2015

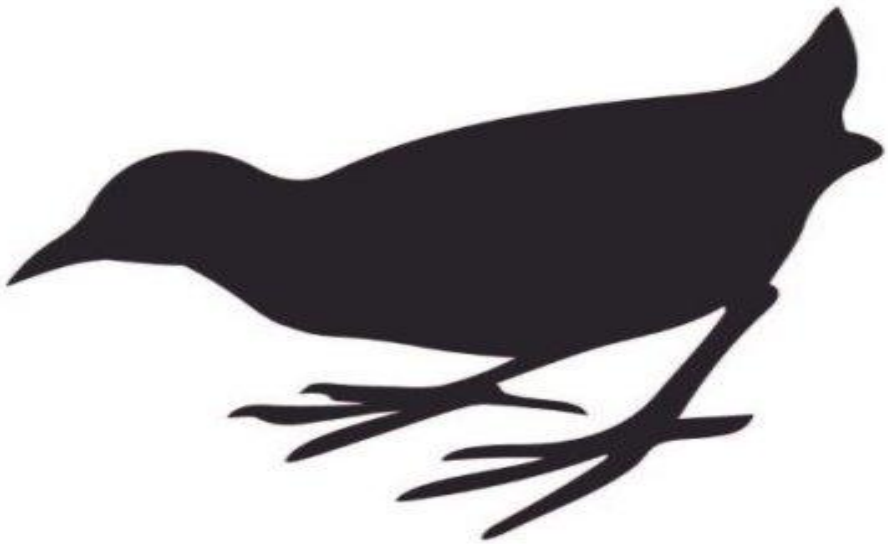
ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:

ΛΟΓΙΣΜΙΚΟ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ WEKA: ΑΝΑΛΥΤΙΚΟ ΕΓΧΕΙΡΙΔΙΟ ΧΡΗΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ



WEKA

ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΣΠΟΥΔΑΣΤΩΝ: ΠΕΠΙΝΙΔΗΣ ΣΤΥΛΙΑΝΟΣ

ΑΝΑΣΤΑΣΙΟΣ ΜΑΤΘΑΙΟΣ ΛΑΖΑΡΟΥ

ΙΩΑΝΝΗΣ ΠΑΝΑΓΙΩΤΗΣ ΧΑΤΖΗΔΑΚΗΣ

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ: ΜΠΑΚΑΛΗΣ ΑΡΗΣ

ΠΑΤΡΑ 2015

«Οι συγγραφείς βεβαιώνουν ότι το περιεχόμενο του παρόντος έργου είναι αποτέλεσμα προσωπικής εργασίας και ότι έχει γίνει η κατάλληλη αναφορά στην εργασία τρίτων, όπου κάτι τέτοιο ήταν απαραίτητο, σύμφωνα με τους κανόνες της ακαδημαϊκής δεοντολογίας.»

ΕΥΧΑΡΙΣΤΙΕΣ- ΑΦΙΕΡΩΣΕΙΣ

Ευχαριστούμε ιδιαίτερα τον Επιβλέποντα καθηγητή μας, κ.Μπακάλη, του οποίου η εμπειρία και οι γνώσεις στο χώρο της θεωρίας και των εφαρμογών της Πληροφορικής, μας βοήθησαν να διεξέλθουμε στην έρευνα αυτή. Ο χρόνος που αφιέρωσε με πολύωρες συζητήσεις για την πτυχιακή, στάθηκε το δημιουργικό εφαλτήριο για τη διατύπωση των κρίσιμων ζητημάτων της έρευνας και τη συγκρότηση θεωρητικού λόγου γύρω από τα καίρια ζητήματά της.

ΠΕΡΙΛΗΨΗ

Στην πτυχιακή εργασία ασχοληθήκαμε με την μελέτη των τεχνικών και των αλγορίθμων που χρησιμοποιούνται κατά την εξόρυξη δεδομένων σε ένα ευρύ φάσμα τομέων όπου υπάρχει σημαντική χρησιμότητα. Η επιστήμη της εξόρυξης γνώσης δημιουργήθηκε τα τελευταία 25 χρόνια δίνοντας λύση στην αξιοποίηση του ταχέως αυξανόμενου όγκου πληροφορίας σε βάσεις δεδομένων και το διαδικτύου. Ταυτόχρονα με αυτό φανέρωσε την γνώση που μπορεί να ωφελήσει επιχειρήσεις, υπηρεσίες ακόμα και επιστήμες.

Η εργασία είναι χωρισμένη σε δύο μέρη. Το Α μέρος περιέχει τα κεφάλαια 1) Εισαγωγή στην θεωρία της εξόρυξης δεδομένων, 2) Εξόρυξη Δεδομένων, 3) Παγκόσμιος ιστός, 4) Εξόρυξη για εκπαιδευτικούς σκοπούς. Στο πρώτο κεφάλαιο γίνεται η εισαγωγή στο θέμα της εργασίας παραθέτοντας ορισμούς και πληροφορίες που χρειάζονται για την καλύτερη κατανόηση της περεταίρω ανάλυσης που ακολουθεί.

Στο δεύτερο κεφάλαιο γίνεται η ανάλυση της κατηγοριοποίησης, της συσταδοποίησης, των κανόνων συσχέτισης και της ανάλυση ακολουθιών που είναι βασικές τεχνικές εξόρυξης από βάσης δεδομένων καθώς και των εργασιών που γίνονται ανάλογα με τον τύπο αλγορίθμου που εφαρμόζεται.

Στο τρίτο κεφάλαιο έχουμε άλλον έναν βασικό τομέα εξόρυξης δεδομένων. Αυτός είναι ο παγκόσμιος ιστός. Οι πληροφορίες που αντλούνται εδώ μπορεί να είναι από υπερσυνδέσμους οι οποίοι είναι μέρος της δομής του ιστού (Δεδομένα Δομής), από τις προτιμήσεις που έχουν οι χρήστες, σύμφωνα με τις περιηγήσεις (Δεδομένα Χρήσης) και από τα περιεχόμενα των σελίδων (Δεδομένα Περιεχομένου). Το κεφάλαιο αυτό περιέχει την ανάλυση της λειτουργίας αλγορίθμων εξόρυξης δεδομένων παγκόσμιου ιστού καθώς και τις τεχνικές που χρησιμοποιούνται ανάλογα με την κατηγορία των δεδομένων.

Στο τέταρτο κεφάλαιο έχουμε την συνοπτική αναφορά στην εξόρυξη δεδομένων στα εκπαιδευτικά δεδομένα όπου παρουσιάζεται η σημασία και οι δυνατότητες που δημιουργούνται με αυτή την τεχνική στον εκπαιδευτικό τομέα.

Τέλος στο Β μέρος της πτυχιακής εργασίας υπάρχει αναλυτικό εγχειρίδιο χρήσης του προγράμματος WEKA που χρησιμοποιείται για την εξόρυξη δεδομένων.

ΛΕΞΕΙΣ-ΚΛΕΙΔΙΑ

ΛΟΓΙΣΜΙΚΟ, ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ, ΕΓΧΕΙΡΙΔΙΟ ΧΡΗΣΗΣ WEKA,ΕΦΑΡΜΟΓΕΣ

ΠΊΝΑΚΑΣ ΠΕΡΙΕΧΟΜΈΝΩΝ

Περιεχόμενα

Ά ΜΕΡΟΣ

ΚΕΦΑΛΑΙΟ 1.

Εισαγωγή στην θεωρία της εξόρυξης δεδομένων

1.1 Εισαγωγή

Η πληροφορία είναι ένας από τους πιο χρήσιμους πόρους των επιχειρήσεων διότι, τους δίνει την δυνατότητα της γνώσης και της πρόβλεψης. Από τα επιχειρησιακά δεδομένα που βρίσκονται στις βάσεις δεδομένων και στα πληροφοριακά συστήματα όπως είναι το ERP, συνήθως αξιοποιείτε από τις επιχειρήσεις ένα μόνο μέρος από τον μεγάλο όγκο πληροφορίας που δημιουργείται εκεί συνεχώς. Αυτό συμβαίνει επειδή δεν μπορεί να αντληθεί εύκολα η γνώση στις περιπτώσεις που ο χρήστης δεν γνωρίζει την δομή και την σημασία των τιμών που εμφανίζονται στα δεδομένα ώστε να μπορούν να γίνουν στοχευμένες ερωτήσεις όπως γίνετε στην στατιστική.

Η εξόρυξη γνώσης αποκαλύπτει αυτή την κριμένη γνώση, καθώς με την χρήση αλγορίθμων γίνετε ο εντοπισμός προτύπων και ο κανονισμός των δεδομένων, φτιάχνοντας έτσι μοντέλα προβλέψεων και συσχετίσεων που εξηγούν τις αλληλεπιδράσεις ανάμεσα στους παράγοντες που παίζουν ρόλο για να επιτευχθούν οι στόχοι τον επιχειρήσεων.

1.2 Τι είναι data mining-Ορισμός

Εξόρυξη δεδομένων η αλλιώς data mining ονομάζεται η σύνθετη διαδικασία εξαγωγής συγκεκριμένης, μη προφανής, άγνωστης μέχρι τώρα και δυνητικά ωφέλιμης γνώσης από μεγάλες βάσεις δεδομένων. Για να επιτευχθεί αυτό γίνετε χρήση αλγορίθμων ομαδοποίησης, κατηγοριοποίησης καθώς και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Εναλλακτικά, θεωρείται και ως η επιστήμη της εξόρυξης χρήσιμης πληροφορίας από σύνολα ή βάσεις δεδομένων μεγάλου μεγέθους. Είναι ένα σημαντικό εργαλείο, το οποίο βοηθάει τον άνθρωπο μέσα από την διαδικασία της εξερεύνησης και της ανάλυσης πολλών δεδομένων με αυτόματα ή ημιαυτόματα μέσα.

Στην διαχείριση επιχειρηματικών πόρων (ERP), το data mining θεωρείται ως η στατιστική και λογική ανάλυση εκτεταμένων συνόλων από δεδομένα συναλλαγών και εργασιών για τον εντοπισμό επαναλαμβανόμενων μοτίβων ή τάσεων προκειμένου να βοηθήσουν στην λήψη αποφάσεων (Ellen Monk, Bret Wagner, 2006). Οι νέες πληροφορίες που προκύπτουν μπορούν να χρησιμοποιηθούν σε διάφορους τομείς όπως για παράδειγμα στην υποστήριξη της λήψης αποφάσεων, στις προβλέψεις και στις εκτιμήσεις σημαντικών επιχειρηματικών αποφάσεων. Γενικά έχουν χρησιμότητα σε τομείς οι οποίοι μπορούν να βοηθήσουν μια

επιχείρηση να αποκτήσει και να διατηρήσει σημαντικό ανταγωνιστικό πλεονέκτημα (Ahmed,S.R.2004).

1.3 Αίτια δημιουργίας του data mining

Η συνεχής πρόοδος της τεχνολογίας στο τομέα της πληροφορικής, παρέχει την δυνατότητα αποθήκευσης τεράστιου όγκου δεδομένων, σε αρχεία, βάσεις δεδομένων, το διαδίκτυο και άλλα μέσα. Οι περισσότερες επιχειρήσεις πλέον χρησιμοποιούν την δυνατότητα αυτή και καταγράφουν το μεγαλύτερο πλήθος των πληροφοριών τους σε ηλεκτρονική μορφή. Αυτό έχει σαν αποτέλεσμα τον διπλασιασμό του όγκου αποθηκευμένων δεδομένων κάθε 3 χρόνια. (Μαστρογιάννης,2009) Έτσι δημιουργήθηκε η ανάγκη για την ανάλυση και την ερμηνεία της σημαντικής πληροφορίας που υπάρχει στις αποθήκες δεδομένων (data warehouse) των επιχειρήσεων.

1.4 Βασικός στόχος

Στόχος της εξόρυξης δεδομένων είναι να εξαχθεί πληροφορία η οποία θα βοηθήσει να παρθούν κατάλληλες αποφάσεις.(Βικιπαίδεια,2015). Για να γίνει αυτό χρειάζεται να υπάρξει η περιγραφή και η πρόβλεψη στα σύνολα δεδομένων. Σκοπός της πρόβλεψης είναι ο υπολογισμός της μελλοντικής αξίας ή συμπεριφοράς των μεταβλητών που μας ενδιαφέρουν και εξαρτούνται από τη συμπεριφορά άλλων μεταβλητών. Σκοπός της περιγραφής είναι η ανακάλυψη προτύπων και η αναπαράσταση των δεδομένων μια πολύπλοκης βάσης δεδομένων με κατανοητό και αξιοποιήσιμο τρόπο. Ανάλογα με τις εφαρμογές εξόρυξης διαφοροποιείται η σημαντικότητα αυτών των δυο. Η περιγραφή είναι πιο σημαντική από τη πρόβλεψη όσων αφορά την εξόρυξη γνώσης, ενώ η πρόβλεψη είναι πιο σημαντική για την αναγνώριση προτύπων και την εφαρμογή μηχανικής μάθησης.(Han,J. & Kamber,M.2006)

1.5 Εφαρμογές της εξόρυξης δεδομένων

Μία από τις χρήσιμες εφαρμογές του data mining είναι να θέτει τα δεδομένα των επιχειρήσεων σαν αρχικούς πόρους και χρησιμοποιώντας προκαθορισμένους αλγόριθμους, να ομαδοποιεί τις τεράστιες ποσότητες αυτών σύμφωνα με τα κριτήρια που επιθυμεί ο εκάστοτε χρήστης ώστε να μπορεί να του είναι χρήσιμα για μελλοντικό μάρκετινγκ και την ανάπτυξη στρατηγικών προώθησης προϊόντων και υπηρεσιών. Μέσα από τις εφαρμογές των τεχνικών εξόρυξης δεδομένων μπορεί μια μεγάλη επιχείρηση να μετατρέψει τις χιλιάδες εγγραφές στις βάσεις δεδομένων των πελατών της σε κάποια συνεκτικού είδους εικόνα για τους πελάτες της. (Chen, 1996).

Μερικοί ακόμα τομείς που υπάρχει εφαρμογή της εξόρυξης δεδομένων είναι:

· Στην τηλεπικοινωνία η εξόρυξη δεδομένων βοηθάει στην διάκριση τηλεπικοινωνιακών προτύπων καταπολέμησης παράνομων δραστηριοτήτων και στην καλύτερη χρήση των πόρων καθώς και στην βελτίωση της ποιότητας των υπηρεσιών. (Βικιπαίδεια,2015)
· Στην αναζήτηση προτύπων (pattern recognition) σε διάφορα προβλήματα τεχνητής νοημοσύνης.
· Στους τομείς της Βιοτεχνολογίας, της Γενετικής και της Ιατρικής έρευνας.
· Στην ανάλυση εικόνας.
· Στην αστρονομία .
· Και σε κάθε τομέα ο οποίος έχει σαν στόχο την αναζήτηση γνώσης. (Γολέμη,Ε.2010.σελ.65)

1.6 Ιστορικά στοιχεία

Αν και ο όρος εξόρυξη δεδομένων εισήχθη το 1990, η έννοια της εξόρυξης δεδομένων έχει τις ρίζες της εδώ και πολλά χρόνια. Η εξόρυξη δεδομένων έφθασε στην σημερινή της μορφή, αφού πρώτα πέρασε από διάφορες φάσεις έρευνας και μελετών. Η ανάπτυξη αυτή ξεκίνησε όταν τα δεδομένα των επιχειρήσεων άρχισαν να αποθηκεύονται στους υπολογιστές. Η διαδικασία συνεχίστηκε με τις προόδους στην τεχνολογία των υπολογιστών συμπεριλαμβανομένης και της αποθήκευσης δεδομένων, της επεξεργαστική ισχύεις, των νέων λογισμικών, των αλγορίθμων κλπ. Καθοριστικό επίσης ήταν για την εξέλιξη της, ότι στο σημερινό ανταγωνιστικό κόσμο των πληροφοριών, όλοι προσπαθούν να κάνουν την καλύτερη χρήση των δεδομένων τους για να κάνουν τις επιχειρήσεις τους να έχουν κέρδος και επιτυχία.

Η συλλογή και αποθήκευση δεδομένων σε υπολογιστές, ταινίες και δίσκους ξεκίνησε το 1960. Το επόμενο εξελικτικό βήμα στην εξόρυξη δεδομένων συνέβη κατά τη διάρκεια της δεκαετίας του 1980 με την εισαγωγή των σχεσιακών βάσεων δεδομένων και τη δομημένη γλώσσα ερωτημάτων. Αυτό βοήθησε τους χρήστες να μάθουν περισσότερα σχετικά με τα δεδομένα που αποθηκεύονται σε σχεσιακές βάσεις δεδομένων και χρησιμοποιούν δομημένη

γλώσσα με ερωτήματα. Έτσι, τα δεδομένα έγιναν διαθέσιμα σε επίπεδο ρεκόρ για την τότε εποχή. Στην συνέχεια η εξέλιξη της αποθήκευσης δεδομένων που συνέβη κατά τη διάρκεια της δεκαετίας του 1990 όπου γίνετε η αναλυτική επεξεργασία και οι πολυδιάστατες βάσεις δεδομένων, συνέβαλαν στην αύξηση της αποθήκευσης δεδομένων.

