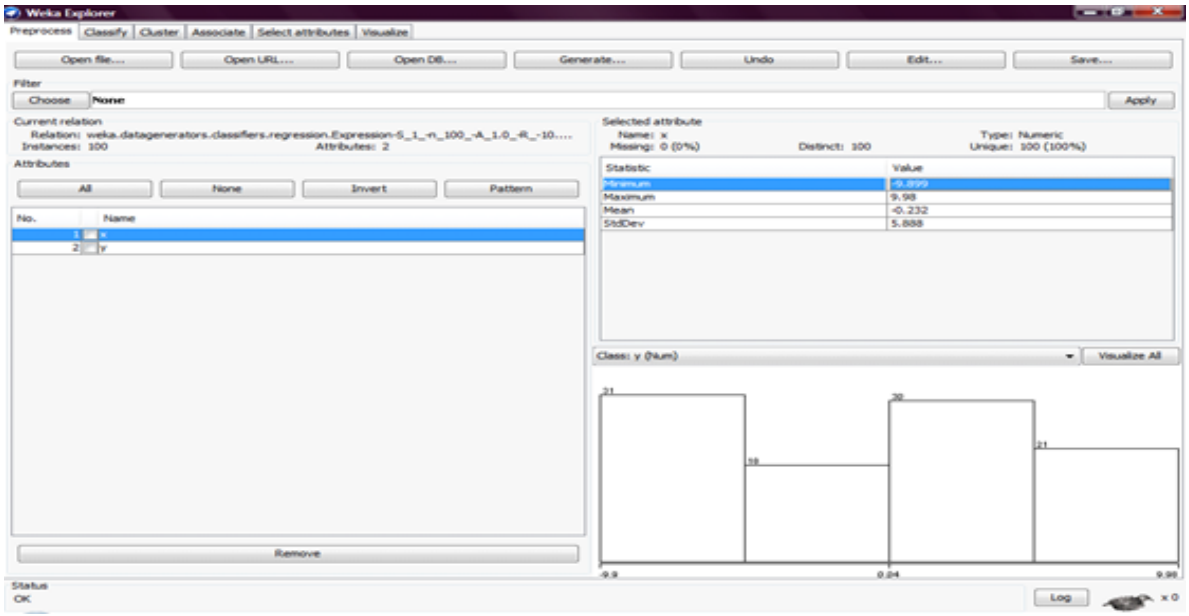
Για την μεταβλητή x:

Μέσος =-9,899

Μέγιστο=9,98

Μέσος=-0,232

Τυπική απόκλιση =5,888



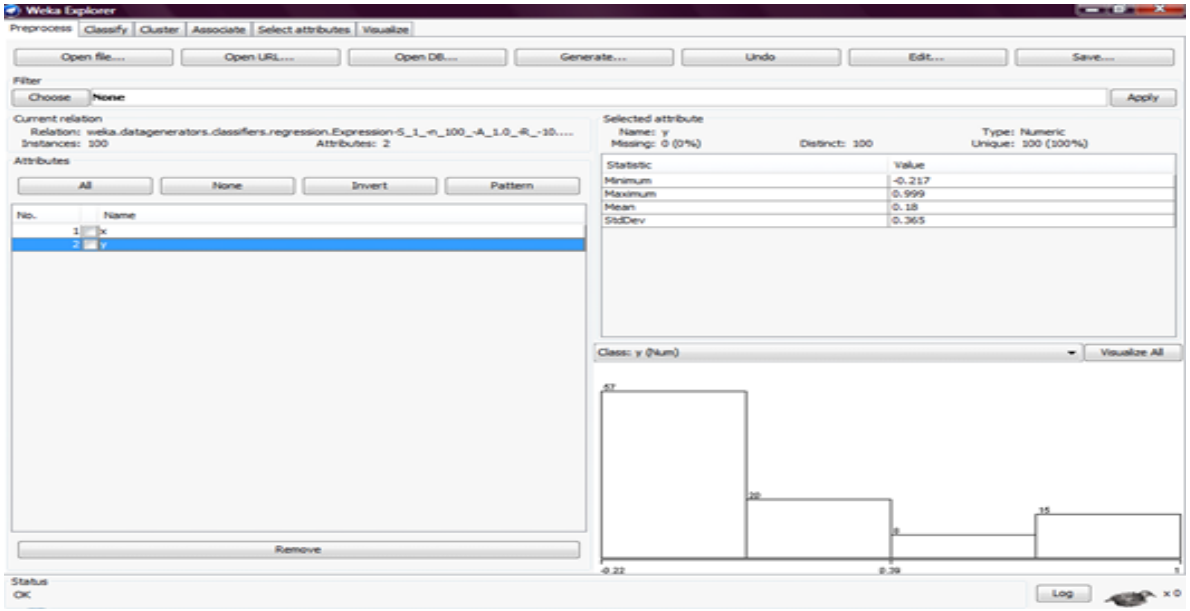
Για την μεταβλητή y:

Μέσος =-0,217

Μέγιστο=0,999

Μέσος=0,18

Τυπική απόκλιση =0,365

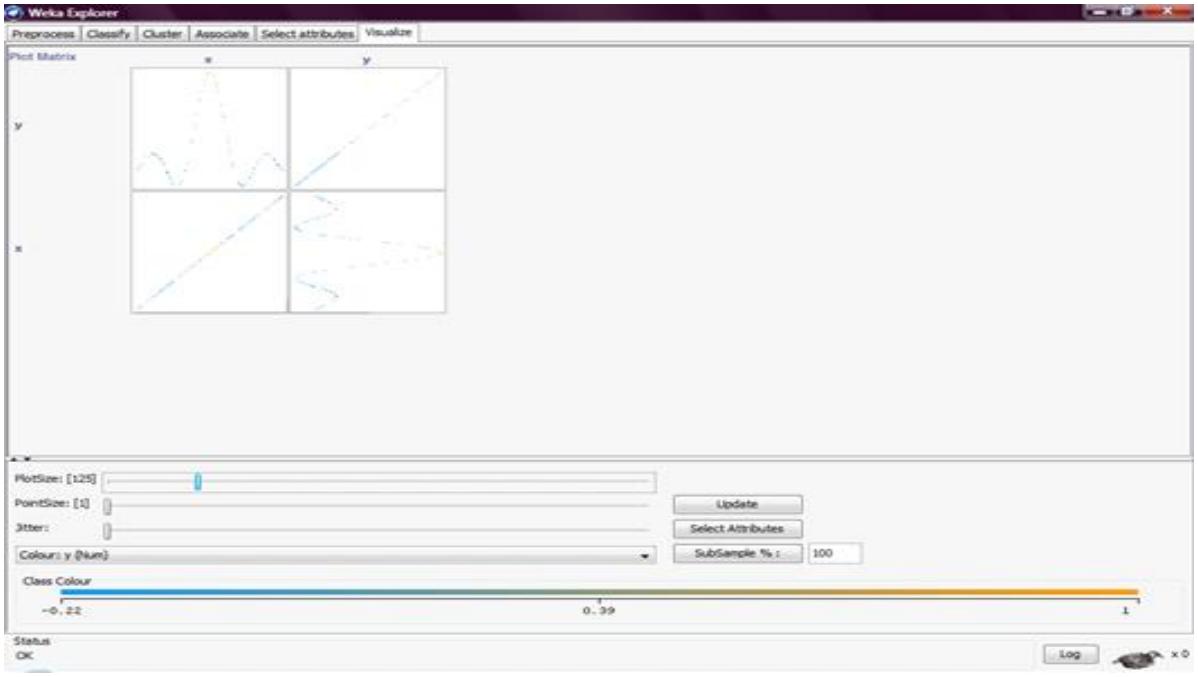


Στην επόμενη καρτέλα Cluster θα εξετάσουμε τα δεδομένων x και y πως εμφανίζονται στα Cluster 0,1,2 και 3 με τα στατιστικά στοιχεία του μέσου αριθμητικού όπου και της τυπικής απόκλισης.