Εάν αναλύσουμε κάθε βήμα της εξέλιξης, είναι πολύ σαφές ότι κάθε βήμα βασίζεται στο προηγούμενο βήμα. Ένα ερώτημα των επιχειρήσεων που θα μπορούσε να απαντηθεί κατά τη διάρκεια του αρχικού σταδίου ήταν "πόσο είναι το σύνολο των εσόδων μου τα τελευταία 3 χρόνια;". Αλλά στο ερώτημα "Τι είναι πιθανό να συμβεί με τις πωλήσεις μου τον επόμενο μήνα και γιατί;" απαντάται πλέον με τεχνικές εξόρυξης δεδομένων.

Κατά τη διάρκεια της δεκαετίας του 1960 τα δεδομένα δεν ήταν κάτι σύνθετο. Πλέον όμως η κατάσταση είναι εντελώς διαφορετική. Τα επιχειρηματικά δεδομένα έχουν μετατραπεί σε επιχειρηματικές πληροφορίες και είναι αρκετά ισχυρά ώστε να απαντήσουν σε πολλές πολύπλοκες επιχειρηματικές ερωτήσεις, ακόμη και να προβλέψουν το μέλλον της επιχείρησης. Η ανάπτυξη των βάσεων δεδομένων συμβαίνει σε τεράστια ποσοστά που απαιτούν μεθόδους για να παρέχουν χρήσιμες πληροφορίες από αυτό το τεράστιο όγκο δεδομένων. Η τεχνολογία εξόρυξης δεδομένων έχει περάσει με την πάροδο των χρόνων από την ανάπτυξη της διαδικασίας σε τρεις διαφορετικούς τομείς που συνέβαλαν στην τρέχουσα μορφή της. Οι τομείς αυτοί είναι οι στατιστικές, η τεχνητή νοημοσύνη και η μηχανική μάθηση.

Το πρώτο εργαστήριο IJCAI για Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων διεξήχθη το 1989 στο Detroit MI, Ηνωμένες Πολιτείες της Αμερικής. IJCAI σημαίνει Διεθνές κοινό Συνέδριο Τεχνητής Νοημοσύνης. Κατά τη διάρκεια του 1991-1994, εργαστήρια για την KDD συζήτησαν για την πρόοδο της ανακάλυψης της γνώσης και εξόρυξης δεδομένων. Από το 1998, έγιναν διεθνείς διασκέψεις για την KDD και την εξόρυξη δεδομένων. Από το 2001, διεξάγονται κάθε χρονιά τα IEEE ICDM (International Conference on Data Mining) και SIAM-DM (conference on discrete mathematics). Εκτός από αυτά, πολλά περιφερειακά συνέδρια, συμπεριλαμβανομένων PAKDD(Pacific-Asia Conference in Knowledge Discovery and Data Mining), ECML/PKDD(European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases) κλπ είναι επίσης σε εξέλιξη.

Σημαντικά περιοδικά στην ιστορία της εξόρυξης δεδομένων είναι: Data Mining and Knowledge Discovery που κυκλοφόρησε το 1997, Knowledge and Information Systems το 1999, IEEE Transactions on Knowledge and Engineering Data, TAPMI, ML, IDA κλπ. Αν και η εξόρυξη δεδομένων θα μπορούσε να λέγεται ότι είναι 10 έως 15 ετών, οι μελέτες συνεχίζονται για να κάνουν την εξόρυξη δεδομένων μια καλύτερη ιδέα. Οι νέες μέθοδοι εξόρυξης δεδομένων που εισάγονται κάνουν την όλη διαδικασία εξόρυξης δεδομένων ολοένα και περισσότερο αποτελεσματική.(History of Data Mining.2015)

1.7 Επιστημονικοί τομείς σε σχέση με το data mining

Η στατιστική είναι το θεμέλιο των περισσότερων τεχνολογιών στις οποίες η εξόρυξη δεδομένων είναι “χτισμένη”, π.χ. ανάλυση παλινδρόμησης, τυπική κατανομή, τυπική απόκλιση, διακριτική ανάλυση, ανάλυση διασποράς, διαστήματα εμπιστοσύνης κτλ. Όλα τα

παραπάνω χρησιμοποιούνται για τη μελέτη των σχέσεων που υπάρχουν στα δεδομένα. Η στατιστική έχει συμβάλει σε μεγάλο βαθμό στην επιχειρηματική ευφυΐα των τελευταίων δεκαετιών. Παρόλα αυτά τα στατιστικά στοιχεία δεν είναι πάντα επιτυχή όσον αφορά την απάντηση σε πολύπλοκες επιχειρηματικές ερωτήσεις στον σημερινό ανταγωνιστικό κόσμο των επιχειρήσεων.

Τεχνητή νοημοσύνη είναι ο τομέας που προσπάθησε να μιμηθεί τον ανθρώπινο τρόπο σκέψης σε στατιστικά προβλήματα. Έννοιες αυτού του εξαιρετικού τομέα συνέβαλαν επίσης στην ανάπτυξη της εξόρυξης δεδομένων. Η εφαρμογή ενός τέτοιου μοντέλου έχει ένα πρόσθετο πλεονέκτημα καθώς το σύστημα βελτιώνεται κάθε φορά που εκτελείται. <<Σκοπός της τεχνικής νοημοσύνης είναι να βγάζει λογικά συμπεράσματα από ανεπεξέργαστα δεδομένα, κάτι που κάνει και ο τομέας της εξόρυξης δεδομένων>> (Τσιράκης,Ν.2006.σελ.22). Μερικά παραδείγματα χρήσης εργαλείων τεχνητής νοημοσύνης από την εξόρυξη δεδομένων είναι τα νευρωνικά δίκτυα, τα δέντρα απόφασης και οι μηχανές διανυσμάτων.

Η μηχανική μάθηση θα μπορούσε να θεωρηθεί ως μια τεχνική που συνδυάζει την κλασική έννοια των στατιστικών στοιχείων και της τεχνητής νοημοσύνης. Με αυτή την τεχνική τα προγράμματα ηλεκτρονικών υπολογιστών “μαθαίνουν” για τα δεδομένα που μελετούν, έτσι ώστε να παίρνουν διαφορετικές αποφάσεις με βάση τις ιδιότητες των υπό μελέτη στοιχείων, με τη χρήση στατιστικών στοιχείων για τις θεμελιώδεις έννοιες και προσθέτοντας πιο προηγμένη τεχνητή νοημοσύνη και αλγόριθμους για να επιτύχουν τους στόχους τους.

Βάση δεδομένων είναι μια συλλογή από δεδομένα, τα οποία έχουν μια ορισμένη δομή ή σχήμα με το οποίο είναι σχετισμένα, έτσι ώστε να αναπαρίστανται με ένα πιο θεωρητικό τρόπο ή μοντέλο δεδομένων. Αυτό το μοντέλο χρησιμοποιείται για να περιγράψει τα δεδομένα, τα χαρακτηριστικά τους και τις σχέσεις μεταξύ τους. Με την ύπαρξη μη καλών συστημάτων διαχείρισης δεδομένων δεν είναι δυνατή η εφαρμογή αλγόριθμων εξόρυξης δεδομένων. Οι δύο αυτοί τομείς έχουν εμφανής σχέση και αυτό φαίνεται, καθώς μεγάλο μέρος των σημερινών ερευνητών που ασχολούνται με την εξόρυξη δεδομένων είναι άτομα προερχόμενα από το τομέα των βάσεων δεδομένων**.** (Τσιράκης, Ν.2006.σελ.22)

1.8 Ορισμός Διαδικασίας KDD (Knowledge Discovery In Database)

Σύμφωνα με τους Frawley, Piatesky-Shaphiro & Matheus(1991) <<KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα>> (Μεγαλοοικονόμου,Β. & Μακρής,Χ.2013)

1.9 Βασικές έννοιες ορισμού της διαδικασίας KDD

· Η διαδικασία KDD είναι τα βήματα που ακολουθούν την λήψη των δεδομένων ,τα οποία γίνονται κατά σειρά. Το πρώτο στάδιο είναι η προεπεξεργασία, μετά η αναζήτηση πρότυπων και τέλος η αξιολόγηση της εξαγόμενης γνώσης.

Ως προεργασία αναφέρονται οι μέθοδοι προετοιμασίας που γίνονται στα δεδομένα για να εφαρμόσουμε την εξόρυξη. Σε αυτό το σημείο έχουμε το ξεκαθάρισμα των δεδομένων, την ολοκλήρωσή τους, την επιλογή ενός συγκεκριμένου συνόλου δεδομένων που θα γίνει η εφαρμογή της εξόρυξης και την τροποποίηση τους σε συγκεκριμένη μορφή που απαιτείται από τον εκάστοτε αλγόριθμο εξόρυξης.

Το κύριο στάδιο εξόρυξης περιλαμβάνει την επιλογή των αλγορίθμων και των παραμέτρων προκειμένου να γίνει η αναζήτηση προτύπων
.
Τέλος, αφού τελικά πάρουμε τα εξαγόμενα αποτελέσματα, ακολουθεί η μετέπειτα διαχείριση και η αξιολόγηση τους.

· Τα δεδομένα είναι ένα σύνολο στιγμιότυπων ή απεικονίσεων καταστάσεων που εμφανίζονται σε μια βάση δεδομένων. Πχ μια συλλογή εγγράφων μιας ασφαλιστικής εταιρείας, όπου κάθε εγγραφή περιλαμβάνει τα πεδία (εισόδημα, οικογενειακή κατάσταση, περιοχή). (Ντούση,Ε.2003)

· Η πληροφορία έχει άμεση σχέση με την έννοια της γνώσης. Πχ πληροφορία από ένα σύνολο δεδομένων πωλήσεων ,είναι ποιο προϊόν πουλάει περισσότερο και πότε. Η γνώση αποκτιέται με τις πληροφορίες που έχουμε σε ένα συγκεκριμένο χρονικό διάστημα και αφορά την εύρεση προτύπων και σχέσεων ανάμεσα στα δεδομένα.(Γολέμη,Ε.2010)

· <<Το πρότυπο (pattern) είναι μια έκφραση Ε σε μια γλώσσα L η οποία περιγράφει ένα υποσύνολο δεδομένων $F_E \leq F$ εκμεταλλευόμενο κοινές ιδιότητες των δεδομένων του. Σε αυτή τη περίπτωση το πρότυπο θεωρείται υποσύνολο του F και αφαίρεση του F. Πχ, ο κανόνας : Εάν οι τηλεφωνικοί συνδρομητές έχουν income >$t ^age [$a_1$, a_2], δηλαδή εισόδημα μεγαλύτερο από μια τιμή t και η ηλικία τους βρίσκεται στο διάστημα τιμών [a_1 , a_2], τότε ανταποκρίνονται στη νέα προσφορά υπηρεσιών». (Χαλκίδη,Μ & Βαζιργιαννης,Μ.2005)
· Εγκυρότητα είναι το κατά πόσο συνεπές είναι το πρότυπο ως προς τον βαθμό της βεβαιότητάς του.

· Χρησιμότητα σημαίνει ότι όταν θα όταν έχουμε μπροστά μας πλέον τα πρότυπα, θα πρέπει να ακολουθηθούν κάποιες χρήσιμες διαδικασίες όπως αξιολόγηση και συνάρτηση χρησιμότητας. Καλό είναι να μπορέσουμε να πάρουμε όσο το δυνατόν περισσότερες πληροφορίες για κάθε στοιχείο.

· Κατανοησιμότητα σε μια εξόρυξης δεδομένων είναι ότι πρέπει τα αποτελέσματα να είναι τόσο κατανοητά, ώστε ακόμα και ένας αδαής να μπορέσει να βγάλει κάποιο χρήσιμο συμπέρασμα.

1.10 Τα βήματα της διαδικασίας KDD

· Ανάπτυξη και κατανόηση της προγενέστερης γνώσης του προς εξέταση τομέα και των τελικών στόχων που θα θέσουμε.

· Ολοκλήρωση των δεδομένων στα διαφορετικά είδη αποθήκευσης πληροφοριών που υπάρχουν για να μπορεί να γίνει ο συνδυασμός τους, προκειμένου να εφαρμοστεί η διαδικασία της εξόρυξης.

· Δημιουργία στόχου-συνόλου δεδομένων**.** Εδώ δηλαδή επιλέγονται τα σύνολα δεδομένων (δειγμάτων δεδομένων και μεταβλητών) στα οποία θα στηριχτεί η διαδικασία εξόρυξης για να εκτελεστεί.

· Καθορισμός και προ-επεξεργασία δεδομένων. Το βήμα αυτό περιλαμβάνει διαδικασίες όπως την αφαίρεση του θορύβου ή των outliers (περιπτώσεις δεδομένων πού διαφέρουν σημαντικά από τα υπόλοιπα λόγω εξαιρέσεων ή λαθών κατά την καταγραφή τους), τη συλλογή των πληροφοριών που θα είναι απαραίτητες στην διαμόρφωση ή τη μέτρηση θορύβου και την απόφαση σχετικά με τις στρατηγικές διαχείρισης των πεδίων δεδομένων που είναι ελλειπή. (Χαλκίδη,Μ. & Βαζιργιαννης,Μ.2005)

· Μετασχηματισμός δεδομένων σε μορφές κατάλληλες πλέον για εξόρυξη. Επίσης γίνεται χρήση μεθόδων για την μείωση των διαστάσεων ή μετασχηματισμού για να μειωθεί ο αριθμός των υπό εξέταση μεταβλητών.

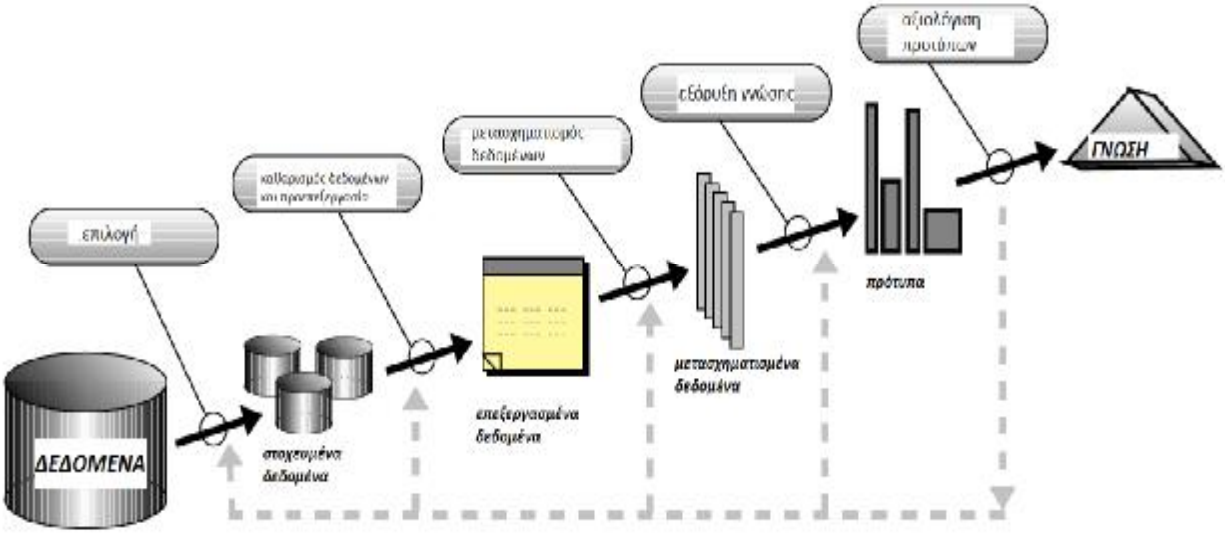
· Επιλογή των στόχων και των αλγόριθμων εξόρυξης δεδομένων. Εδώ αποφασίζεται ο στόχος της διαδικασίας KDD με βάση την επιλογή των στόχων εξόρυξης που θέλουμε να πετύχουμε. Επίσης επιλέγονται οι μέθοδοι που θα χρησιμοποιήσουμε και περιλαμβάνεται η επιλογή του κατάλληλου μοντέλου και παραμέτρων. <<Οι παράμετροι του μοντέλου, που είναι γνωστές από τα πρότυπα ή τα δεδομένα που προσδιορίζονται, αντιπροσωπεύουν την γνώση που έχει εξαχθεί από ένα σύνολο δεδομένων>> (Χαλκίδη, Μ & Βαζιργιαννης,Μ.2005)

· Εξόρυξη δεδομένων εφαρμόζοντας ευφυείς μεθόδους, ψάχνουμε για ενδιαφέροντα πρότυπα γνώσης, τα οποία μπορεί να είναι συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου τέτοιων πχ κανόνες κατηγοριοποίησης, δέντρα αποφάσεων, παλινδρόμηση, συσταδοποίηση κτλ. Η απόδοση και τα αποτελέσματα εξαρτώνται άμεσα από τα προηγούμενα βήματα.

· Αξιολόγηση των προτύπων χρησιμοποιώντας κάποια μέτρα για να προσδιοριστεί ποια από τα εξαγόμενα πρότυπα είναι κατάλληλα, ώστε να αντιπροσωπεύουν την γνώση.

· Σταθεροποίηση και παρουσίαση της γνώσης. Εδώ η γνώση που εξορύξαμε ενσωματώνεται στο σύστημα και χρησιμοποιούνται τεχνικές αντιπροσώπευσης γνώσης για να παρουσιαστεί η εξορυγμένη γνώση στον χρήστη. Τέλος γίνεται έλεγχος για πιθανές διαφορές/συγκρούσεις με παλιότερη εξορυγμένη γνώση.

Τα βήματα της KDD είναι μια διαλογική και επαναληπτική διαδικασία που πολλές φορές χρειάζεται και την ανθρώπινη παρέμβαση ώστε να ολοκληρωθεί με επιτυχία. Ο εκάστοτε χρήστης μπορεί να τροποποιεί τα μετρά αξιολόγησης και τα δεδομένα ώστε να παίρνει διαφορετικά και καταλληλότερα αποτελέσματα.



Σχήμα 1.1 Απεικόνιση ροής των βημάτων διαδικασίας KDD

ΚΕΦΑΛΑΙΟ 2.

Εξόρυξη Δεδομένων

2.1 Εξαγωγή προτύπων-Μοντέλα συναρμολογήσεων

Η εξόρυξη δεδομένων είναι από τα πιο σημαντικά και ενδιαφέροντα βήματα της KDD, καθώς περιέχει τα μοντέλα συναρμολογήσεων των υπό εξέταση δεδομένων, για την εξαγωγή προτύπων από αυτά.

Κατά την χρήση αλγορίθμων εξόρυξης για την εξαγωγή προτύπων πρέπει να έχουμε τη περιγραφή των μοντέλων η οποία χωρίζεται σε δύο κατηγορίες. Η πρώτη είναι η περιγραφή ως προς την λειτουργία του μοντέλου, δηλαδή ο καθορισμός των βασικών στόχων καθ' όλη την διάρκεια της διαδικασίας της εξόρυξης γνώσης. Η δεύτερη είναι η ως προς την παραστατική μορφή του μοντέλου ώστε να καθορίζει το ταίριασμα του με την απεικόνιση των δεδομένων και την δυνατότητα να ερμηνευτεί το μοντέλο με κατανοητούς όρους. Όσο πιο πολύπλοκα είναι τα μοντέλα, τόσο καλύτερα ταιριάζουν στα δεδομένα, αλλά ταυτόχρονα τόσο πιο δύσκολο είναι να γίνουν κατανοητά σε πραγματικές συνθήκες(Tan,P.N. et al.2010).

Στην συνέχεια χρειάζεται η αξιολόγηση του μοντέλου. Έχοντας υπόψη κάποια βασικά κριτήρια αξιολόγησης, μπορούμε να καθορίσουμε το κατά πόσο καλά ταιριάζει στα κριτήρια της KDD διαδικασίας το συγκεκριμένο μοντέλο για να δούμε την εγκυρότητα και την ακρίβεια του.

Τέλος σημαντικό για την εξόρυξη είναι, οι προδιαγραφές του αλγόριθμου αναζήτησης, να είναι οι κατάλληλες ώστε να μπορεί να βρεί συγκεκριμένους παραμέτρους και μοντέλα. Υπάρχου δύο τύποι αλγορίθμων αναζήτησης. Αυτοί της αναζήτησης παραμέτρων που είναι οι αλγόριθμοι που ψάχνουν για παραμέτρους, οι οποίες βελτιστοποιούν ένα κριτήριο αξιολόγησης για το μοντέλο. Σε αυτή την περίπτωση ο στόχος αναζήτησης εκτελείται παίρνοντας ως είσοδο ένα σύνολο δεδομένων και μια απεικόνιση μοντέλου. Ο δεύτερος τύπος αναζήτησης μοντέλων εκτελεί μια επαναληπτική διαδικασία αναζήτησης για την αντιπροσώπευση δεδομένων. Όταν θέλουμε συγκεκριμένη απεικόνιση ενός μοντέλου, εφαρμόζεται η μέθοδος αναζήτησης παραμέτρων και η ποιότητα των αποτελεσμάτων αξιολογείται.

2.2 Απαιτήσεις εξόρυξης δεδομένων

Για μια αποτελεσματική διαδικασία εξόρυξης δεδομένων, εξετάζεται το είδος των χαρακτηριστικών που αναμένεται να έχει το σύστημα εξόρυξης δεδομένων καθώς και τις απαιτήσεις που λαμβάνονται υπόψη στην ανάπτυξη των τεχνικών εξόρυξης δεδομένων. Οι κύριες απαιτήσεις είναι (Agrawal,R.1998):

Ο χειρισμός των διαφορετικών τύπων δεδομένων. Τα συστήματα εξόρυξης χρησιμοποιούν διαφορετικούς τύπους δεδομένων οπότε θα πρέπει να μπορούν να εφαρμόζονται αποτελεσματικά σε αυτούς. Επειδή πολλές βάσεις είναι «συγγενείς», το εκάστοτε σύστημα θα πρέπει να είναι αποτελεσματικό και αποδοτικό αλλά κυρίως να έχει τις σωστές τεχνικές για να μπορέσει να τις επεξεργαστεί. Πολλά συστήματα βάσεων δεδομένων έχουν πιο πολύπλοκους τύπους δεδομένων (υπερκείμενο, σύνθετα αντικείμενα κτλ.) οπότε αυτά τα συστήματα πρέπει να είναι λειτουργικά, άσχετα με το τύπο των δεδομένων που έχουν να επεξεργαστούν. Επομένως αυτή η διαφοροποίηση των τύπων δεδομένων αλλά και οι διαφορετικοί στόχοι της εξόρυξης μας δυσκολεύουν, ενώ θα μπορούσαν να είναι πιο ρεαλιστικά συστήματα για συγκεκριμένους τύπους δεδομένων.

Οι αλγόριθμοι θα πρέπει να είναι προσαρμοσμένοι στα σύνολα των δεδομένων ώστε να είναι αποδοτικοί, δηλαδή ο χρόνος εκτέλεσης των αλγορίθμων θα πρέπει να είναι αποδεκτός και να μην ξεπερνάει κάποια πλαίσια.

Τα αποτελέσματα που παίρνουμε από την εξόρυξη θα πρέπει να είναι ακριβής απεικόνιση της βάσης που χρησιμοποιήσαμε. Υπάρχουν τεχνικές που χρησιμοποιούνται για να εξακριβώσουμε την αποτελεσματικότητά τους. Ο θόρυβος και outliers που αφορούν τις εξαιρέσεις (σφάλματα) θα πρέπει να αντιμετωπιστούν με τις ανάλογες τεχνικές από τα συστήματα εξόρυξης που αυτά μας δίνει το κίνητρο να κάνουμε στατιστικές μελέτες, ανάλυση δεδομένων, μοντέλα προσομοίωσης πάνω στην εξορυγμένη γνώση.

Από μεγάλες βάσεις δεδομένων μπορούμε να εξορύξουμε πολλούς διαφορετικούς τύπους γνώσεων και μπορούμε να τις αποδώσουμε με διάφορες μορφές. Αυτό έχει σαν συνέπεια την ανάγκη να μπορούμε να εκφράσουμε τις υποερωτήσεις τους σε μια μορφή που να είναι εφαρμόσιμη από μη ειδικούς και η εξορυγμένη γνώση να είναι άμεσα εφαρμόσιμη από τους χρήστες. Σημαντική απαίτηση για τη σωστή παρουσίαση της γνώσης είναι το σύστημα να

μπορεί να αναπαραστήσει τη γνώση με τις σωστές εκφραστικές τεχνικές. (Νανόπουλος,Α & Μανωλόπουλος,Γ.2008)

Η διαλογική ανακάλυψη προσφέρει στο χρήστη τη δυνατότητα αλληλεπίδρασης με το σύστημα με άμεσο αποτέλεσμα να θέσει τις ερωτήσεις εξόρυξης δεδομένων με σκοπό να αλλάξει την εστίαση των δεδομένων που οδηγεί τη διαδικασία εξόρυξης σε πιο λεπτομερές επίπεδο ώστε να μπορούμε να δούμε τα αποτελέσματα από διαφορετικές σκοπιές.

2.3 Είδη δεδομένων

Τα συστήματα εξόρυξης δεδομένων μπορούν να ταξινομηθούν σύμφωνα με τα είδη βάσεων δεδομένων που εφαρμόζεται η εξόρυξη. Γενικότερα ένα σύστημα εξόρυξης δεδομένων ταξινομείται σε διάφορους τύπους συστημάτων δεδομένων όπως τα αντικειμενοστραφή, χωροχρονικές βάσεις δεδομένων, σχεσιακά συστήματα κ.α.

· Τα αντικειμενοστραφή βασίζονται στον αντικειμενοστραφή προγραμματισμό.

· Οι Χωροχρονικές βάσεις δεδομένων σχετίζονται με αντικείμενα που είναι κινητά ή μη και έχουν διαφορετικές θέσεις στην πάροδο του χρόνου και έχουν χωρικές ιδιότητες όπως σχήμα ,έκταση και θέση σε συνάρτηση με τον χρόνο. Ένα απλό παράδειγμα είναι η χρήση GPS/GSM/GPRS για να μελετήσουμε την κίνηση ενός ασθενοφόρου.

· Τα σχεσιακά συστήματα αποτελούνται από πίνακες όπου ο κάθε ένας από αυτούς μπορεί να αποθηκευτεί ξεχωριστά και να μπορούμε να τον χρησιμοποιήσουμε και μόνο του. Πολλές φορές χρησιμοποιείται γλώσσα επερωτήσεων (query languages).

2.4 Τεχνικές πρόβλεψης και περιγραφής των δεδομένων για την παραγωγή προτύπων

Υπάρχουν πολλοί αλγόριθμοι εξόρυξης γνώσης οι οποίοι χρησιμοποιούν έννοιες και τεχνικές από και άλλους επιστημονικούς τομείς. Το βασικό τους στοιχείο όμως που τους διαφοροποιεί από άλλες παρόμοιες τεχνικές που υιοθετούνται στην μηχανική μάθηση και τη στατιστική, είναι ότι οι αλγόριθμοι της εξόρυξης δεδομένων έχουν σχεδιαστεί με έμφαση στην εξελιξιμότητα όσον αφορά το μέγεθος του συνόλου δεδομένων εισαγωγής.

· Κατηγοριοποίηση (Classification): Ταξινόμιση των δεδομένων σε προκαθορισμένες κατηγορίες.
· Συσταδοποίηση (Clustering): Ομαδοποίηση των δεδομένων βάση ενός συνόλού κοινων χαρακτηριστηκων
· Κανόνες συσχέτισης (Association rules): Ανακάληψη σχέσεων ανάμεσα στα δεδομένα.
· Παλινδόμηση (Regression): Στατιστική εκμάθηση μιας συνάρτησης που απεικονίζει ένα στοιχείο σε μία πραγματική τιμή. Χρησιμοποιείται για μελλοντικές αριθμητικές προβλέψεις με την εφαρμογή νέων δεδομένων.

· Ανάλυση ακολουθιών((Sequence analisys): Εκτίμηση μοντέλου για ασυνεχείς σειρές.(Μακρής,Α.2015)
· Ανίχνευση ανωμαλιών(Anomaly detection): Αναγνώριση πρωτύπων σε σύνολα δεδομένων που παρουσιάζουν διαφορετική συμπεριφορα απο την αναμενώμενη. (Fayyad,U.1996)

2.5 Κατηγοριοποίηση

Η κατηγοριοποίηση έχει σαν βάση την εξέταση ενός μη κατηγοριοποιημένου αντικειμένου και την αντιστοίχηση του σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που θα κατηγοριοποιηθούν αναπαριστούνται από τις εγγραφές στη βάση δεδομένων και η κατηγοριοποίηση γίνεται σύμφωνα με την ανάθεση της κάθε εγγραφής στις προκαθορισμένες κατηγορίες. Η τεχνική αυτή χρησιμοποιείται για την κατασκευή μοντέλων που θα μπορούν να κατηγοριοποιήσουν νέα δεδομένα των οποίων η κατηγοριοποίηση είναι άγνωστη. Για να γίνει αυτό, το σύνολο των διαθέσιμων δεδομένων χωρίζονται σε ένα σύνολο δεδομένων εκπαίδευσης-training data και σε ένα σύνολο ελέγχου. Στο πρώτο στάδιο χρησιμοποιείται ένας αλγόριθμος κατηγοριοποίησης για την ανάλυση των δεδομένων ώστε να κατασκευαστεί το μοντέλο. Στο δεύτερο στάδιο το μοντέλο χρησιμοποιεί τα δεδομένα ελέγχου για να υπολογιστεί η ακρίβεια του. Με αυτή την διαδικασία μπορούμε να προβλέπουμε ένα πεδίο κλάσης έχοντας την βοήθεια των υπόλοιπων πεδίων τα οποία είναι οι παράμετροι υπολογισμού του. (Dunham,M.H.2004) Τα Δέντρα Αποφάσεων και Νευρωνικά δίκτυα είναι δυο τύποι αλγορίθμων κατηγοριοποίησης που βασίζονται στην εκπαίδευση ενός αντιπροσωπευτικού δείγματος του συνολικού όγκου δεδομένων. Με την εφαρμογή αυτή έχουμε τον καθορισμό προτύπων για τις κατηγορίες των δεδομένων, με αποτέλεσμα να μπορούμε να κατηγοριοποιήσουμε εύκολα τα νέα στοιχεία. Στις τεχνικές που εφαρμόζονται στα νευρωνικά δίκτυα παρατηρείται το φαινόμενο της αμφίδρομης αναμετάδοσης και επεξεργασίας δεδομένων, ενώ στις συμβολικές τεχνικές υπάρχουν μοντέλα δέντρων αποφάσεων ή μοντέλα IF/THEN/ELSE.

2.6 Συσταδοποίηση

Συσταδοποίηση είναι ο καταμερισμός ενός συνόλου δεδομένων σε συστάδες ομοίων χαρακτηριστικών. Οι διαφορές με την κατηγοριοποίηση είναι ότι στην κατηγοριοποίηση διαιρούμε το σύνολο των δεδομένων σε κατηγορίες τοποθετώντας τα δεδομένα σε προκαθορισμένες κατηγορίες σύμφωνα με το μοντέλο που αναπτύχτηκε από την εκπαίδευση του με παραδείγματα που είχαν κατηγοριοποιηθεί από την αρχή. Στην συσταδοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες, αλλά γίνεται ομαδοποίηση συνόλων σύμφωνα με την ομοιότητα που παρουσιάζεται μεταξύ τους, παράδειγμα ο διαχωρισμός σε συστάδες ενός πλήθους με βάση τα κινηματογραφικά τους ενδιαφέροντα. Τα βασικά βήματα της συσταδοποίησης είναι (Fayyad,U.1996) η επιλογή χαρακτηριστικών γνωρισμάτων, ο αλγόριθμος συσταδοποίησης ,η επικύρωση αποτελεσμάτων και τέλος η ερμηνεία των αποτελεσμάτων.

2.7 Κανόνες συσχέτισης

Η εξαγωγή κανόνων συσχέτισης είναι από τα σημαντικότερα κομμάτια της εξόρυξης δεδομένων. Χρησιμοποιούνται για να ανακαλύψουν κρυμμένες συσχετίσεις μεταξύ ενός συνόλου δεδομένων και παρουσιάζονται στη μορφή: A->B όπου A και B τα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα.

Έστω I={i1,i2,…in} ένα σύνολο από διακριτά στοιχεία. Ακόμα D={f1,f2,…,fm} ένα σύνολο από συναλλαγές, όπου κάθε συναλλαγή F είναι ένα σύνολο από αντικείμενα και ισχύει FÍ I. Κάθε συναλλαγή ταυτίζεται με ένα μοναδικό αναγνωριστικό που καλείται FID.