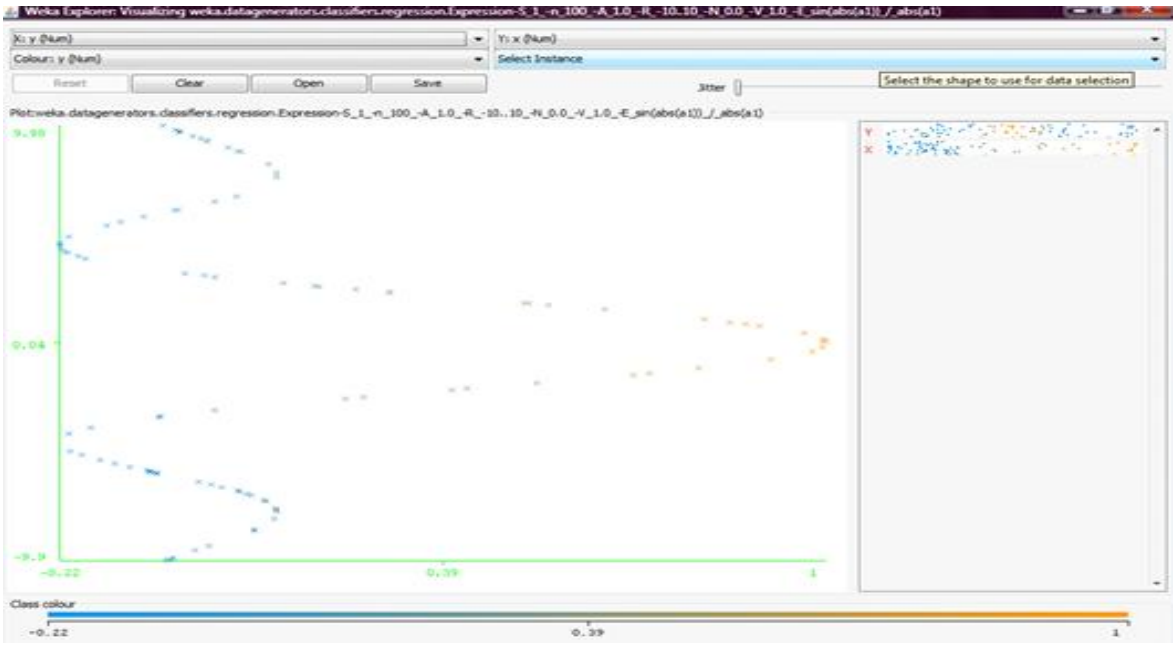
Τα στοιχεία που απαρτίζουν τις μεταβλητές x και y τοποθετούνται στα 0,1,2 και 3 σε ποσοστά 34% , 37% , 8% και 21 % αντίστοιχα . Πιο αναλυτικά και με τα στατιστικά μέτρα που προαναφέρθηκαν φαίνονται στον επόμενο πίνακα.

x	0	1	2	3
Mean	6,4363	-6,6426	0,0079	0,2877
Std. Dev.	2,4318	1,8892	0,2669	1,8114
y				
Mean	-0,0335	-0,0128	0,9882	0,5568
Std. Dev .	0,1177	0,0998	0,0114	0,2458

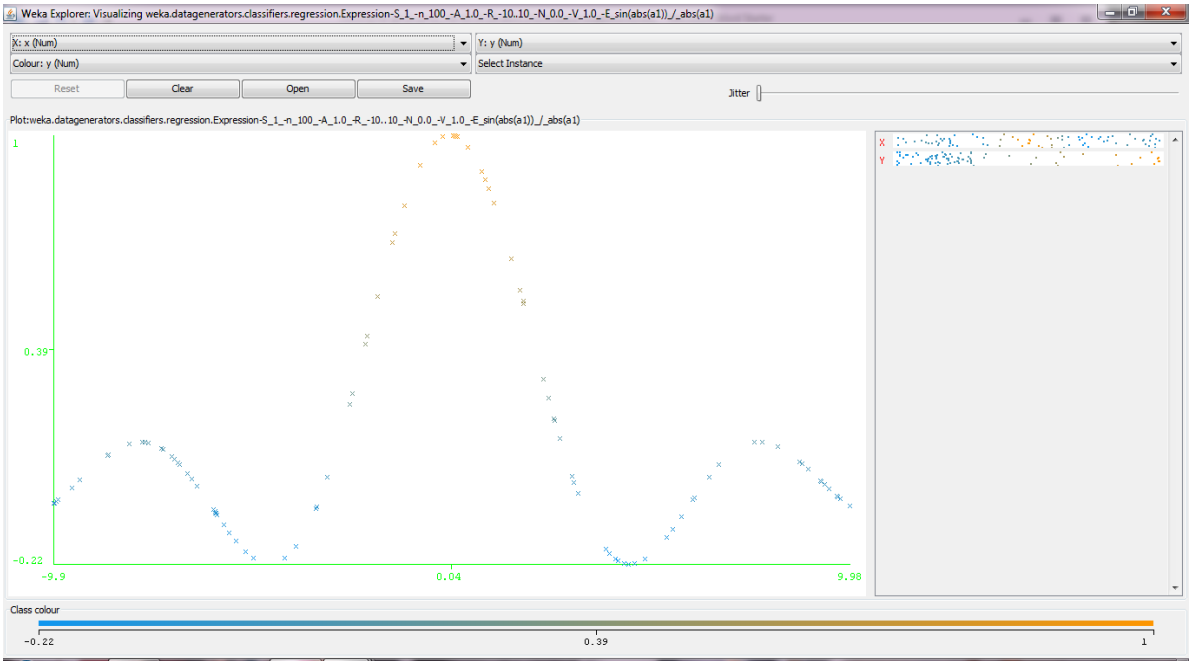
Στην καρτέλα Visualize έχουμε την οπτικοποίηση των δεδομένων σε δισδιάστατα γραφήματα



Ενδεικτικά για ανεξάρτητη μεταβλητή X η y και εξαρτημένη μεταβλητή Y η x προκύπτει το παρακάτω διάγραμμα διασποράς και παρατηρούμε πως οι κλάσεις κινούνται από το -9,9 μέχρι 9,98.



Στην αντίθετη περίπτωση δηλαδή όταν ανεξάρτητη μεταβλητή είναι η x και εξαρτημένη μεταβλητή y οπότε προκύπτει το παρακάτω διάγραμμα διασποράς και παρατηρούμε πως οι κλάσεις κινούνται από το -9,9 μέχρι 9,98.



5^{o} Παράδειγμα: WEATHER

Περιγραφή δεδομένων

Επιλέγοντας από την επιλογή Generate → Choose διαλέγουμε το παράδειγμα που μας ενδιαφέρει και στη συνέχεια Generate. Στο συγκεκριμένο παράδειγμα έχουμε επιλέξει την μεταβλητή Weather που περιλαμβάνει 5 γνωρίσματα-attributes τα οποία είναι: outlook, temperature, humidity, windy, play.

Αποτελέσματα

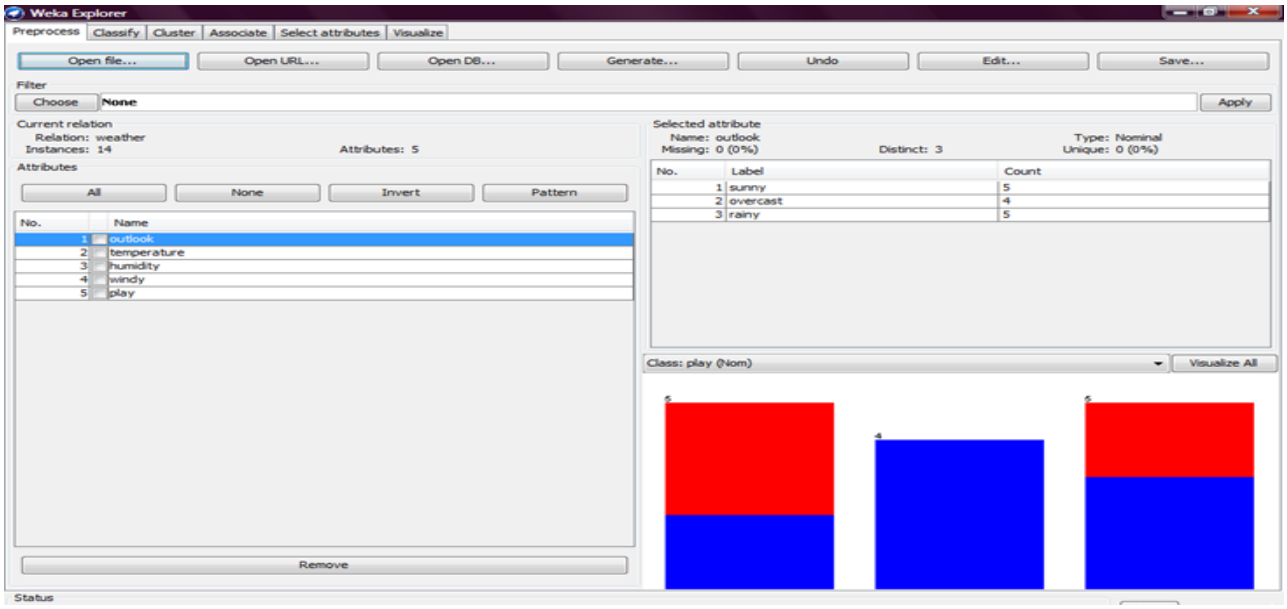
Στο παράθυρο Preprocess εμφανίζονται τα αποτελέσματα γραφικά και αριθμητικά όπως φαίνεται στην παρακάτω εικόνα, πιο αναλυτικά:

Για την μεταβλητή outlook οι τιμές είναι:

Sunny 5

Overcast 4

Rainy 5



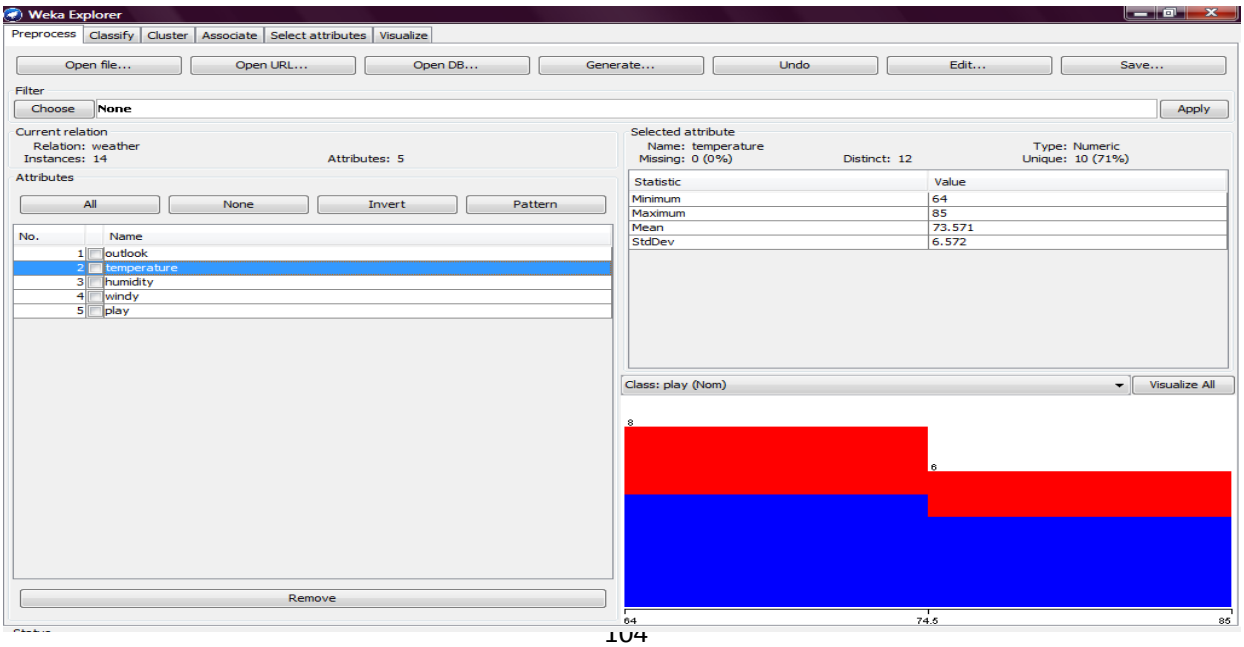
Για την μεταβλητή temperature οι τιμές είναι:

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων 64

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=85

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=73.571

StdDeav η τυπική απόκλιση των δεδομένων=6572



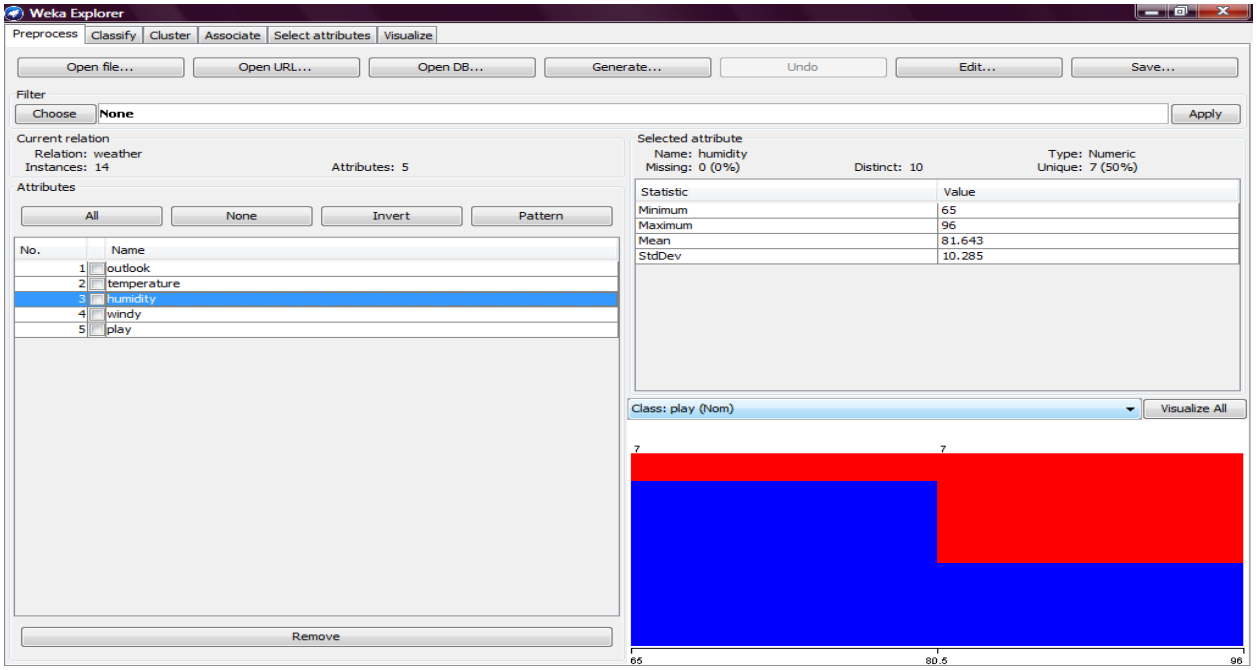
Για την μεταβλητή humidity οι τιμές είναι