Ένας κανόνας είναι μία συσχέτιση της μορφής XÞ Y όπου X Í I , Y Í I και X ∩ Y=Ø. Το πρώτο μέλος του κανόνα ονομάζεται υπόθεση ενώ το δεύτερο ονομάζεται συμπέρασμα.

Υπάρχουν δύο βασικές μετρικές στον συσχετισμό κανόνων. Η υποστήριξη(support) που σημαίνει ότι ο κανόνας X Þ Y έχει υποστήριξη s, αν το s% των δοσοληψιών στο D περιέχουν το (X È Y)και η εμπιστοσύνη που σημαίνει ότι ο κανόνας X Þ Y ισχύει στο D, αν το c% των δοσοληψιών στο D που περιέχουν το X, περιέχουν επίσης και το Y. Από τα παραπάνω προκύπτει ότι ο κανόνας X Þ Y έχει υποστήριξη s, όταν sup(X Þ Y)=sup(XÈY) και εμπιστοσύνη c, όταν conf(X Þ Y) = sup(XÈY)/sup(X). (Βικιπαίδεια,2015)

2.8 Ανάλυση ακολουθιών

Η ανάλυση ακολουθιών αφορά την εξόρυξη προτύπων ή ακολουθιών σχετικά με τον χρόνο. Στα πρότυπα αυτά μπορούμε να εφαρμόσουμε περιορισμούς ώστε να βρίσκουμε πρότυπα τα οποία μας ενδιαφέρουν περισσότερο. Για παράδειγμα (B|D)F(G|D) δηλαδή ο χρήστης με αυτό τον κανόνα ενδιαφέρεται να βρεί τα γεγονότα B και Δ χωρίς να τον ενδιαφέρει η σχετική τους κατάληξη και ακολουθούνται από το γεγονός F ,το οποίο ακολουθείται από τα γεγονότα G και D.

2.9 Βασικοί τύποι παρουσίασης αποτελεσμάτων των αλγορίθμων εξόρυξης δεδομένων

Οι βασικότεροι τύποι αλγορίθμων είναι (Ahmed, S.R.2004):

- Τα νευροτικά δίκτυα (Κατηγοριοποίησης)
- Τα δέντα αποφάσεων (Κατηγοριοποίησης)
- Ο Naive Bayes (Κατηγοριοποιήσης)
- Οι γενετικοί αλγόριθμοι
- Η επαγωγή κανόνων
- Η μέθοδος του κοντινότερου γείτονα
- Η απεικόνιση δεδομένων

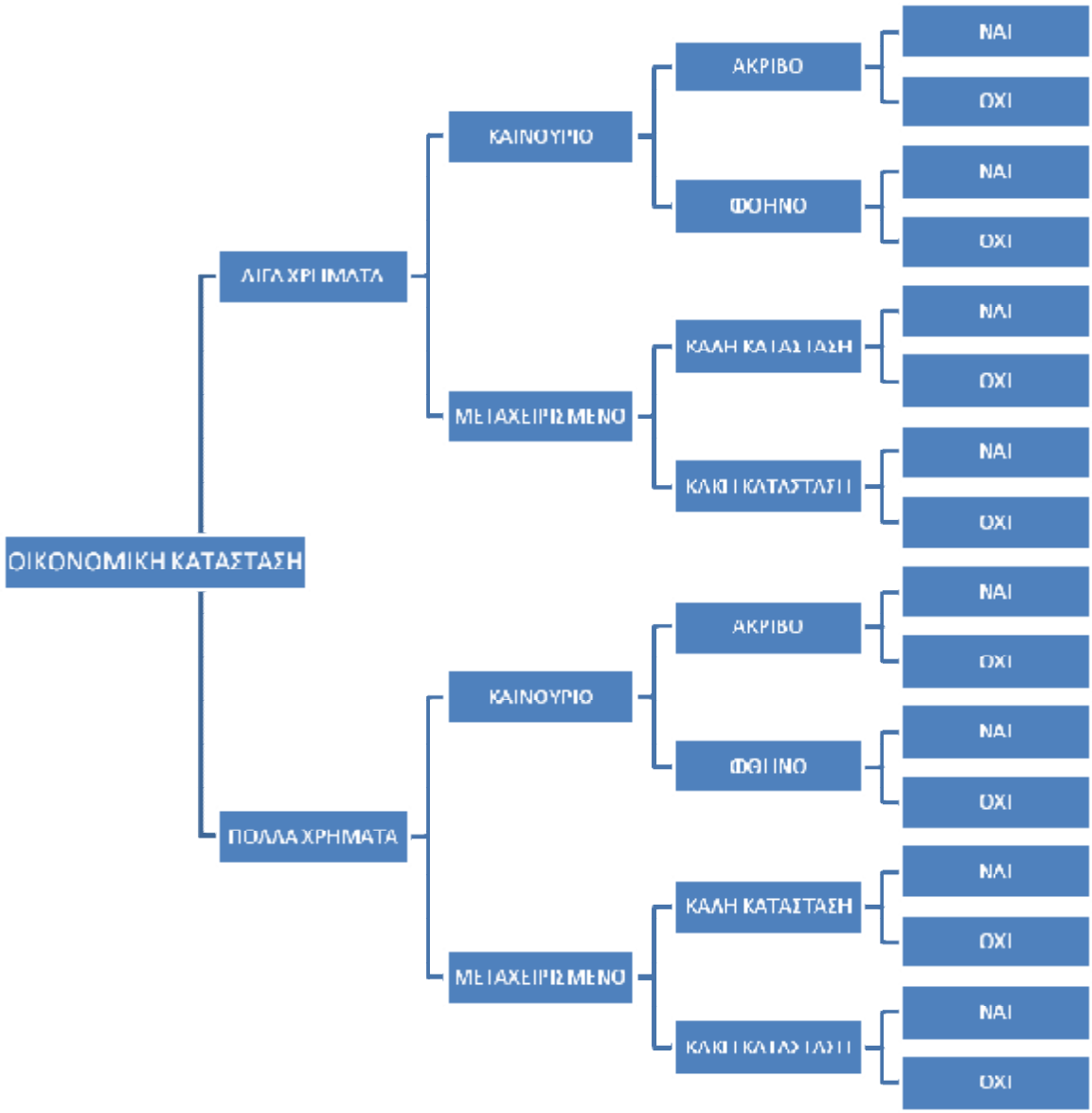
2.10 Δέντρα αποφάσεων

Χρησιμοποιούμε δέντρα αποφάσεων κυρίως για περιπτώσεις ταξινόμησης και πρόβλεψης. Αντιπροσωπεύεται από κανόνες IF – THEN – ELSE με αρχή τη ρίζα του δέντρου και κατάληξη στα φύλλα του (Murthy,S.1998). Τα γνωρίσματα του προβλήματος εμπεριέχονται στους κόμβους του δέντρου. Οι δυνατές τιμές των γνωρισμάτων περιέχονται στις ακμές και οι πιθανές κλάσεις του προβλήματος βρίσκονται στα φύλλα. Για την κατασκευή ενός δέντρου είναι απαραίτητο ένα σύνολο από στιγμιότυπα εκπαίδευσης. Το κάθε στιγμιότυπο έχει χαρακτηριστικά και επίσης την κλάση του προβλήματος όπου ανήκει.

Κατά την κατασκευή δέντρων αποφάσεων ξεκινάμε από την ρίζα, όπου ο αλγόριθμος διαχωρίζει το σύνολο των στιγμιότυπων σε υποσύνολα με βάση τη βέλτιστη ιδιότητα του κόμβου, η οποία καθορίζεται από κριτήρια όπως informetiongain,gain ratio κτλ. Στην συνεχεία για το κάθε ένα από αυτά τα υποσύνολα επαναλαμβάνεται η ίδια διαδικασία μέχρι το σημείο που τα στιγμιότυπα ανήκουν στην ίδια κλάση ή έχουν εξαντληθεί όλα τα γνωρίσματα. Επίσης υπάρχουν και τα στιγμιότυπα ελέγχου τα οποία ελέγχουν την απόδοση του δέντρου, δηλαδή το πόσο ακριβείς είναι η απάντηση του προβλήματος της ταξινόμησης. Στην περίπτωση αυτή το δέντρο παίρνει σαν είσοδο τις τιμές των γνωρισμάτων και παίρνουμε σαν απάντηση την τάξη του στιγμιότυπου. Οι απαντήσεις που παίρνουμε μας καθορίζουν την ακρίβειά του από το πόσες απαντήσεις δεν είχαν την πραγματική κλάση και είχαν κάποια άλλη. (Ντούση,Ε.2003)

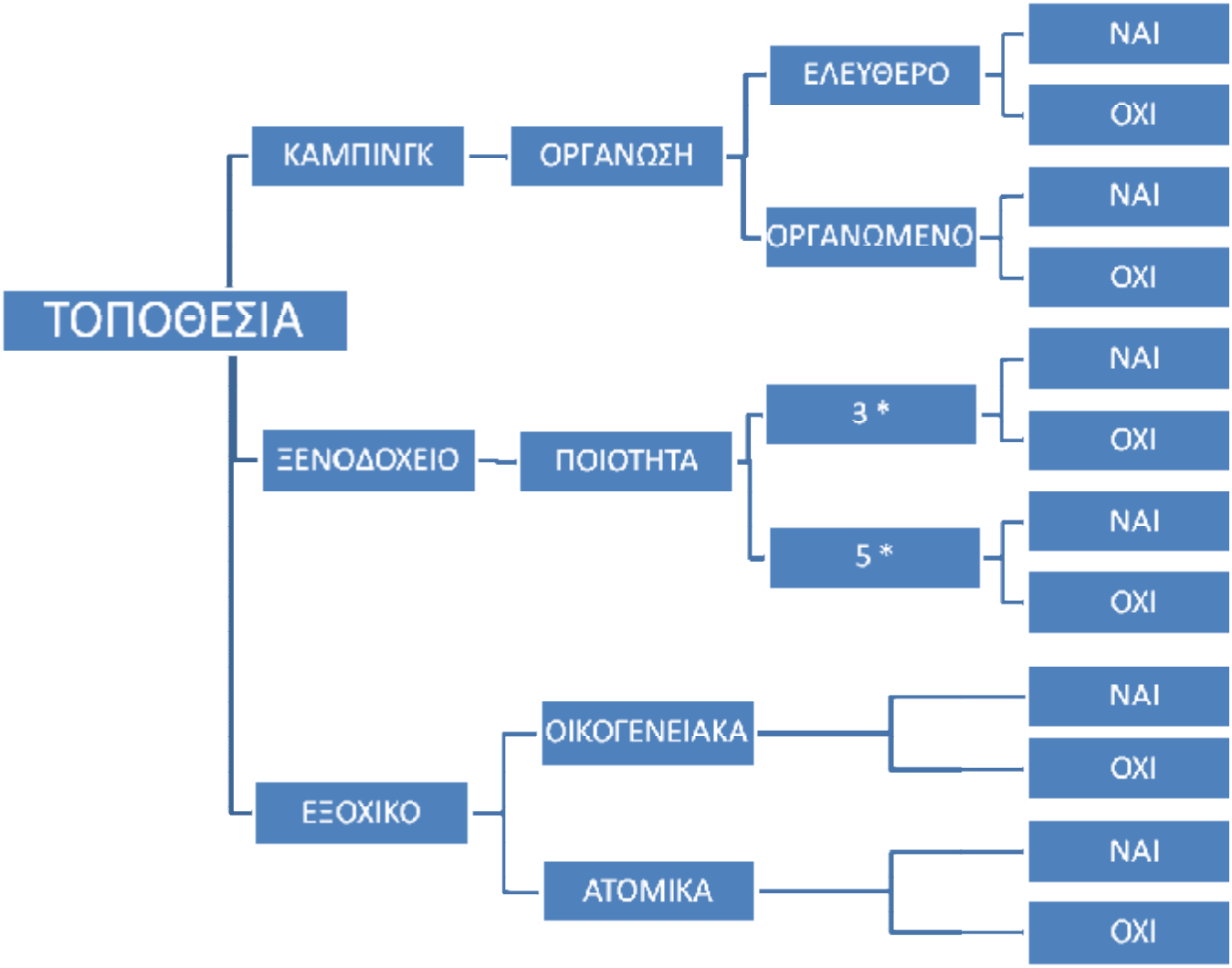
Έστω παράδειγμα στο ερώτημα «ΕΠΙΛΟΓΗ ΑΥΤΟΚΙΝΗΤΟΥ» με κλάσεις «ΝΑΙ» ή «ΟΧΙ»

Η απάντηση που θα πάρουμε εξαρτάται από τους παράγοντες «ΟΙΚΟΝΟΜΙΚΗ ΚΑΤΑΣΤΑΣΗ» με πιθανές τιμές « ΛΙΓΑ ΧΡΗΜΑΤΑ», «ΠΟΛΛΑ ΧΡΗΜΑΤΑ», «ΚΑΙΝΟΥΡΙΟ», «ΜΕΤΑΧΕΙΡΙΣΜΕΝΟ» με πιθανές τιμές «ΑΚΡΙΒΟ», «ΦΘΗΝΟ» ΚΑΙ «ΚΑΛΗ ΚΑΤΑΣΤΑΣΗ», «ΚΑΚΗ ΚΑΤΑΣΤΑΣΗ»



Σχήμα 2.1 Παράδειγμα 1 <<ΕΠΙΛΟΓΗ ΑΥΤΟΚΙΝΗΤΟΥ >>-Δέντρο απόφασης

Παράδειγμα «ΠΟΥ ΝΑ ΜΕΙΝΩ ΣΤΙΣ ΔΙΑΚΟΠΕΣ» με πιθανές κλάσεις «ΝΑΙ» ή «ΟΧΙ» με παράγοντες «ΤΟΠΟΘΕΣΙΑ» με πιθανές τιμές «ΚΑΜΠΙΝΓΚ» και «ΞΕΝΟΔΟΧΕΙΟ» και «ΕΞΟΧΙΚΟ» παράγοντες «ΟΡΓΑΝΩΣΗ» με πιθανές τιμές «ΕΛΕΥΘΕΡΟ» ή «ΟΡΓΑΝΩΜΕΝΟ» για το $1^{ο}$, «ΠΟΙΟΤΗΤΑ» με πιθανές τιμές «3*» και «5*» για το $2^{ο}$ και «ΟΙΚΟΓΕΝΕΙΑΚΑ» ή «ΑΤΟΜΙΚΑ» για το $3^{ο}$.



Σχήμα 2.2 Παράδειγμα 2 << ΠΟΥ ΝΑ ΜΕΙΝΩ ΣΤΙΣ ΔΙΑΚΟΠΕΣ >>-Δέντρο απόφασης

2.11 Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα είναι μια τεχνική που εφαρμόζεται για πρόβλεψη, ταξινόμηση και τμηματοποίηση. Έχουν την δυνατότητα να μαθαίνουν από τα δεδομένα τους και χρησιμοποιούνται για την εξαγωγή προτύπων και για να προσδιορίσουμε τάσεις οι οποίες είναι τόσο πολύπλοκες που δεν μπορούν να προσδιοριστούν από ανθρώπους ή από άλλες τεχνικές υπολογισμού (Κωνσταντίνος,Δ.2007). Σύμφωνα με παραπάνω, για κάθε νέο στιγμιότυπο σε ένα πρόβλημα μπορούμε να κάνουμε έγκυρη πρόβλεψη, εφόσον το νευρωνικό μας δίκτυο είναι εκπαιδευμένο. Τα νευρωνικά δίκτυα χρησιμοποιούν κόμβους όπως το ανθρώπινο μυαλό χρησιμοποιεί νευρώνες. Οι κόμβοι συνδέονται μεταξύ τους σε ένα

δίκτυο που αναγνωρίζει τα πρότυπα, όταν αυτά παρουσιαστούν σε ένα σύνολο δεδομένων. Η διαφορά των νευρωνικών δικτύων με τα παραδοσιακά προγράμματα υπολογιστών είναι ότι τα πρώτα μαθαίνουν από την εμπειρία, όπως οι άνθρωποι, ενώ τα δευτέρα ακολουθούν οδηγίες σύμφωνα με μια καθορισμένη σειρά. Η βασική μονάδα ενός νευρωνικού δικτύου είναι το perceptron, το οποίο παίρνει ως είσοδο ένα διάνυσμα πραγματικών τιμών, υπολογίζει ένα γραμμικό συνδυασμό των εισόδων και δίνει ως έξοδο 1 αν το αποτέλεσμα είναι μεγαλύτερο από κάποιο κατώφλι θ ή μηδέν διαφορετικά. Τα νευρωνικά δίκτυα εκπαιδεύονται με τρόπο ,όπου μια είσοδος οδηγεί σε μια συγκεκριμένη έξοδο. Μετέπειτα το δίκτυο μας ρυθμίζεται σύμφωνα με σύγκριση της τρέχουσας εξόδου με την επιθυμητή μέχρι να ταιριάξουν. Ο πιο διαδεδομένων αλγόριθμος εξορυξης δεδομένων είναι ο Back Propagation. (Ντούση,Ε.2003)

Κύρια βήματα κατασκευής μοντέλου κατηγοριοποίησης/πρόβλεψης με βάση τα νευρωνικά δίκτυα:

· Αναγνώριση των χαρακτηριστικών εισόδου και εξόδου.
· Σωστή επιλογή τοπολογίας για την κατασκευή του
· Επιλογή σωστού συνόλου εκπαίδευσης
· Επιλογή αντιπροσωπευτικού συνόλου δεδομένων για την εκπαίδευση του δικτύου. Θα πρέπει τα δεδομένα μας να απεικονίζονται με τρόπο που θα βελτιστοποιεί την δυνατότητα να αναγνωρίζει πρότυπα.
· Θα πρέπει να ελέγξουμε το δίκτυο χρησιμοποιώντας ένα σύνολο ελέγχου το οποίο είναι ανεξάρτητο από το σύνολο εκπαίδευσης.