Minimum η ελάχιστη τιμή που εντοπίστηκε στο σύνολο δεδομένων 65

Maximum η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων=96

Mean η μέση τιμή που εντοπίστηκε στο σύνολο των δεδομένων=81643

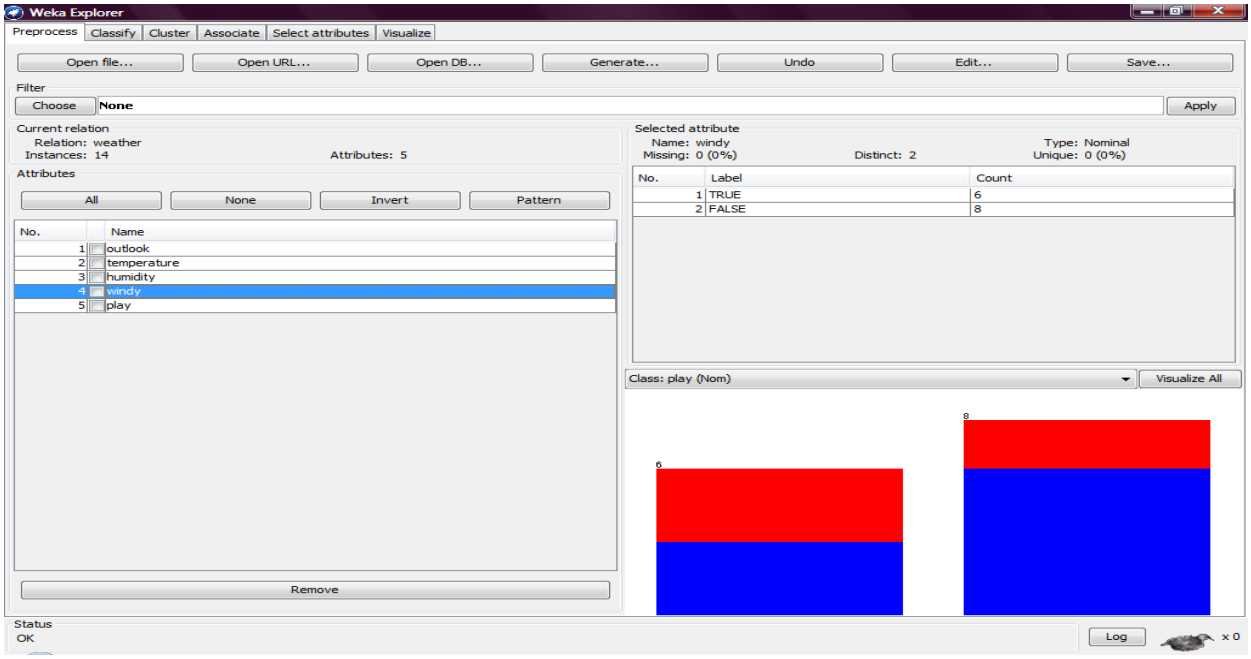
StdDeav η τυπική απόκλιση των δεδομένων=10285



Για την μεταβλητή windy οι τιμές είναι:

True=6

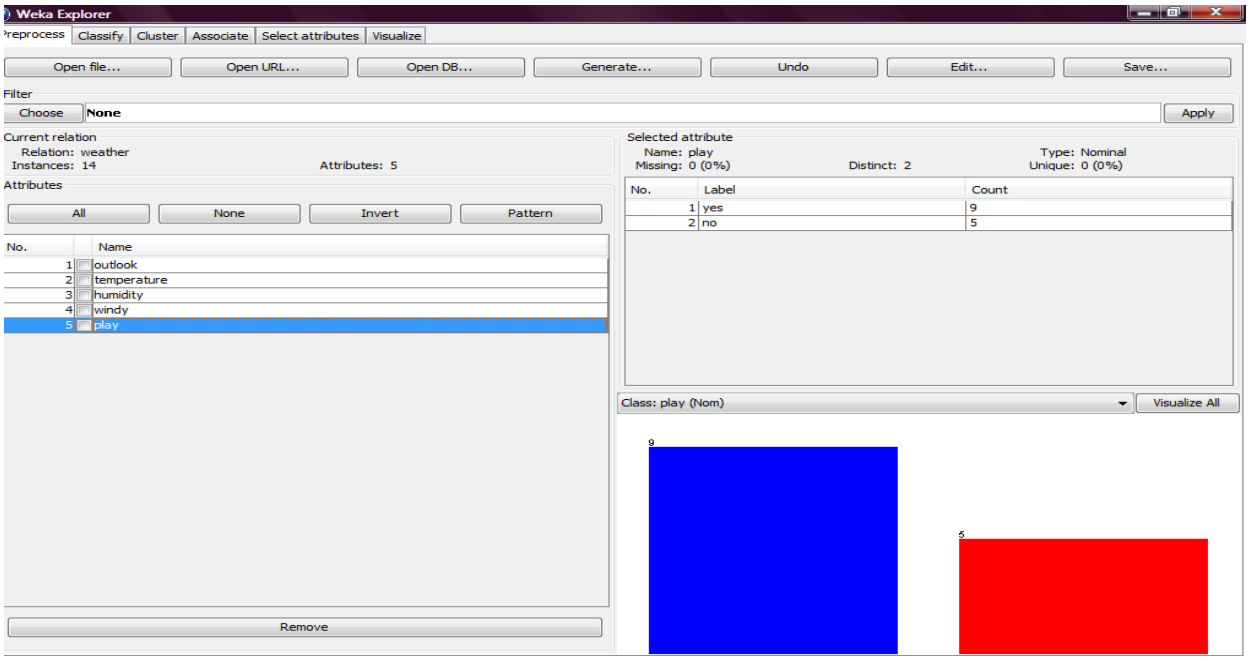
False=8



Για την μεταβλητή play οι τιμές είναι

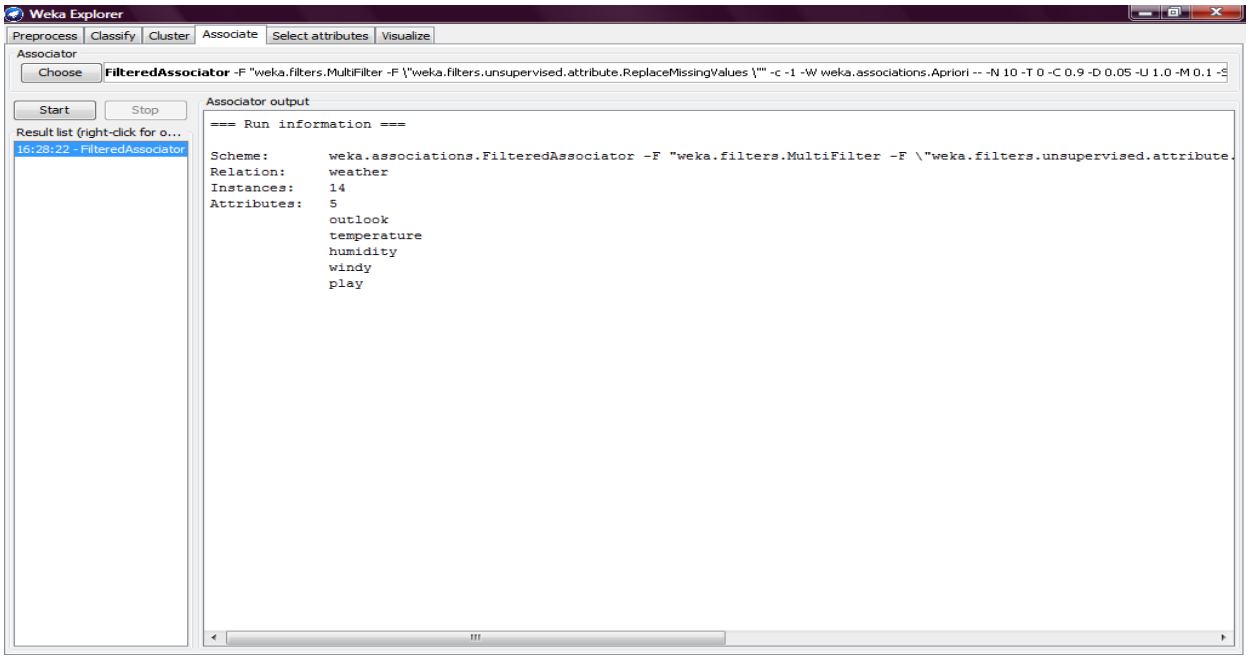
Yes= 9

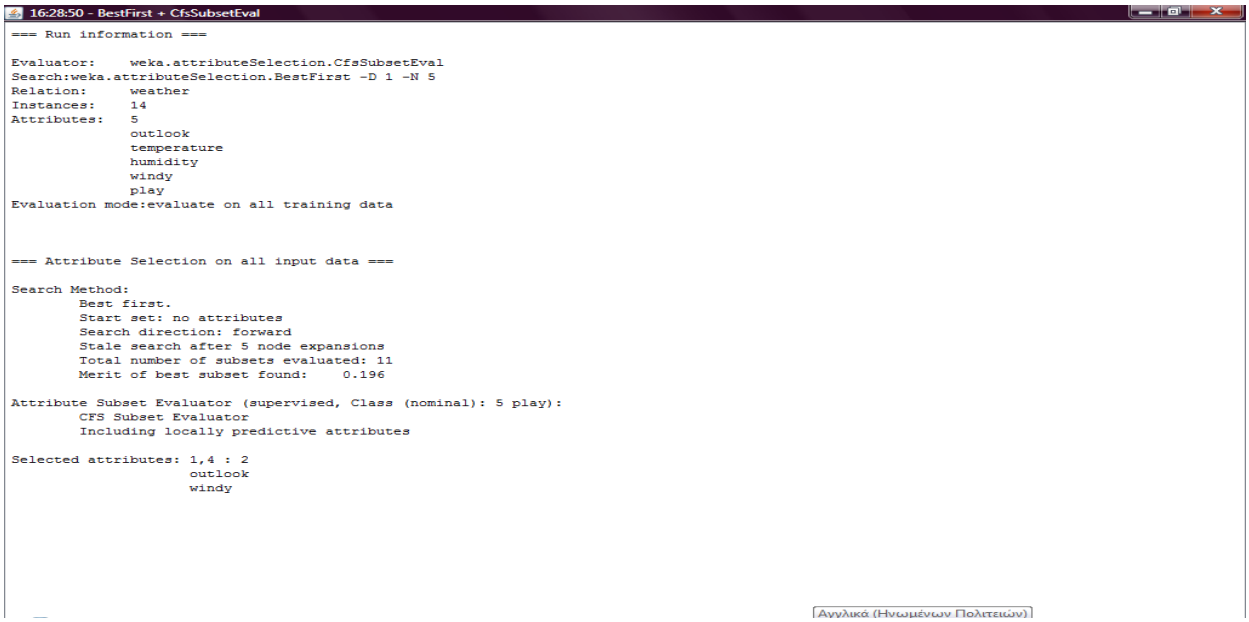
No= 5



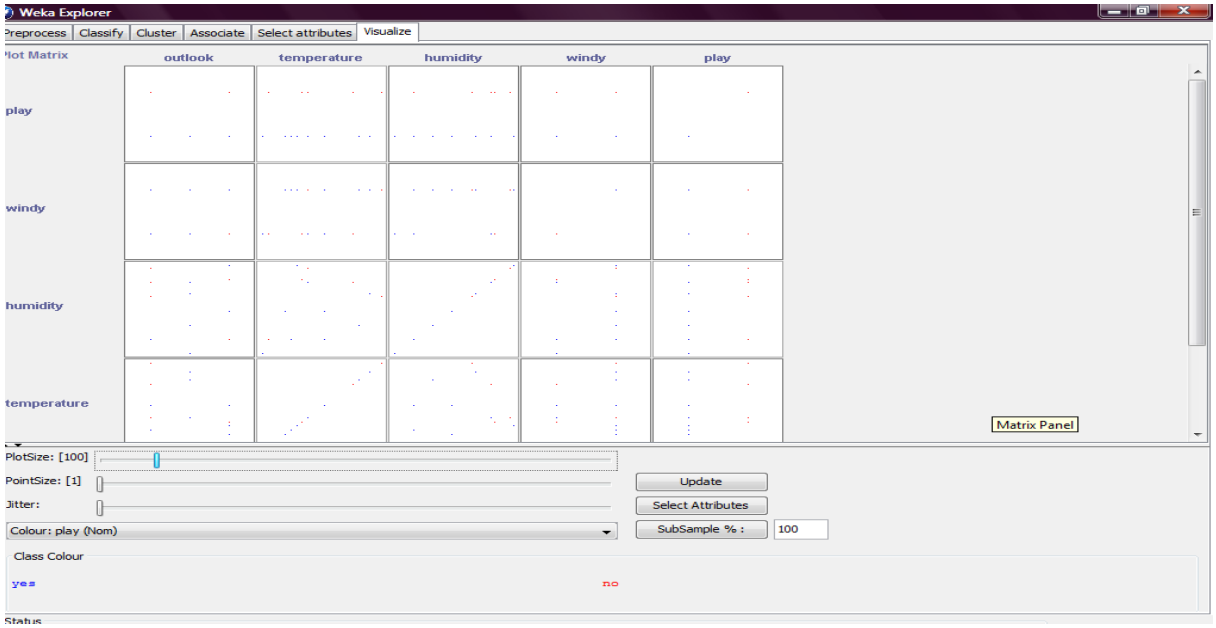
<u>Διαδικασία επιλογής αλγορίθμου</u>

Στην καρτέλα Associate επιλέγοντας τον αλγόριθμο παρατηρούμε ότι σε αυτό το παράδειγμα δεν προκύπτει συσχέτιση μεταξύ των attributes όπως στα δύο προηγούμενα παραδείγματα. Για να γίνει η επιλογή του αλγορίθμου που επιθυμούμε στη προκειμένη περίπτωση χρειαζόμαστε τον αλγόριθμο Apriori. Για την επιλογή του ακολουθούμε τη διαδρομή Choose→ Weka, associations→Apriori και στη συνέχεια πατάμε το κουμπί Start για να τρέξει ο αλγόριθμος για να προκύψει η συσχέτιση μεταξύ των αντικειμένων.

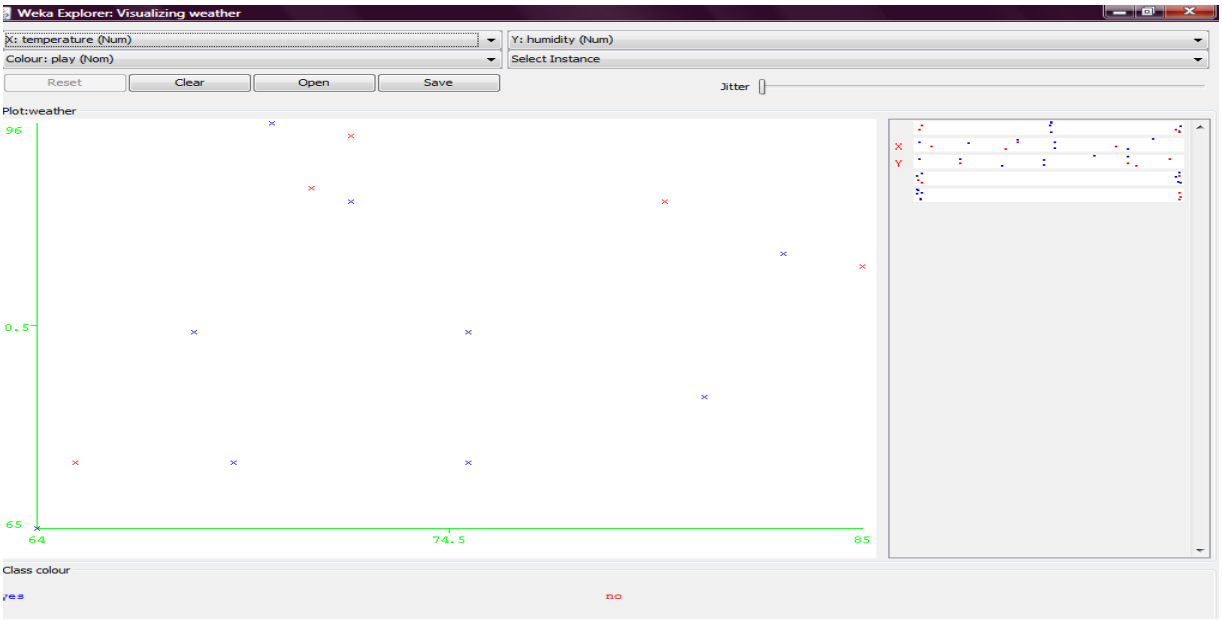

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
FilteredAssociator -F "weka.filters.MultiFilter -F \"weka.filters.unsupervised.attribute.ReplaceMissingValues \"" -c -1 -W weka.associations.Apriori -- -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S
Start
Stop
Result list (right-click for o...
16:28:22 - FilteredAssociator
Associator output
=== Run information ===
Scheme: weka.associations.FilteredAssociator -F "weka.filters.MultiFilter -F \"weka.filters.unsupervised.attribute.
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play


16:28:50 - BestFirst + CfsSubsetEval
=== Run information ===
Evaluator: weka.attributeSelection.CfsSubsetEval
Search:weka.attributeSelection.BestFirst -D 1 -N 5
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Evaluation mode:evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 11
Merit of best subset found: 0.196
Attribute Subset Evaluator (supervised, Class (nominal): 5 play):
CFS Subset Evaluator
Including locally predictive attributes
Selected attributes: 1,4 : 2
outlook
windy

Στην καρτέλα Visualize έχουμε την οπτικοποίηση των δεδομένων σε δισδιάστατα γραφήματα.



Ενδεικτικά για ανεξάρτητη μεταβλητή X η temperature και εξαρτημένη μεταβλητή Y η humidity προκύπτει το παρακάτω διάγραμμα διασποράς και παρατηρούμε πως οι κλάσεις κινούνται από το 65 μέχρι 96. Επίσης παρατηρούμε ότι δεν υπάρχει θετική συσχέτιση μεταξύ της εξαρτημένης και της ανεξάρτητης μεταβλητής, αφού οι τιμές τις μεταβλητής X δεν επηρεάζονται από την Y.



Κεφάλαιο 4

Συμπεράσματα

Η συνεχή ανάπτυξη και βελτίωση τη τεχνολογίας καθώς και η ογκώδη παραγωγή πληροφοριών έχει σαν αποτέλεσμα τα δεδομένα να αποθηκεύονται σε βάσεις δεδομένων. Για το λόγο αυτό δημιουργήθηκε η ανάγκη της διαδικασίας Εξόρυξης Δεδομένων (Data Mining). Ο κύριος λόγος που γίνεται αυτή η διαδικασία είναι για να γίνει εξόρυξη χρήσιμης γνώσης ή προτύπων, να μετατραπούν δηλαδή τα δεδομένα σε πληροφορία που θα είναι χρήσιμη γνώση για μελλοντικές χρήσεις, με τη βοήθεια κάποιων αλγορίθμων, όπου θα βοηθήσει τον άνθρωπο να αναλύσει και να εκτιμήσει τη συμπεριφορά των παραμέτρων για να προβεί στην πιο αποτελεσματική λήψη αποφάσεων. Επιπρόσθετα η διαδικασία εξόρυξης δεδομένων έχει βοηθήσει στην επίλυση πολλών προβλημάτων του πραγματικού κόσμου σε διάφορους τομείς όπως η επιστήμη (όπως βιολογία, χημεία, ιατρική, αστρονομία), οικονομία (όπως είναι χρηματιστηριακές εφαρμογές, πωλήσεις) , ασφάλεια (διάφορες διαδικτυακές και οικονομικές απάτες) , τηλεπικοινωνίες (όπως είναι το φαξ, κινητό, ηλεκτρονικό ταχυδρομείο) και τέλος διαδίκτυο (Google, yahoo).

Η διαδικασία της εξόρυξης γνώσης στηρίζεται σε αλγορίθμους που αναλύθηκαν στο 2 και 3 κεφάλαιο και συνδυάζει στοιχεία από διάφορους τομείς όπως είναι η στατιστική, η οπτικοποίηση, βάσεις δεδομένων, μηχανική μάθηση και τεχνητή νοημοσύνη, για την εξαγωγή πολύτιμης γνώσης που από μόνος του ο χρήστης δεν θα μπορούσε να ανακαλύψει γιατί δεν είναι προφανής. Επίσης στην συνέχεια αναφέρθηκαν οι 4 τεχνικές εξόρυξης δεδομένων οι οποίες είναι: η κατηγοριοποίηση, η συσταδοποίηση, η παλινδρόμηση και οι κανόνες συσχέτισης.