2.12 Bayesian κατηγοριοποίηση

Το θεώρημα του Bayes έχει σαν στόχο να κατηγοριοποιεί ένα σύνολο δεδομένων D σε μία κατηγορία C με την χρησιμοποιώντας ένα μοντέλο πιθανοτήτων.

Η θεωρία του bayes έχει ως εξής(Ντούση,Ε.2003):

· *P* είναι η διαμοίραση πιθανότητας
· *D* είναι μια συλλογή στιγμιότυπων για τα οποία γνωρίζουμε την κλάση τους
· *h* είναι μια υπόθεση, όπως για παράδειγμα τα δεδομένα D να ανήκουν σε μία συγκεκριμένη κλάση *C*

Εάν γνωρίζουμε ότι:

· *P(h),* την A- priori πιθανότητα η υπόθεση *h* να είναι σωστή
· *P(D),* την πιθανότητα να παρατηρηθούν τα δεδομένα *D*
· *P(D|h),* την posteriori πιθανότητα να παρατηρηθούν τα δεδομένα *D* με την προϋπόθεση ότι η υπόθεση *h* είναι σωστή

τότε σύμφωνα με το θεώρημα αυτό μπορούμε να υπολογίσουμε την πιθανότητα P(h/D) δηλαδή την πιθανότητα η υπόθεσης μας h να είναι σωστή όταν παρατηρούνται τα δεδομένα D.Η σχέση μας δίδεται από τον ακόλουθο τύπο:

P(h|D)=P(D|h)*P(h)/P(D)

Ένα παράδειγμα για αυτό είναι :

Ας υποθέσουμε ότι έχουμε έναν αγρότη που παρακολουθεί μια συγκεκριμένη αγελάδα για 3 βδομάδες. Ας υποθέσουμε ότι αυτή η αγελάδα να είναι έγκυος με

- P(έγκυος =”ναι”)=x
- P(έγκυος =”όχι”)=y
- Με βασικό κανόνα χ+y=1

Επίσης υποθέτουμε ότι η αγελάδα μπορεί να είναι άρρωστη την συγκεκριμένη περίοδο με

- P(άρρωστη =”ναι”)=z
- P(άρρωστη =”όχι”)=d
- Με βασικό κανόνα z+d=1

Ας δούμε λίγο τους πιθανούς συνδυασμούς :

- Υποθέτουμε ότι η αγελάδα είναι έγκυος:

Ποια είναι η πιθανότητα να παρατηρήσει ο αγρότης μας ότι είναι και άρρωστη:

- P(άρρωστη =”ναι”| έγκυος =”ναι”)=x1
- P(άρρωστη =”όχι”| έγκυος =”ναι”)=x2
- Με βασικό κανόνα x1+x2 =1

Υποθέτουμε ότι η αγελάδα δεν είναι έγκυος:

Ποια είναι η πιθανότητα να παρατηρήσει ο αγρότης μας ότι είναι και άρρωστη:

- P(άρρωστη =”ναι”| έγκυος =”όχι”)=y1
- P(άρρωστη =”όχι”| έγκυος =”όχι”)=y2
- Με βασικό κανόνα y1+y2 =1

Επειδή το να είναι έγκυος έχει πιθανότητα, λόγω της κατάστασης της, 50% να είναι και αντίστοιχα 50% να μην είναι έγκυος.

Στο παρακάτω πίνακα φαίνονται και οι τυχαίες πιθανότητες για όλες τις άλλες καταστάσεις:

Έγκυος =”ναι”	Έγκυος =”όχι”
X=0,5	Y=0,5

	Άρρωστη =”ναι”	Άρρωστη =”όχι”
Έγκυος =”ναι”	X1=0,15	X2=0,85
Έγκυος =”όχι”	Y1=0,70	Y2=0.30

Ας υπολογίσουμε χρησιμοποιώντας την μέθοδο bayes την πιθανότητα να είναι έγκυος και να είναι άρρωστη.

Σύμφωνα με τον τύπο P(h|D)=P(D|h)*P(h)/P(D) έχουμε:

P(hi|Dj)=P(Dj|hi)P(hi) / (Dj|hi)P(h1)+ (Dj|hi)P(h2) + ….+(Dj|hn)P(hn)

P(έγκυος =”ναι”| άρρωστη =”ναι”)

P(άρρωστη =”ναι”| έγκυος =”ναι”)*(P(έγκυος =”ναι”)/ P(άρρωστη =”ναι”| έγκυος =”ναι”)*(P(έγκυος =”ναι”))+ P(άρρωστη =”ναι” |έγκυος =”όχι”)*(P(έγκυος =”όχι”))=

0,15*0,5/0,15*0,5+0,7 *0,5=

0,075/0,075+0,35=0,1764

Και τώρα την πιθανότητα να είναι έγκυος και να μην είναι άρρωστη.

P(έγκυος =”ναι” | άρρωστη =”όχι”)

P(άρρωστη =”όχι” | έγκυος =”ναι”)*(P(έγκυος =”ναι”)/ P(άρρωστη =”όχι” | έγκυος =”ναι”)*(P(έγκυος =”ναι”))+ P(άρρωστη= ”όχι”| έγκυος =”όχι”)*(P(έγκυος =”όχι”))=

0,85*0,5/0,85*0,5+0,3*0,5=

0,425/0,425+0,15=0,7391

2.13 Ο Απλός (Naïve) Bayes κατηγοριοποιητής

Ο Naïve bayes κατηγοριοποιητής είναι ο απλούστερος Bayesian κατηγοριοποιητής. Βασίζεται στην υπόθεση ότι η παρουσία/απουσία μιας συγκεκριμένης ιδιότητας κλάσης δεν σχετίζεται με την παρουσία/απουσία κάποιας άλλης. Έστω για παράδειγμα ότι θέλουμε να υπολογίσουμε την πιθανότητα να ένας άνθρωπος που έχει κάποια συμπτώματα (κρυάδες ,υγρή μύτη, πονοκέφαλος, πυρετός) να έχει γρίπη.

Η λειτουργία του Naive Bayes κατηγοριοποιητή συνοψίζεται στα ακόλουθα (Χαλκίδη,Μ & Βαζιργιαννης,Μ.2005):

Κάθε δείγμα Χ από το σύνολο S του προβλήματος αντιπροσωπεύετε από διανύσματα γνωρισμάτων x1,x2,…xn, δηλαδή X= <x1,x2,…xn>

Με την υπόθεση ότι το πρόβλημα έχει m κλάσεις, C1,C2,…Cm. στο δείγμα δεδομένων Χ του προβλήματος, για το οποίο δε γνωρίζουμε σε ποια κλάση ανήκει, ο κατηγοριοποιητής προβλέπει ότι το Χ ανήκει στην κλάση με τη μεγαλύτερη posteriori πιθανότητα. Ο κατηγοριοποιητής αναθέτει ένα άγνωστο στιγμιότυπο Χ του προβλήματος στην κλάση Ci αν ισχίει P(Ci|X)> P(Cj|X) για 1≤j≤m, j≠i

Η κλάση Ci στην οποία η πιθανότητα P(Ci|X) μεγιστοποιείται ονομάζεται μέγιστη μεταγενέστερη υπόθεση. Η μεγιστοποίηση της πιθανότητας P(Ci|X), γίνετε βάσει της υπόθεσης του θεωρήματος Bayes δίνεται από τη σχέση P(Ci|X)=P(X1|Ci)… P(Xn|Ci).

Κάθε μία από τις πιθανότητες P(Ci|X) υπολογίζεται από τα δεδομένα εκπαίδευσης.

Πίνακας 2.1 Παράδειγμα Naïve Bayes κατηγοριοποιητής

ΚΡΥΑΔΕΣ	ΥΓΡΗ ΜΥΤΗ	ΠΟΝΟΚΕΦΑΛΟΣ	ΠΥΡΕΤΟΣ	ΓΡΙΠΗ
Ναι	**Όχι**	**Μέτριος**	**Ναι**	**ΟΧΙ**
Ναι	**Ναι**	**Όχι**	**Όχι**	**ΝΑΙ**
Ναι	**Όχι**	**Δυνατός**	**Ναι**	**ΝΑΙ**
Όχι	**Ναι**	**Μέτριος**	**Ναι**	**ΝΑΙ**
Όχι	**Όχι**	**Όχι**	**Όχι**	**ΟΧΙ**
Όχι	**Ναι**	**Δυνατός**	**Ναι**	**ΝΑΙ**
Όχι	**Ναι**	**Δυνατός**	**Όχι**	**ΟΧΙ**
Ναι	**Ναι**	**Μέτριος**	**Ναι**	**ΝΑΙ**

Ρ (γρίπη = Ναι) = 0,625

Ρ (κρυάδες = Ναι | γρίπη = Ναι) =0,6

Ρ (κρυάδες = όχι | γρίπη = Ναι) =0,4

Ρ (υγρή μύτη = Ναι | γρίπη = Ναι) =0,8

Ρ (υγρή μύτη = όχι | γρίπη = Ναι) =0,2

Ρ (πονοκέφαλος = μέτριος | γρίπη = Ναι) =0,4

Ρ (πονοκέφαλος = όχι | γρίπη = Ναι) =0,2

Ρ (πονοκέφαλος = δυνατός | γρίπη = Ναι) =0,4

Ρ (πυρετός = ναι | γρίπη = Ναι) =0,8

Ρ (πυρετός = όχι | γρίπη = Ναι) =0,2

Ρ (γρίπη = Οχι) = 0,375

Ρ (κρυάδες = Ναι | γρίπη =όχι) =0,333

Ρ (κρυάδες = όχι | γρίπη = όχι) =0,666

Ρ (υγρή μύτη = Ναι | γρίπη = όχι) =0,333

Ρ (υγρή μύτη = όχι | γρίπη = όχι) =0,666

Ρ (πονοκέφαλος = μέτριος | γρίπη = όχι) =0,333

P (πονοκέφαλος = όχι | γρίπη = όχι) =0,333

P (πονοκέφαλος = δυνατός | γρίπη = όχι) =0,333

P (πυρετός = ναι | γρίπη = Ναι) =0,333

P (πυρετός = οχι | γρίπη = Ναι) =0,666

Υπόθεση 1η :

Κρυάδες	Υγρή μύτη	Πονοκέφαλος	Πυρετός	**γρίπη**
Ναι	Οχι	Μέτριος	Οχι	**Ναι**

· P (γρίπη = Ναι) * P (κρυάδες = Ναι | γρίπη = Ναι) * P (υγρή μύτη = όχι | γρίπη = Ναι) * P (πονοκέφαλος = μέτριος | γρίπη = Ναι) * P (πυρετός = όχι | γρίπη = Ναι) =0,006

Κρυάδες	Υγρή μύτη	Πονοκέφαλος	Πυρετός	**γρίπη**
ναι	όχι	Μέτριος	Οχι	**Όχι**

· P (γρίπη = Όχι) * P (κρυάδες = Ναι | γρίπη = Όχι) * P (υγρή μύτη = όχι | γρίπη = Όχι) * P (πονοκέφαλος = μέτριος | γρίπη = Όχι) * P (πυρετός = όχι | γρίπη = Όχι) =0,0185

Οπότε πιο πιθανό είναι ο ασθενής να μην έχει τελικά γρίπη γιατί 0,006 < 0,0185

Υπόθεση 2η :

Κρυάδες	Υγρή μύτη	Πονοκέφαλος	Πυρετός	**γρίπη**
Ναι	Ναι	Δυνατός	Οχι	**Ναι**

· P (γρίπη = Ναι) * P (κρυάδες = Ναι | γρίπη = Ναι) * P (υγρή μύτη = Ναι | γρίπη = Ναι) * P (πονοκέφαλος = δυνατός | γρίπη = Ναι) * P (πυρετός = όχι | γρίπη = Ναι) = 0,024

Κρυάδες	Υγρή μύτη	Πονοκέφαλος	Πυρετός	**γρίπη**
Ναι	Ναι	Δυνατός	Οχι	**Όχι**

· P (γρίπη = Όχι) * P (κρυάδες = Ναι | γρίπη = Όχι) * P (υγρή μύτη = Ναι | γρίπη = Όχι) * P (πονοκέφαλος = δυνατός | γρίπη = Όχι) * P (πυρετός = όχι | γρίπη = Όχι) = 0,0092

Οπότε πιο πιθανό είναι ο ασθενής να έχει γρίπη γιατι 0,024 < 0,0092

ΚΕΦΑΛΑΙΟ 3

Παγκόσμιος ιστός

3.1 ΕΙΣΑΓΩΓΗ

Εξόρυξη δεδομένων του παγκόσμιου ιστού, ορίζεται η χρήση τεχνικών εξόρυξης γνώσης για την αυτόματη ανακάλυψη και εξαγωγή δεδομένων από κείμενα και υπηρεσίες του παγκόσμιου ιστού.(Etzioni,O.1996)

<<Ο παγκόσμιος ιστός είναι πλέον ο δημοφιλέστερος τρόπος διάδοσης πληροφοριών και επικοινωνίας .Ουσιαστικά είναι μια πλατφόρμα από ερευνητικά άρθρα, forums επικοινωνίας και επικαιρότητας καθώς και δοσοληψιών μέσω του ηλεκτρονικού εμπορίου.>>(Χαλκίδη,Μ & Βαζιργιαννης,Μ.2005) Λόγω της ταχύρυθμης και χαοτικής ανάπτυξης του διαδικτύου δεν υπάρχει αυστηρή δομή και οργάνωση, για αυτό η αναζήτηση της χρήσιμης πληροφορίας από τον χρήστη δεν είναι εύκολη.

Στην εξόρυξη γνώσης από δεδομένα του παγκόσμιου ιστού (web mining) υπάρχει συνεισφορά από μεθόδους των ερευνητικών περιοχών των βάσεων δεδομένων ,της ανάκτησης πληροφοριών και της τεχνητής νοημοσύνης.

Τα δεδομένα του παγκόσμιου ιστού κατατάσσονται σε τρείς βασικές κατηγορίες ανάλογα με τα δεδομένα που δίνονται σαν είσοδο στις διαδικασίες εξόρυξης γνώσης. Πολλές φορές όμως χρησιμοποιούνται τα ίδια δεδομένα για διαφορετικούς σκοπούς για αυτό κάποιες φορές τα όρια μεταξύ τους μπορεί να γίνουν δυσδιάκριτα.

3.2 Κατηγορίες δεδομένων εξόρυξης γνώσης

Δεδομένα Δομής (structure data) ονομάζονται οι πληροφορίες που προέρχεται από υπερσυνδέσμους οι οποίοι είναι μέρος της δομής του ιστού. Από αυτούς τους μπορούμε να βρούμε ιστοσελίδες, κοινότητες χρηστών, τα ενδιαφέροντα τους κτλ. Κάθε σελίδα του ιστού είναι σαν ένας κόμβος οπου οι άκρες του είναι οι υπερσύνδεσμοι που την φέρνουν σε επαφή με άλλες σελίδες.

Από τα Δεδομένα Χρήσης (usage data), έχουμε την γνώση που αποκομίζεται από τα πρότυπα που συσχετίζονται με τις προτιμήσεις που έχουν οι χρήστες, σύμφωνα με τις περιηγήσεις του σε ιστοσελίδες ή από τα δεδομένα που εισάγει.

Τα Δεδομένα Περιεχομένου (content data) αφορούν την εξόρυξη προτύπων μέσα από τα περιεχόμενα των σελίδων που είναι κείμενα, εικόνες, ήχος ή και δομημένα δεδομένα. Έτσι γίνεται μια ομαδοποίηση τους ανάλογα με την θεματολογία τους.