Η κατηγοριοποίηση είναι από τις σημαντικότερες τεχνικές εξόρυξης. Πραγματοποιείται ανάλυση των δεδομένων - μη κατηγοριοποιημένα για να ενταχθούν στις κατάλληλες κλάσεις. Μερικές από τις μεθόδους που χρησιμοποιεί η κατηγοριοποίηση και έχουμε αναλύσει είναι : τα Δέντρα Αποφάσεων, Νευρωνικά Δίκτυα, Αλγόριθμοι διανυσμάτων Υποστήριξης. Ουσιαστικά κατηγοριοποιούνται τυχαία δείγματα με δικά μας δεδομένα και μετέπειτα το μοντέλο συγκρίνει την κατηγορία που ανήκουν τα δεδομένα με την πρόβλεψη που έγινε μέσω του μοντέλου μας. Αν το μοντέλο είναι αξιόπιστο τότε χρησιμοποιείται για την ταξινόμηση μελλοντικών δειγμάτων. Κάτι που μπορεί να επισημανθεί είναι ότι η κατηγοριοποίηση

δεδομένων μπορεί να περιγραφεί σε δυο βήματα αρχικά την Εκμάθηση και στη συνέχεια στη κατηγοριοποίηση.

Η συσταδοποίηση είναι από τις πλέον βασικές εργασίες στην διαδικασία εξόρυξης γνώσης έχοντας στόχο την ανακάλυψη κατανομών ή προτύπων που μπορεί να παρουσιάζουν ενδιαφέρον. Η συσταδοποίηση ουσιαστικά είναι ο διαχωρισμός μιας συστάδας/ ομάδες σε υποδιαίστερες ομάδες/συστάδες ίδιων χαρακτηριστικών/γνωρισμάτων. Πιο αναλυτικά η συσταδοποίηση είναι η ανακάλυψη ομάδων ή αλλιώς οι ονομαζόμενες συστάδες αντικειμένων όπου με την βοήθεια των αλγόριθμων συσταδοποίησης τα δεδομένα χωρίζονται και ταξινομούνται σε κάθε συστάδα ανάλογα με τα όμοια χαρακτηριστικά των συστάδων. Έτσι πετυχαίνουμε και μείωση των μεγάλων όγκων δεδομένων ώστε να αντλήσουμε πιο εύκολα χρήσιμη πληροφορία.

Η παλινδρόμηση είναι μια πιο παλιά τεχνική σε σχέση με τις άλλες. Κυρίως την χρησιμοποιούμε για την μοντελοποίηση και την ανάλυση ποσοτικών δεδομένων γιατί σε κατηγορικά δεδομένα δεν εφαρμόζεται καλά και αποτελεί μια πιο στατιστική τεχνική, αφού αποτελεί μια συνάρτηση συσχέτισης της εξαρτημένης μεταβλητής με τις ανεξάρτητες. Αναφορικά υπάρχει η απλή, γραμμική παλινδρόμηση και η πολλαπλή, γραμμική παλινδρόμηση.

Στη συνέχεια στο Κεφάλαιο 2 αναφέρθηκε πιο αναλυτικά η τέταρτη τεχνική, οι κανόνες συσχέτισης, οι οποίοι είναι χρήσιμοι γιατί αποτελεί μια σύγχρονη μέθοδος για την εξαγωγή χρήσιμης γνώσης. Για αυτό άλλωστε τον λόγο έχει δημιουργηθεί μια πλειάδα αλγορίθμων που παράγουν κανόνες συσχέτισης και πρότυπα με πολλαπλές εφαρμογές. Με τους κανόνες συσχέτισης μπορούν να αποκαλυφθούν κρυμμένες συσχετίσεις μέσα σε ένα σύνολο δεδομένων και να εξάγει τις αγοραστικές τάσεις των καταναλωτών. Οι κανόνες αυτοί αναφορικά για να γίνει πιο κατανοητική η εφαρμογή, σε ένα πολυκατάστημα μπορεί να χρησιμοποιηθεί στην προώθηση προϊόντων, την τοποθέτηση προϊόντων στα ράφια καταστημάτων και την διαχείριση αποθεμάτων.

Ένας αλγόριθμος που αντιπροσωπεύει τη διαδικασία εξαγωγής κανόνων συσχέτισης είναι ο Apriori, ο οποίος έχει αναλυθεί στο Κεφάλαιο 2 και είναι ένας αποτελεσματικός αλγόριθμος γιατί μπορεί να δώσει λύσεις στις μεγάλες λίστες που προκύπτουν. Είναι ένας αλγόριθμος που μπορεί η διαδικασία του να επαναληφθεί όσες φορές χρειαστεί στη βάση δεδομένων για να εξάγει χρήσιμες πληροφορίες. Ένας αλγόριθμος, ο οποίος είναι μια παραλλαγή του Apriori είναι ο AprioriTID. Στον αλγόριθμο AprioriTID εξάγονται αποτελέσματα σε μεταγενέστερες

επαναλήψεις σε σύγκριση με το Apriori που δίνει αποτελέσματα από τις πρώτες κιόλας επαναλήψεις και ο οποίος βασίζεται στη φιλοσοφία του Apriori με μόνη διαφορά ότι ο αρχικός πίνακας δοσοληψιών D διαβάζεται μόνο μια φορά και αυτό είναι στην αρχή. Η διαφοροποίηση αυτή έχει στόχο να αυξήσει την αποδοτικότητα της διαδικασίας της εξόρυξης κανόνων συσχέτισης.

Ένα πρόγραμμα που μπορούμε να βρούμε τον αλγόριθμο Apriori αλλά και άλλους αλγόριθμους είναι το Weka. Το Weka είναι ένα πρόγραμμα που μελετάει μεγάλες βάσεις δεδομένων, αναπαριστά γραφικά τα σύνολα δεδομένων και εξάγει χρήσιμα συμπεράσματα με βάσει τις παραμέτρους που βάζει ο χρήστης. Το Weka αποτελείται από πολλούς σύγχρονους αλγορίθμους μηχανικής μάθησης, όπου ο κάθε αλγόριθμος κάνει και μια διαφορετική λειτουργία. Επίσης περιέχει και μεθόδους προεπεξεργασίας δεδομένων, οπτικοποίηση με δισδιάστατα γραφήματα, ταξινόμηση, συσταδοποίηση, παλινδρόμηση και εύρεση κανόνων συσχέτισης. Το συμπέρασμα είναι ότι το πρόγραμμα Weka είναι ένα ισχυρό πακέτο με εναλλακτικές χρήσεις που επεξεργάζεται μεγάλους όγκους δεδομένων γρήγορα με αποτέλεσμα να εξάγει χρήσιμες πληροφορίες ανάλογα με τι κριτήρια έχει επιλέξει ο κάθε χρήστης, αφού τα σύνολα των δεδομένων ποικίλουν. Είναι σημαντικό να αναφέρουμε και σε αυτό το σημείο ότι κανένας αλγόριθμος δεν ταιριάζει για όλα τα σύνολα δεδομένων και για αυτό το λόγο το πρόγραμμα weka, το οποίο περιέχει αρκετούς αλγόριθμους αποτελεί ένα από τα σημαντικότερα και ισχυρότερα εργαλεία για την εξόρυξη δεδομένων.