3.3 Η εξέλιξη του τρόπου αναζήτησης στον Παγκόσμιο Ιστό

Στα πρώτα χρόνια του παγκόσμιου ιστού η πλοήγηση γινόταν με μηχανές αναζήτησης που χρησιμοποιούσαν λίστες τις οποίες κατασκεύαζαν άνθρωποι και οι οποίες περιείχαν τα πιο διαδεδομένα θέματα. Όταν αυξήθηκαν πολύ οι σελίδες στον Παγκόσμιο Ιστό, τότε

δημιουργήθηκαν οι αυτοματοποιημένες μηχανές αναζήτησης οι οποίες βασίζονται σε ομοιότητες μεταξύ λέξεων - κλειδιών. Στην συνέχεια αυτές οι μηχανές αναζήτησης για να παρέχουν πιο ποιοτικά αποτελέσματα στους χρήστες πρόσθεσαν κάποιες απλές ευρεστικές μεθόδους που λαμβάνουν υπό όψιν τους, την συχνότητα παρουσίασης ενός όρου μέσα στο κείμενο, αν εμφανίζεται στην αρχή του κειμένου ή σε περιοχές που θεωρούνται σημαντικές. Πλέον η αναζήτηση βασίζεται στη σημαντικότητα μιας σελίδας και είναι ανάλογη με το πλήθος των συνδέσμων που τη δείχνουν, ειδικά όταν οι σύνδεσμοι πηγάζουν από σημαντικές σελίδες. Αυτό φαίνεται στην δομή του γράφου του Π.Ι. Οι βασικότεροι αλγόριθμοι που έχουν αυτόν τον τρόπο αξιολόγησης είναι οι PageRank και HITS.

Το πιο δημοφιλές παράδειγμα εξόρυξης δεδομένων στο διαδίκτυο είναι η Google η οποία χρησιμοποιεί τον αλγόριθμο PageRank. Ο όγκος της πληροφορίας που υπάρχει μέχρι τώρα στο διαδίκτυο είναι αδύνατο να μετρηθεί με ακρίβεια. Οι σελίδες που κάθε φορά ερευνά η Google είναι πάνω από 100 δισεκατομμύρια. Κάθε ερώτημα στη μηχανή αναζήτησης δεν ξεπερνά σε χρόνο τα δυο δευτερόλεπτα. Η Google και γενικά ο τομέας της εξόρυξης δεδομένων στο διαδίκτυο έχουν σήμερα τεράστια επιτυχία, γιατί καταφέρνουν να (Τσιράκης.Ν):

1. μπορούν να κάνουν αναζήτηση σε τεράστια ποσότητα δεδομένων σε πολύ σύντομο χρόνο.
2. μπορούν να επιστρέψουν σε κάθε ερώτημα τα πρώτα αποτελέσματα που είναι πιο χρήσιμα, σύμφωνα με τις πληροφορίες που έχει αποκομίσει από τον εκάστοτε χρήστη.

Έτσι, τελικά ο χρήστης λαμβάνει γρήγορα και εύκολα μόνο την ουσιώδη πληροφορία που θέλει.

3.4 Ο γράφος του Π.Ι.

Ο Παγκόσμιος Ιστός είναι ένας κατευθυνόμενος γράφος στον οποίο οι κόμβοι είναι οι ιστοσελίδες και ακμές οι υπερσύνδεσμοι τους. Η δομή του έχει έναν ισχυρά συνδεδεμένο πυρήνα 56 εκατομμυρίων σελίδων στο κέντρο και δυο ακόμα τμήματα 44 εκατομμυρίων σελίδων στα δυο άκρα, όπου το ένα περιέχει σελίδες που δείχνουν προς το SCC (το σύνολο ΠΡΟΣ) και το άλλο περιέχει σελίδες που δείχνονται από το SCC (το σύνολο ΑΠΟ). Επιπρόσθετα κάποιοι “σωλήνες” ενώνουν απευθείας τα σύνολα ΠΡΟΣ και ΑΠΟ. Τέλος, υπάρχουν και κάποια μικρότερα τμήματα (ομάδες σελίδων) τα οποία δεν μπορούν να προσπελασθούν από κανένα άλλο τμήμα αυτής της δομής λόγω της έλλειψης συνδεσμολογίας. (Χαλκίδη,Μ & Βαζιργιαννης,Μ.2005)

3.5 Στόχοι των αξόνων εξόρυξης δεδομένων

Στόχος στην εξόρυξη γνώσης από δεδομένα δομής (web structure mining) είναι η ταξινόμηση των σελίδων και η εξαγωγή πληροφορίας σχετικά με τις σχέσεις μεταξύ τους, όπως αυτές προκύπτουν από την τοπολογία των υπερσυνδέσμων στον γράφο του παγκόσμιου ιστού. Τα τελευταία χρόνια , η περιοχή αυτή επικεντρώνεται στην ανακάλυψη τερματικών σελίδων (authorities), δηλαδή σελίδων που θεωρούνται σημαντικές πηγές πληροφορίας από πολλούς χρήστες. Σχετική με αυτή την περιοχή είναι η επεξεργασία του γράφου του παγκόσμιου ιστού.

Η εξόρυξη γνώσης από δεδομένα χρήσης (web usage mining) είναι η διαδικασία ανακάλυψης και επεξεργασίας προτύπων που περιγράφουν την πλοήγηση των χρηστών μέσα σε ένα δικτυακό τόπο (navigational/browsing patterns). Αυτή η διαδικασία δέχεται ως είσοδο τα δεδομένα χρήσης που βρίσκονται στο αρχείο του εξυπηρετητή του εκάστοτε δικτυακού τόπου(web server logs) και τα οποια καταγράφουν τις επισκέψεις των χρηστών . Η εκτεταμένη έρευνα σε αυτό τον τομέα, οδήγησε στη δημιουργία του σχετιζόμενου ερευνητικού τομέα, εξατομίκευση του παγκόσμιου ιστού(web personalization). Αυτή η περιοχή χρησιμοποιεί τα αποτελέσματα της εξόρυξης γνώσης από δεδομένα χρήσης, ώστε να παρέχει με δυναμικό τρόπο προτάσεις πλοήγησης (recommendations) προς τους χρήστες.

Η εξόρυξη γνώσης από δεδομένα περιεχομένου (web content mining) ασχολείται με την ανάκτηση του περιεχομένου των δικτυακών τόπων, καθώς επίσης και με τη δεικτοδότηση τους (indexing) για τη διευκόλυνση της αναζήτησης. Το περιεχόμενο μπορεί να είναι αδόμητο (απλό κείμενο), ημιδομημένο (html/xml κείμενο), ή δομημένο με εγγραφές που προέρχονται από βάση δεδομένων και εμφανίζονται σε δυναμικές σελίδες.

<<Οι περισσότερες ερευνητικές προσπάθειες σήμερα προτείνουν συστήματα και αλγόριθμους που συνδυάζουν μεθόδους και από τις τρεις κατηγορίες, ενώ υπάρχει τάση για υιοθέτηση ολοένα και περισσότερο της σημασιολογίας (semantics). Αυτά τα σημασιολογικά δεδομένα μπορούν να οριστούν με εργαλεία τα οποία προκύπτουν στο γενικότερο πλαίσιο του Σημασιολογικού Ιστού, όπως XML, RDF και οντολογίες.>> (Χαλκίδη, Μ & Βαζιργιαννης,Μ.2005)

3.6 Περιγραφή της μεθόδου PageRank

Στην εφαρμογή της μεθόδου PageRank χρησιμοποιείται ένας γράφος που έχει κόμβους τις ιστοσελίδες(websites) και ακμές τους υπερσύνδεσμους (links). Σε αυτή τη μέθοδο η βασική ιδέα είναι ότι η σημαντικότητα – εγκυρότητα μιας σελίδας είναι ανάλογη με το πλήθος των υπερσυνδέσμων που έχει από άλλες σελίδες. Επίσης, ισχύει ότι οι υπερσύνδεσμοι από σημαντικούς κόμβους έχουν μεγαλύτερη αξία από ότι λιγότερο σημαντικοί κόμβοι. (Leskovec, Jure et al. 2014)

Στόχος της μεθόδου PageRank είναι ο υπολογισμός μιας επίδοσης (score) για κάθε κόμβο έτσι ώστε οι σημαντικότερες σελίδες να έχουν μεγαλύτερο βαθμό.Το PageRank παρακολουθεί το μονοπάτι που ακολουθεί ο κάθε χρήστης και μετράει τη πιθανότητα που υπάρχει ξεκινώντας από ένα τυχαίο website και φτάνοντας, ακολουθώντας links, σε μία συγκεκριμένη ιστοσελίδα. Όσο μεγαλύτερο είναι το PageRank, μετρώντας σε score, μιας συγκεκριμένης ιστοσελίδας, τόσο μεγαλύτερη και η πιθανότητα στο χρήστη να φτάσει σε αυτήν. Βασικό παράγοντα στην μέτρηση αύτη έχουν τα links. Κάθε link υπολογίζεται σαν «ψήφος εμπιστοσύνης» από μια ιστοσελίδα προς μία άλλη και για αυτό τον λόγο όσο περισσότερα links έχει μία ιστοσελίδα τόσο περισσότερες πιθανότητες έχει για να την ανακαλύψουν οι χρήστες και κατά συνέπεια και μεγαλύτερο PageRank.

<<Το PageRank υπολογίζεται με βάση μία κλίμακα από το 0 έως το 10. Όσο μεγαλύτερο το νούμερο τόσο μεγαλύτερο το PageRank. Το 0 αντιστοιχεί συνήθως στις καινούργιες ιστοσελίδες ή στις ιστοσελίδες που δεν έχουν πολλά links. Το 10 αντιστοιχεί σε ελάχιστες ιστοσελίδες κορυφαίων websites όπως αυτή της κεντρικής ιστοσελίδας του Google.

Η κλίμακα αυτή δεν είναι γραμμική. Δηλαδή, εάν μία ιστοσελίδα έχει PageRank 2 και 40 links δεν σημαίνει αυτόματα και ότι με 80 links θα έχει PageRank 4. Αντιθέτως για να ανεβάσει μια σελίδα το επίπεδό της στην PageRank, θα πρέπει να έχει όλο και περισσότερα links. Γενικά, ισχύει ο κανόνας ότι όσο ανεβαίνει επίπεδα θα πρέπει να ανεβαίνει εκθετικά σε αριθμό links.>>

Ή κάθε σελίδα που θέλει να αυξήσει το score της, ώστε να βελτιώσει τη κατάταξή της ανάμεσα στις υπόλοιπες, θα πρέπει να εξετάσει πόσο σχετικές είναι αυτές με το αντικείμενο της δικής της πριν επιδιώξει να αποκτήσει links από αυτές, με όσο το δυνατόν καλύτερο PageRank. <<Οι ιστοσελίδες με μεγαλύτερο PageRank έχουν κατά κανόνα περισσότερα links επειδή προσελκύουν το ενδιαφέρον περισσότερων χρηστών, χωρίς αυτό να σημαίνει ότι γίνεται λόγω της ποιότητας και του καλού περιεχομένου τους.>> (Καψωμενάκης,Ν.2015)

Ο πίνακας πιθανοτήτων μετάβασης P ορίζεται ως P= (1-e)*(T+Z) + e*S

3.7 Θεμελιώδεις έννοιες της μεθόδου

Πίνακας πιθανοτήτων μετάβασης (Transition probability matrix) P είναι ένας τετραγωνικός πίνακας που περιέχει τις πιθανότητες μετάβασης από μια κατάσταση i σε μια κατάσταση j σε οποιαδήποτε στιγμή με ένα βήμα. Ο πίνακας P πρέπει να έχει τα ακόλουθα χαρακτηριστικά:

· Ο P πρέπει να είναι στοχαστικός ως προς τις γραμμές.

· Pij >= 0, για κάθε κατάσταση i , ,j

Πιθανότητα μετάβασης Pij είναι η πιθανότητα $P[X(t_{n+1})=x_j \mid X(t_n)=x_i]$ δηλαδή πιθανότητα μετάβασης από την κατάσταση Xi στην Xj.

Στάσιμη πιθανότητα μετάβασης είναι η μετάβαση σε μία κατάσταση ανεξάρτητη του χρόνου.

Στοχαστικός ως προς τις γραμμές είναι ένας πίνακας , όταν το άθροισμα των στοιχείων κάθε γραμμής είναι 1.Δηλαδή ο γράφος δεν έχει κόμβους καταβόθρες.

Σελίδες καταβόθρες ονομάζονται οι σελίδες που δεν έχουν κανένα εξερχόμενο σύνδεσμο. Συνεπώς αν i είναι μια σελίδα καταβόθρα, τότε το άθροισμα του Pij (για κάθε στήλη j) είναι μηδέν.

Τυχαίος πλοηγητής (Random Surfer) ονομάζεται ένας ιδεατός χρήστης που κάνει τυχαίους περιπάτους σε ένα γράφημα. Όταν ο ιδεατός χρήστης βρίσκεται σε έναν κόμβο του γραφήματος, τότε μπορεί να ακολουθήσει τους εξερχόμενους συνδέσμους της σελίδας (να κάνει μία μετάβαση μέσω κάποιας ακμής) ή να μεταβεί τυχαία σε κάποιον άλλο κόμβο (να κάνει μια τυχαία προσπέλαση). Συνήθως η τυχαία προσπέλαση συμβαίνει με πιθανότητα e=0,15 περίπου και αποτρέπει την παγίδευση της διαδικασίας σε έναν κόμβο καταβόθρα.

Τυχαίος περίπατος ορίζεται η διακριτή Μαρκοβιανή αλυσίδα που χαρακτηρίζεται από στάσιμες πιθανότητες μετάβασης οι οποίες προκύπτουν από την εφαρμογή του PageRank στους κόμβους του γράφου. Επιπλέον ένα άλλο χαρακτηριστικό του είναι ότι το σύνολο των δεικτών είναι φυσικοί αριθμοί που αντιπροσωπεύουν τα τυχαία βήματα που κάνει ο τυχαίος πλοηγητής.

Η Μαρκοβιανή αλυσίδα είναι μια στοχαστική διαδικασία που έχει την Μαρκοβιανή ιδιότητα.

Στοχαστική διαδικασία είναι ένα σύνολο τυχαίων μεταβλητών {X(t), t ϵ T}. Στις περισσότερες περιπτώσεις το t αναπαριστά το χρόνο. Το σύνολο δεικτών (index set) T της στοχαστικής διαδικασίας και οι τιμές που παίρνουν οι τυχαίες μεταβλητές X(t) αποτελούν το χώρο καταστάσεων (state space) S.

Μία στοχαστική διαδικασία έχει την Μαρκοβιανή ιδιότητα αν για κάθε σύνολο n+1 τιμών από το σύνολο δεικτών $t_1 < t_2 < \ldots < t_{n+1}$ και για κάθε σύνολο n+1 καταστάσεων $\{x_1, x_2 \ldots x_{n+1}\}$ από τον χώρο καταστάσεων, τότε ισχύει ότι : $P[X(t_{n+1}) = x_{n+1} \mid X(t_1) = x_1 , X(t_2) = x_2 , \ldots, X(t_n) = x_n] P[X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n]$ Δηλαδή η τιμή στην κατάσταση X_{n+1} εξαρτάται μόνο από την αμέσως προηγούμενη X_n .

Διακριτή Μαρκοβιανή αλυσίδα είναι μία αλυσίδα με διακριτό σύνολο δεικτών T.

Αμείωτη Μαρκοβιανή αλυσίδα είναι όταν όλες οι καταστάσεις της είναι αμοιβαία προσβάσιμες. Δηλαδή για κάθε ζεύγος καταστάσεων I,j μια διαδικασία μπορεί να μεταβεί από την κατάσταση I στην κατάσταση j σε πεπερασμένο αριθμό βημάτων.

Εργοδική Μαρκοβιανή αλυσίδα είναι μια διακριτή Μαρκοβιανή αλυσίδα που είναι χρονικά ομοιογενής , αμείωτη, μη περιοδική και θετικά επαναλαμβανόμενη.

Ο τυχαίος περίπατος που κάνει ο τυχαίος πλοηγητής είναι μια εργοδική Μαρκοβιανή αλυσίδα και έτσι έχει μοναδικές στάσιμες πιθανότητες κατάστασης. Με αυτόν τον τρόπο , οι βαθμοί που προκύπτουν από την εφαρμογή του PageRank είναι μοναδικοί και κάθε σελίδα θα έχει μια στάσιμη πιθανότητα κατάστασης.

3.8 Υπολογισμός πιθανότητας επίσκεψης σε μία σελίδα

Στον τύπο υπολογισμού πιθανοτήτων μετάβασης P= (1-e) (T+Z) + e*S έχουμε τους παρακάτω πίνακες :

A είναι ο πίνακας γειτνίασης(adjacency matrix) του γραφήματος. Αν υπάρχει ακμή που ξεκινάει από τον κόμβο i και καταλήγει στον j τότε Aij=1 αλλιώς if Aij=0 και ισχιεί: Tij = 1 / outdegree(i) , if Aij=1 and Tij= 0 , if Aij=0

H συνάρτηση outdegree υπολογίζει το πλήθος των εξερχόμενων συνδέσμων μιας σελίδας.

S είναι ένας πίνακας που περιγράφει το τυχαίο πήδημα (random jump) και ισούται με e*s.

s ορίζεται ως ο πίνακας τυχαίας πρόσβασης διαστάσεων 1*n και κάθε στοιχείο του είναι ίσο με 1/n (όπου n το πλήθος των κόμβων του γραφήματος).

e είναι ένας (n*1) πίνακας κάθε στοιχείο του οποίου είναι ίσο με 1

Z είναι ο πίνακας που μας εξασφαλίζει ότι η Μαρκοβιανή αλυσίδα είναι στοχαστική ως προς τις γραμμές και υπολογίζεται ως z*s

z είναι ένας (n*1) πίνακας που έχει :

zi=1 αν μία σελίδα i δεν έχει εξερχόμενους συνδέσμους και

zi=0 διαφορετικά

Οι στάσιμες πιθανότητες κατάστασης μιας Μαρκοβιανής αλυσίδας ,που ταυτίζονται με τους βαθμούς κάθε σελίδας του PageRank ,μπορούν να υπολογιστούν εφαρμόζοντας τη δυναμομέθοδο (power method) δηλαδή το $\chi^{(\kappa+1)}$ μπορεί να υπολογιστεί στο κ+1 βήμα με τον πολλαπλασιασμό πινάκων $\chi^{\kappa} * P$. Η επανάληψη τελειώνει όταν $| \chi^{(\kappa+1)} - \chi^{\kappa} | <= \delta$, είναι ένας πολύ μικρός αριθμός (Brin,S & Page,L.1998).

3.9 Μέθοδος Hits

Σε αυτή τη μέθοδο οι σελίδες χαρακτηρίζονται από τις επιδόσεις τους ως κεντρικές (hub score – y_i) και ως έγκυρες (authority score – x_i) ιστοσελίδες.

Τα βήματα για την υλοποίηση του αλγόριθμου είναι τα παρακάτω (Χαλκίδη,Μ & Βαζιργιαννης,Μ.2005)

:

· Στο πρώτο βήμα συγκεντρώνονται οι ιστοσελίδες οι οποίες περιέχουν έναν, όλους ή συγκεκριμένους όρους του ερωτήματος. Έστω Ν1 το σύνολο αυτών των σελίδων. Φτιάχνουμε ένα Ν1 υπογράφημα του γραφήματος Ν, το οποίο θα έχει Ν1 κόμβους και ακμές οι οποίες ξεκινούν και καταλήγουν σε κόμβους του συνόλου Ν.

· Μετά επεκτείνουμε το Ν1 γράφημα , προσθέτοντας και άλλους κόμβους, στο γράφημα Ν και δείχνουν ή δείχνονται από κόμβους, οι οποίοι ανήκουν στο γράφημα Ν1.Έστω τώρα Ν2 ο νέος αυτός γράφος. Επειδή ο Ν2 μπορεί να γίνει πολύ μεγάλος, βάζουμε ένα ανώτατο όριο στους αριθμούς των εισερχόμενων και των εξερχόμενων ακμών σε κάθε κόμβο. (Έστω Ε το σύνολο των ακμών e_{ij}).

· Υπολογίζουμε τον γειτονικό πίνακα L με στοιχεία $L_{ij}=1$ αν υπάρχει ακμή e_{ij} και $L_{ij}=0$ αν δεν υπάρχει. Υπολογίζουμε τα score και μετά από την κανονικοποίηση τους προκύπτει $\chi^{(\kappa)}=L^T * y^{(k-1)} =L^T * L * x^{(k-1)}$ και $y^{(\kappa)} =L* \chi^{(\kappa)}=L* L^T * y^{(k-1)}$

· Έτσι μπορούμε να χρησιμοποιούμε τη δυναμομέθοδο για τον υπολογισμό του επικρατέστερου χ ιδιοδιανύσματος με authority score και του επικρατέστερου y ιδιοδιανύσματος με hub score.

· Και τέλος ταξινομούμε τις ιστοσελίδες σε λίστες (μια για κάθε επίδοση) βάσει των παραπάνω επιδόσεων.

3.9.1Πλεονεκτήματα – Μειονεκτήματα της μεθόδου Hits

Τα σημαντικά πλεονεκτήματα της μεθόδου είναι ότι:

· δίνει ως έξοδο δύο ταξινομημένες λίστες άρα και περισσότερες επιλογές στον χρήστη, ως προς το ποια ταξινόμηση να διαλέξει.

· βρίσκει τα επικρατέστερα ιδιοδιανύσματα μικρών πινάκων αντιμετωπίζοντας το πρόβλημα “ανάκτηση πληροφορίας” από το παγκόσμιο ιστό.

Τα μειονεκτήματα είναι :

· Για κάθε ερώτημα θα πρέπει να δημιουργείται ένα υπογράφημα N2

· Είναι επιρρεπής στο spamming. Υπάρχει η δυνατότητα κάποιος χρήστης του web να μπορεί να βάλει στην ιστοσελίδα του συνδέσμους που την δείχνουν σε άλλες ιστοσελίδες και έτσι να μεγαλώνει το hub score της δικιάς του, όποτε και το authority score.

· Topic drift. Εδώ το πρόβλημα συναντάται στη δημιουργία ενός υπογραφήματος N2 για κάποιο ερώτημα, μια σελίδα με μεγάλο authority score, αλλά εκτός θέματος, είναι πιθανό να συνδέεται με κάποια σελίδα που περιέχει όρους του ερωτήματος αυτού. Αυτή η εκτός θέματος σελίδα μπορεί να έχει όμως τόσο μεγάλο authority score, που να επικρατήσει μαζί με τις γειτονικές της σελίδες στην ταξινομημένη λίστα.(Χαλκκίδη,Μ. & Βαζιργιάννης,Μ.2005)

3.10 Εξόρυξη γνώσης από τον παγκόσμιο ιστό με βάση το περιεχόμενο

Η εξόρυξη γνώσης από τον παγκόσμιο ιστό με βάση το περιεχόμενο, συνήθως χρησιμοποιείται σε συνδυασμό με την εξόρυξη γνώσης με βάση την δομή .Ο συνδυασμός των δυο τεχνικών, μας δίνει καλύτερα αποτελέσματα στην οργάνωση και την εξαγωγή της πληροφορίας. Στόχος της εξόρυξης γνώσης με βάση το περιεχόμενο, είναι να αποκτήσουν τα κείμενα των ιστοσελίδων επεξεργάσιμο μορφότυπο. Για να το πετύχουμε αυτό χρησιμοποιούνται τρία στάδια από την τεχνική εξόρυξης γνώσης των κειμένων. Τα στάδια αυτά είναι η προεπεξεργασία κειμένων, η αναπαράσταση κειμένων και τέλος η εξαγωγή χαρακτηριστικών γνωρισμάτων από κείμενα.

3.11 Στάδια εξόρυξης γνώσης των κειμένων

Η προεπεξεργασία κειμένων αποτελείται από δύο στάδια. Το πρώτο είναι η αφαίρεση τετριμμένων λέξεων (stop-work removal). Αυτό γίνεται, γιατί αν στην ανάλυση ή τον χαρακτηρισμό ενός κειμένου συμπεριλάβουμε άρθρα, συνδέσμους (και ,όμως, κτλ), αντωνυμίες ή και επιρρήματα που δεν έχουν ιδιαίτερη σημασιολογική πληροφορία, τότε αυτά θα λειτουργήσουν σαν θόρυβος και θα μειώσουν την απόδοση του αλγόριθμου θα χρησιμοποιηθεί για την εξόρυξη γνώσης. Το δεύτερο στάδιο είναι η διαδικασία αναγνώρισης των ριζών των λέξεων (stemming). Σε αυτό το στάδιο αντιστοιχούνται σε μόνο μία οι λέξεις που έχουν την ίδια ρίζα, αλλά διαφέρουν στον ή την πτώση που βρίσκονται.

Η αναπαράσταση κειμένων ,γίνεται με την δημιουργία ενός διανυσματικού χώρου, στον οποίο κάθε κείμενο αποτελεί και ένα διάνυσμα. Αυτό μας επιτρέπει να αναπαραστήσουμε τα κείμενα σαν ένα σύνολο όρων, όπου μπορεί ο καθένας να έχει διαφορετικό βάρος. Πριν το στάδιο της αναπαράστασης πρέπει να έχει γίνει η προεπεξεργασία, καθώς ο διανυσματικός

χώρος θα έχει διαστάσεις όσο είναι οι διαφορετικοί όροι στο κείμενο. Οι δύο βασικοί τρόποι αναπαράστασης κειμένων είναι η αναπαράσταση Boolean, στην οποία η τιμή που μπορεί να πάρει η κάθε διάσταση είναι {0,1} ανάλογα με τον αν υπάρχει αντίστοιχα ένας όρος στο κείμενο ή όχι. Δηλαδή αν υποθέσουμε ότι έχουμε πέντε διαφορετικούς όρους στο κείμενο d1={t1,t2,t3,t4,t5} τότε θα έχουμε και πέντε διαστάσεις D=5 και μπορεί να αναπαρασταθεί σαν Boolean διάνυσμα ως d1=(1,1,1,1,1). Αν όμως το κείμενο d2={t2.t3,t4,t5} τότε στον ίδιο χώρο, θα έχει Boolean αναπαράσταση ως d2=(0,1,1,1,1). Ο άλλος τρόπος αναπαράστασης κειμένων είναι η αναπαράσταση βασισμένη στον αριθμό εμφανίσεων των όρων στο κείμενο. Οι διαστάσεις του διανύσματος θα είναι ίδιες με αυτές στη Boolean αναπαράσταση, αλλά οι τιμές για κάθε διάσταση θα είναι ένας αριθμός που θα προκύπτει σε σχέση με τον αριθμό εμφανίσεων του αντίστοιχου όρου στο κείμενο.

Η εξαγωγή χαρακτηριστικών γνωρισμάτων, είναι το τελευταίο και το πιο σημαντικό στάδιο, πριν την εφαρμογή της εξόρυξης γνώσης από κείμενα. Σε αυτό το στάδιο διατηρούμε μόνο τους καλύτερους όρους, που σύμφωνα με κάποιο μέτρο ποιότητας θα αποτελούν τους όρους που περιγράφουν καλύτερα το κείμενο. Έτσι θα μειώσουμε τον μεγάλο χώρο μνήμης που χρειάζονται τα διανύσματα των κειμένων για να αποθηκευτούν. Συνεπώς θα μπορούμε πιο εύκολα και γρήγορα να έχουμε την εξαγωγή γνώσης. Μερικά από τα πιο δημοφιλή μέτρα ποιότητας είναι το TF-IDF και το CHI.

Στην μέθοδο TF-IDF έχουμε την ποσότητα TF που μετρά την συχνότητα εμφανίσεις των όρων t_i ενός κειμένου d_k και συμβολίζεται με TF(d_k , t_i). Επίσης έχουμε την ποσότητα IDF που λειτουργεί σαν ένα βάρος σημαντικότητας ενός όρου ως προς το κείμενο, σε σχέση όμως με ολόκληρη την συλλογή κειμένων που ανήκει αυτό. Με αυτό τον τρόπο διακρίνονται οι όροι που χαρακτηρίζουν καλύτερα το κείμενο. Το IDF(t_i) για έναν όρο θα είναι μεγάλο αν ο όρος αυτός εμφανίζεται σε λίγα κείμενα. Σύμφωνα με τον τύπο της μεθόδου TF(d_k , t_i) * IDF(t_i) θα έχουμε μεγάλο βάρος σε έναν όρο που εμφανίζεται συχνά σε ένα κείμενο d_k και που συνολικά στην συλλογή είναι σπάνιος.(Kaufmann,M. & Chakrabarti,S. 2002)

3.12 Συσταδοποιήση εγγράφου από τον παγκόσμιο ιστό

Η συσταδοποίηση είναι μια τεχνική ανάλυσης των δεδομένων που μπορεί να αυξήσει την αποδοτικότητα, την αποτελεσματικότητα και να ταξινομήσει τα αποτελέσματα της ανάκτησης. Στις συλλογές εγγράφων για να γίνει τμηματοποίηση συνήθως χρησιμοποιείται το μοντέλο Vector Space το οποίο αναπαριστά τα έγγραφα σαν ένα διάνυσμα χαρακτηριστικών , το οποίο στην συνέχεια χρησιμοποιείται σαν είσοδο στην συσταδοποίηση μαζί με κάποιον αλγόριθμο που επιλέγεται με βάση τα χαρακτηρίστηκα γνωρίσματα των κειμένων ή τις διαφορετικές προσεγγίσεις που μπορεί να είναι λέξεις , φράσεις, σύνδεσμοι ή να είναι υβριδικές και να λαμβάνουν υπόψη τους και το περιεχόμενο και τους συνδέσμους των εγγράφων .Οι αλγόριθμοι αυτοί χωρίζονται στις κατηγορίες, διαιρετικοί, ιεραρχικοί, βασιζόμενοι σε γράφους, βασιζόμενοι σε νευρωτικά δίκτυα και οι βασιζόμενοι σε πιθανότητες.(Bing Liu & Springer 2011)

3.13 Εξόρυξη γνώσης από δεδομένα του παγκόσμιου ιστού - εξατομίκευση

Εξατομίκευση του παγκόσμιου ιστού είναι μια διαδικασία ενός διαδικτυακού τόπου να προσαρμόζεται στις ανάγκες του κάθε χρήστη, αναλύοντας την συμπεριφορά του ως προς την πλοήγηση (δεδομένα χρήσης) σε συσχέτιση με διάφορες πληροφορίες που συλλέγονται από το υπόλοιπο web πχ δεδομένα δομής, περιεχομένου και προφίλ χρηστών. Λόγω της ραγδαίας ανάπτυξης του διαδικτύου, αυτή η λειτουργία είναι ένα σημείο αιχμής τόσο για τον κλάδο της έρευνας, όσο και του εμπορίου. Έτσι ο παγκόσμιος ιστός χρειάζεται νέες μεθόδους σχεδιασμού και ανάπτυξης των online υπηρεσιών. Λόγω μεγέθους και πολυπλοκότητας των δομών του παγκόσμιου ιστού όμως, οι χρήστες δυσκολεύονται να βρουν ακριβώς αυτό που ψάχνουν ή λαμβάνουν διαφορετικά αποτελέσματα από αυτά που ψάχνουν. Αυτό θεωρείται μια απαίτηση από τους χρήστες η οποία δεν περνά απαρατήρητη.

Ορισμός λοιπόν της εξατομίκευσης του παγκόσμιου ιστού (web personalizing) είναι η όποια ενέργεια προσαρμογής των πληροφοριών και υπηρεσιών που παρέχονται σε ένα διαδικτυακό τόπο σε συνδυασμό με τα περιεχόμενα και τη δομή του, στις ανάγκες του κάθε χρήστη και τα ενδιαφέροντα του. Δουλεία και “υποχρέωση” του παγκόσμιου ιστού είναι να παρέχει τις πληροφορίες που θέλουν οι χρήστες, χωρίς καν αυτοί να χρειαστεί να το ζητήσουν.

Είναι αναγκαίο σε αυτό το σημείο να τονιστεί η διαφορά μεταξύ της προσαρμογής (customization) της εμφάνισης διαδικτυακού τόπου ανάλογα με το τι θέλει ο κάθε χρήστης και την εξατομίκευση (personalization). Στην περίπτωση της προσαρμογής η εμφάνιση του εκάστοτε διαδικτυακού τόπου ρυθμίζεται σύμφωνα με τις προτιμήσεις του χρήστη (πχ. Χρώμα, αλλαγή εμφάνισης θεματικών περιοχών κτλ). Αυτό πάει να πει ότι με το μπαίνει στο σύστημα ένας εγγεγραμμένος χρήστης (log on), το site θα φορτώνει την αρχική του σελίδα με βάση τις ανάγκες του και τις προτιμήσεις που είχε προσαρμόσει στο παρελθόν. Σε αυτή τη περίπτωση η διαδικασία προσαρμογής γίνεται με δύο τρόπους. Χειροκίνητα και ημιαυτόματα. Στα συστήματα εξατομίκευσης τώρα, οι αλλαγές που αφορούν το περιεχόμενο ή και την δομή ενός site, γίνονται δυναμικά και χωρίς κάποια παρέμβαση από τον χρήστη. (Παπαρριζος,Κ.Ι.2009)

3.14 Διαδικασία και στάδια εξατομίκευσης του παγκόσμιου ιστού.