Ελληνική Βιβλιογραφία:

1. Καλούδη Β. Σφυρογιαννάκη Ε. Τσουρδιου Α., Πάτρα 2009, «Αλγόριθμοι Ομαδοποίησης στην Εξόρυξη Δεδομένων», Πτυχιακή εργασία.
2. Μαλέσκου Π., Κουμπή Ε., Πάτρα 2012, «Οι Αλγόριθμοι της εξόρυξης δεδομένων: Περιγραφή, Εφαρμογές, Προοπτικές», Πτυχιακή εργασία.
3. Χαλκίδη Μ. & Βαζιργιάννης Μ. «Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό».
4. Γεράσιμος Ε. Σταυλίωτης, Πανεπιστήμιο Πειραιώς, Εξόρυξη δεδομένων(Data Mining) αναγνώριση προτύπων σε κατηγορίες δεδομένων μέσω συσταδοποίησης.
5. Πολύδωρου Καμπακτσή, Θεσσαλονίκη 2004, Χρήση τεχνικών εξόρυξης δεδομένων για την πρόβλεψη ενεργειών σε επίπεδο συμπεριφοράς πρακτόρων, Διπλωματική Εργασία.
6. Μποφιλάκη Αικατερίνης, ΘΕΣΣΑΛΟΝΙΚΗ 2007, «Εξόρυξη Δεδομένων στον Παγκόσμιο Ιστό βασισμένη στη Χρήση», Πτυχιακή εργασία.
7. Γουρδούλης Ιωάννης-Πρόδρομος, Πάτρα, «Αλγόριθμοι Εξόρυξης δεδομένων για χειρισμό πολλαπλών υποστηρίξεων και αρνητικών συσχετίσεων», Διπλωματική Εργασία.
8. Λίνα Μάσσου, Εθνικό Μετσόβιο Πολυτεχνείο 2008, «Αλγόριθμοι Εξόρυξης Πληροφορίας», Εργασία, Διαθέσιμο: http://dataminingntua.files.wordpress.com/2008/05/cea4ce95ce9bce99ce9ace97_ce95cea1ce93ce91cea3ce99ce911.pdf
9. Μηνάς Καραολής, Πανεπιστήμιο Κύπρου 2010, «ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΜΕ ΕΞΑΓΩΓΗ ΚΑΝΟΝΩΝ ΣΕ ΚΑΡΔΙΑΓΓΕΙΑΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ».
10. Διαφάνειες που στηρίζονται στο P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining», Addison Wesley, 2006, «Κανόνες Συσχέτισης Ι» & «Κανόνες Συσχέτισης ΙΙ», Εργασία.
11. Ανδρέας Λ. Συμεωνίδης, Θεσσαλονίκη 2011-2012 «Αναγνώριση Προτύπων Ομαδοποίηση, Ταξινόμηση, Παλινδρόμηση», Εργασία.
12. Παύλος Εφραιμίδης, Θεσσαλονίκη, «Εξόρυξη Δεδομένων από Βάσεις Δεδομένων», Εργασία, Διαθέσιμο: http://multimine.iti.gr/seminar5%20presentations/thewritiko_meros/DPTh_Efremidis_DataMining.pdf

13. Τζιραλής Γεώργιος, ΕΜΠ ΜΜ ΒΔΕΕ 2007, «Εισαγωγή στο Data Mining από τα δεδομένα στη γνώση», Εργασία.
14. Βλαχάβας Ι., Κεφαλάς Π., Βασιλειάδης Ν., Κόκκορας Φ., Σακελλαρίου Η., Τεχνητή Νοημοσύνη-Β΄Έκδοση, Μηχανική Μάθηση-Κεφάλαιο 18.
15. Μακρής Αριστομένης, Εξόρυξη Δεδομενων - Data Mining, Εργασία, Διαθέσιμο: http://amacris.ode.unipi.gr/DST/07_DST_DM.pdf
16. Ιστοσελίδα: Πως εφαρμόζετε η εξόρυξη δεδομένων, Διαθέσιμο: http://www.trainmor-knowmore.eu/F5048108.el.aspx
17. ΝΙΚΟΛΑΟΣ Χ. ΤΣΙΡΑΚΗΣ, Πάτρα 2006, «Αλγόριθμοι και τεχνικές εξόρυξης δεδομένων από ροές δεδομένων στο παγκόσμιο ιστό», Μεταπτυχιακή εργασία, Διαθέσιμο: http://nemertes.lis.upatras.gr/jspui/bitstream/10889/542/1/Nimertis_Tsirakis.pdf
18. Παρουσίαση εξόρυξης δεδομένων, διαθέσιμο: http://www.cs.uoi.gr/~pitoura/courses/dm/introspring11.pdf
19. Ουγιάρογλου Στέφανου, Θεσσαλονίκη 2006, «Κατηγοριοποίηση με βάση δυναμικό αριθμό κοντινότερων γειτόνων», Διπλωματική εργασία, διαθέσιμο: http://users.sch.gr/stoug/papers/final_work_msc.pdf
20. Κουρής Ν. Γίαννης, Πάτρα 2006, «Εφαρμογή Τεχνικών Data Mining σε συστήματα Ηλεκτρονικού Εμπορίου», Διδακτορική Διατριβή, διαθέσιμο: http://nemertes.lis.upatras.gr/jspui/bitstream/10889/1610/1/Nimertis_Kouris.pdf
21. Δέσποινα, Ι. Κουκουλά, Αθήνα 2008, «Αλγόριθμοι Συσταδοποίησης Ενεργειακών Καταναλώσεων μέσω Διαδικτύου, για παροχή Ηλεκτρονικών Ενεργειακών Υπηρεσιών από ESCOs», Διπλωματική Εργασία, Διαθέσιμο: http://artemis.cslab.ntua.gr/el_thesis/artemis.ntua.ece/DT2008-0206/DT2008-0206.pdf
22. Addison Wesley, Συσταδοποίηση-2006, «Introduction to Data Mining», διαθέσιμο: http://www.cs.uoi.gr/~pitoura/courses/dm/cluster1-11.pdf
23. Βικιπαιδεία

- ✓ http://el.wikipedia.org/wiki/%CE%9A%CE%B1%CF%84%CE%B7%CE%B3%CE%BF%CF%81%CE%B9%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7
- ✓ http://en.wikipedia.org/wiki/Naive_Bayes_classifier
- ✓ http://el.wikipedia.org/wiki/%CE%9D%CE%B5%CF%85%CF%81%CF%89%CE%BD%CE%B9%CE%BA%CF%8C_%CE%B4%CE%AF%CE%BA%CF%84%CF%85%CE%BF
- ✓ http://en.wikipedia.org/wiki/Apriori_algorithm
- ✓ http://en.wikipedia.org/wiki/Weka_(machine_learning)

Διεθνής Βιβλιογραφία:

1. Ramakrishnan Srikant & Rakesh Agrawal, Mining Generalized Association Rules.
2. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases.
3. Machine Learning Group at the University of Waikato, Weka 3: Data Mining Software in Java, Διαθέσιμο: http://www.cs.waikato.ac.nz/ml/weka/
4. DePaul University 2005, Association Rule Mining with WEKA, Διαθέσιμο: http://maya.cs.depaul.edu/~Classes/Ect584/Weka/associate.html
5. BMI-Data mining with WEKA, Part 2: Classification and clustering, Διαθέσιμο: http://www.ibm.com/developerworks/opensource/library/os-weka2/index.html
6. Rakesh Agrawal & Ramakrishnan Srikant, «Fast Algorithms for Mining Association Rules», Διαθέσιμο: http://www.cs.ubc.ca/~rap/teaching/504/2005/slides/association-rules.pdf
7. Anita Wasilewska, APRIORI Algorithm - Lecture Notes, διαθέσιμο: http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf
8. Andrew Kusiak, The Apriori Algorithm, διαθέσιμο: http://www.engineering.uiowa.edu/~comp/Public/Apriori.pdf
9. Γρήγορη Αλγόριθμοι για Κανόνες συσχέτισης, διαθέσιμο: http://www.eecs.berkeley.edu/~fox/summaries/database/fast_mining.html