Η εξατομίκευση του παγκοσμίου ιστού περιλαμβάνει τα εξής:

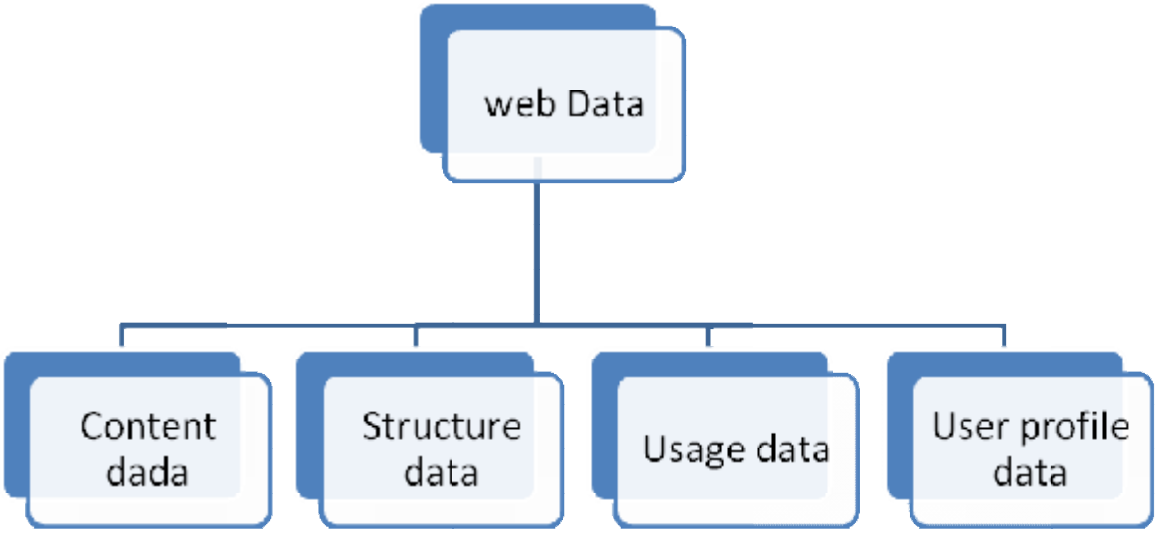
1. Κατηγοριοποίηση και επεξεργασία των δεδομένων του παγκόσμιου ιστού.
2. Εξαγωγή συσχετίσεων μεταξύ των διαφορετικών ειδών αυτών των δεδομένων.
3. Καθορισμός των ενεργειών που προτείνει το σύστημα εξατομίκευσης.

Τα Δεδομένα του παγκόσμιου ιστού που μπορούν να συλλεχτούν και να χρησιμοποιηθούν στην εξατομίκευση του παγκόσμιου ιστού ταξινομούνται σε 4 κατηγορίες:

1. Content dada (Τα δεδομένα περιεχομένου)
2. Structure data (Τα δεδομένα δομής)

3. Usage data (δεδομένα χρήσης)
4. User profile data (δεδομένα προφίλ χρήστη)

Σχήμα 3.1 Κατηγορίες δεδομένων εξατομίκευσης παγκόσμιου ιστού



Η συνολική διαδικασία εξατομίκευσης του παγκόσμιου ιστού βασισμένης στα δεδομένα χρήσης του, διαιρείται σε 5 επιμέρους τμήματα, τα οποία ανταποκρίνονται στα αντίστοιχα στάδια αυτής της διαδικασίας. Αυτά είναι:

1. User profiling (Δημιουργία προφίλ χρηστών): στον παγκόσμιο ιστό, με αυτόν τον όρο ονομάζουμε την διαδικασία συλλογής πληροφοριών άμεσα και έμμεσα για κάθε επισκέπτη. Το προφίλ του χρήστη περιλαμβάνει δημογραφικά στοιχεία σχετικά με αυτόν, ενδιαφέροντά του, μέχρι και το πώς συμπεριφέρεται κατά την πλοήγησή του στο site. Η σωστή εκμετάλλευση αυτών των πληροφοριών οδηγεί στην προσαρμογή του περιεχομένου και την δομή στου site στις συγκεκριμένες και εξατομικευμένες ανάγκες του κάθε επισκέπτη.

2. Log analysis (ανάλυση αρχείων πρόσβασης) και web usage mining (εξόρυξη γνώσης από δεδομένα χρήσης του παγκόσμιου ιστού): σε αυτή τη διαδικασία η πληροφορία που βρίσκεται αποθηκευμένη στα αρχεία πρόσβασης (log files) του εξυπηρετητή (web server) ενός site, επεξεργάζεται εφαρμόζοντας τεχνικές και μεθόδους εξόρυξης δεδομένων με σκοπό: i) την εξαγωγή στατιστικών δεδομένων και την ανακάλυψη ενδιαφερόντων και χρήσιμων προτύπων. ii) την κατάταξη των χρηστών σε ομάδες σύμφωνα με την συμπεριφορά της πλοήγησής τους και iii) την ανακάλυψη πιθανών συσχετισμών μεταξύ ιστοσελίδων και ομάδων χρηστών. Αυτή η διαδικασία εξαγωγής πληροφορίας, σχετικής με την συμπεριφορά των χρηστών στην διάρκεια της πλοήγησης τους, μπορεί να θεωρηθεί ως μέρος της διαδικασίας δημιουργίας του προφίλ του χρήστη.

3. Διαχείριση περιεχομένου**:** η διαδικασία αυτή αφορά την κατάταξη του περιεχομένου ενός site σε σημασιολογικές (semantic) κατηγορίες με σκοπό τη διευκόλυνση της διαδικασίας της ανάκτησης και παρουσίασης των πληροφοριών στον χρήστη. Η διαχείριση του περιεχομένου είναι ιδιαίτερα σημαντική ειδικά σε sites που το περιεχόμενό τους αυξάνεται ή

αλλάζει σε καθημερινή βάση, όπως πχ οι πύλες ενημέρωσης και αναζήτησης (portals) το Facebook κτλ.

4. Web site publishing (παρουσίαση δικτυακού τόπου): ο μηχανισμός αυτός χρησιμοποιείται στην παρουσίαση του περιεχομένου που είναι αποθηκευμένο τοπικά σε έναν εξυπηρετητή ή/και δεδομένων που προέρχονται από άλλες πηγές δικτύου, με ομοιομορφία στον τελικό χρήστη.

5. Απόκτηση πληροφορίας και αναζήτηση: υπάρχουν περιπτώσεις που τα δεδομένα που παρέχονται από ένα site δεν είναι αποθηκευμένα στον εξυπηρετητή του site. Στην περίπτωση μιας δικτυακής πύλης το ενδιαφέρον των χρηστών τείνει σε πληροφορίες που προέρχονται και από άλλες πηγές. Οι υπεύθυνοι του site έχουν υποχρέωση να ψάξουν στον παγκόσμιο ιστό για τέτοιου είδους περιεχόμενο, το οποίο θα πρέπει μετά να ταξινομηθεί σε θεματικές κατηγορίες. Τόσο κατά τη διαδικασία απόκτησης συναφούς πληροφορίας, όσο και κατά την παρουσίαση των κατάλληλων δεδομένων σε αντίστοιχες ομάδες χρηστών, χρησιμοποιούνται τεχνικές αναζήτησης (searching) και κατάταξης (ranking) με βάση τη σχετικότητα.

Και έτσι ένα σύστημα εξατομίκευσης του παγκόσμιου ιστού που βασίζεται στην χρήση του χρησιμοποιεί αυτά τα δεδομένα έτσι ώστε να τροποποιήσει ένα site. Η εξατομίκευση των sites επιτυγχάνεται μέσω της αλληλεπίδρασης των παραπάνω τμημάτων.

3.15 Εφαρμογές web usage mining

Σκοπός του web usage mining είναι να συλλέγει και να συνδυάζει πληροφορίες που αφορούν τα πρότυπα των περιηγήσεων των χρηστών μέσα στις ιστοσελίδες. Από αυτό βγαίνουν συμπεράσματα που αποσκοπούν στη βελτίωση των ιστοσελίδων αλλά και την ομαδοποίηση των χρηστών αναλόγως με τα ενδιαφέροντά τους. Τα αποτελέσματα που παράγονται από την ανάλυση των web log αρχείων χρησιμοποιούνται για πολλούς σκοπούς όπως πχ στη προσωποποίηση του περιεχομένου των σελίδων, στη βελτίωση της εμπειρίας της περιήγησης των χρηστών με προανάκληση (prefetching) και εναποθήκευση (caching) δεδομένων (βάζοντας έναν προσωπικό κωδικό το site προσαρμόζεται στις προτιμήσεις που έχει δείξει ότι έχει ή έχει δηλώσει), στη βελτίωση σχεδίασης των σελίδων και τέλος σε σελίδες σαν αυτές του ηλεκτρονικού εμπορίου. Παρακάτω αυτά αναλύονται.

1) Προσωποποίηση του περιεχομένου: Οι τεχνικές που υπάρχουν χρησιμοποιούνται με σκοπό να παρέχουν προσωποποιημένα δεδομένα στου χρήστες δίνοντας τους την αίσθηση πως το περιεχόμενο στις σελίδες είναι ειδικά για αυτούς και τα ενδιαφέροντά τους. Πχ μπορεί να προβλεφθεί η συμπεριφορά ενός χρήστη σε πραγματικό χρόνο εάν συγκριθούν τα ήδη υπάρχοντα πρότυπα περιήγησης με τα τυπικά πρότυπα που έχουν παλαιότερα εξαχθεί στο παρελθόν από τα web log αρχεία. Σε αυτήν την περιοχή τα συστήματα προτείνουν (recommendation systems) σε χρήστες συνδέσμους με περιεχόμενο που πιθανό τους ενδιαφέρει. Οι προσωποποιημένοι χάρτες είναι ένα σύστημα που προτείνει συνδέσμους. Πχ ένα τέτοιο σύστημα είναι το Google. Κάθε υπηρεσία που προσφέρεται εκεί, έχει ειδική περιοχή που προτείνει συνδέσμους με χρήση αρχείων cookies. Ακόμα και οι αναζητήσεις στη μηχανή αναζήτησής της είναι προσωποποιημένες διαφορετικά ανάλογα τον υπολογιστή που τις εμφανίζει. Λαμβάνει υπ' όψιν παλιότερες αναζητήσεις και δίνει αντίστοιχα αποτελέσματα

στις μελλοντικές, που είναι πιο ενδιαφέροντα για κάθε χρήστη. Άλλο παράδειγμα είναι αυτό των ηλεκτρονικών καταστημάτων που επεξεργάζονται προηγούμενες αγορές του χρήστη και του προτείνει προσφορές της κατηγορίας προϊόντων που προτιμάει.

2) Προανάκληση και εναποθήκευση των δεδομένων: Αποτελέσματα προερχόμενα της εφαρμογής τεχνικών web usage mining χρησιμοποιούνται για τη βελτίωση της απόδοσης του διακομιστή και γενικά των εφαρμογών διαδικτύου. Τυπικά οι τεχνικές αυτές χρησιμοποιούνται για τη δημιουργία των κατάλληλων στρατηγικών προανάκλησης και εναποθήκευσης δεδομένων για τη μείωση του χρόνου απόκρισης των διακομιστών.

3) Υποστήριξη πάνω στο σχεδιασμό σελίδων: Η ευχρηστία (usability) είναι ένα βασικό ζήτημα στο σχεδιασμό και την υλοποίηση των σελίδων. Με τις τεχνικές που υπάρχουν μπορούμε να δώσουμε κατευθύνσεις στη βελτίωση του σχεδιασμού των εφαρμογών διαδικτύου. Στις προσαρμοστικές σελίδες η δομή και το περιεχόμενο αλλάζει δυναμικά, προσαρμόζεται και αναδιοργανώνεται ανάλογα πάντα με την συμπεριφορά του χρήστη.

4) Στο ηλεκτρονικό εμπόριο: Η εξόρυξη γνώσης στις εμπορικές σελίδες είναι μια πολύ σημαντική δυνατότητα ως προς τη βελτίωση των παρεχόμενων υπηρεσιών, τη προσφερόμενη ικανοποίηση του κάθε πελάτη ξεχωριστά και φυσικά στην αύξηση των κερδών της επιχείρησης. Η διαχείριση της σχέσης πελάτη-επιχείρηση (customer relationship management) υποστηρίζεται με τη χρήση τεχνικών web usage mining. Με αυτή τη λογική δίνεται έμφαση:

- Στη προσέλκυση πελατών
- Στη διατήρηση πελατών
- Στην ανταλλαγή πωλήσεων
- Στην ενεργή παρουσία πελατών

Πίνακας 3.1 Συσχέτιση μεταξύ τεχνικών και εφαρμογών

Προσωποποίηση	Συσταδοποίηση, Πρότυπα Ακολουθιών, Ασαφής Συσταδοποίηση, Κανόνες συσχέτισης, Σχεσιακά Μοντέλα Marcov
Εναποθήκευση	Κανόνες Αυτοσυσχέτισης, Σχεσιακά Μοντέλα Marcov
Σχεδιασμός	Μοντέλα Marcov, Πρότυπα Ακολουθιών, Κατηγοριοποίηση, Κανόνες συσχέτισης

Ηλεκτρονικό Εμπόριο	Κατηγοριοποίηση, Πρότυπα Ακολουθιών, Ασαφής Λογική, Συσταδοποίηση, Γενετικοί Αλγόριθμοι

Ο παραπάνω πίνακας είναι ένα παράδειγμα κάποιων εφαρμογών και τις τεχνικές που έχουν χρησιμοποιηθεί Βλέπουμε οτι δεν υπάρχει κάποια αυστηρή συσχέτιση μεταξύ τεχνικών και εφαρμογών. Οι πιο πολλές από τις προσεγγίσεις χρησιμοποιούνται στις πιο πολλές εφαρμογές.

3.16 Σύστημα Εξατομίκευσης του Παγκόσμιου Ιστού

Η εξατομίκευση ενός site ορίζεται ως η διαδικασία προσαρμογής του περιεχομένου και της δομής αυτού στις συγκεκριμένες απαιτήσεις και ανάγκες κάθε χρήστη, χρησιμοποιώντας τα δεδομένα που προκύπτουν από τη συμπεριφορά πλοήγησης του χρήστη. Τα βήματα αυτής της διαδικασίας είναι τα εξής:
α) συλλογή δεδομένων από τον Παγκόσμιο Ιστό,
β) μοντελοποίηση και κατηγοριοποίηση αυτών των δεδομένων (φάση προεπεξεργασίας)
γ) ανάλυση αυτών των δεδομένων
δ) καθορισμός των ενεργειών που πρέπει να γίνουν.

Οι διαφορετικές πρακτικές που χρησιμοποιούνται για την ανάλυση των δεδομένων που έχουν συλλεχθεί είναι:

- το φιλτράρισμα σύμφωνα με το περιεχόμενο (content-based filtering),
- το συνεργατικό φιλτράρισμα (collaborative filtering),
- το φιλτράρισμα με βάση κανόνες (rule-based filtering)
- η εξόρυξη γνώσης από τη χρήση του Παγκόσμιου Ιστού (Web usage mining).

Ο δικτυακός τόπος μπορεί να εξατομικευτεί κάνοντας πιο ενεργούς τους ήδη υπάρχοντες υπέρ-συνδέσμους, εισάγοντας δυναμικά νέους υπέρ-συνδέσμους οι οποίοι εκτιμάται ότι ενδιαφέρουν τον χρήστη ή δημιουργώντας νέες σελίδες περιεχομένων.

Τα συστήματα που βασίζονται στο φιλτράρισμα σύμφωνα με το περιεχόμενο (content based filtering), βασίζονται μόνο στις προτιμήσεις εκάστοτε χρήστη. Το σύστημα “διαβάζει” τη συμπεριφορά του κάθε χρήστη κατά την πλοήγηση και την συσχετίζει τα συγκεκριμένα αντικείμενα με εκείνα που προτιμήθηκαν από το χρήστη στο παρελθόν.
Τα συστήματα συνεργατικού φιλτραρίσματος (collaborative filtering) καλούν τους χρήστες να βαθμολογήσουν αντικείμενα ή να δηλώσουν τα ενδιαφέροντα και τις προτιμήσεις τους και στη συνέχεια επιστρέφουν πληροφορίες οι οποίες ενδέχεται ότι θα τους ενδιαφέρουν. Αυτή η διαδικασία στηρίζεται στην υπόθεση ότι χρήστες με παρόμοια συμπεριφορά πλοήγησης (π.χ. χρήστες που δείχνουν προτίμηση σε παρόμοια αντικείμενα) έχουν και αντίστοιχα ενδιαφέροντα.

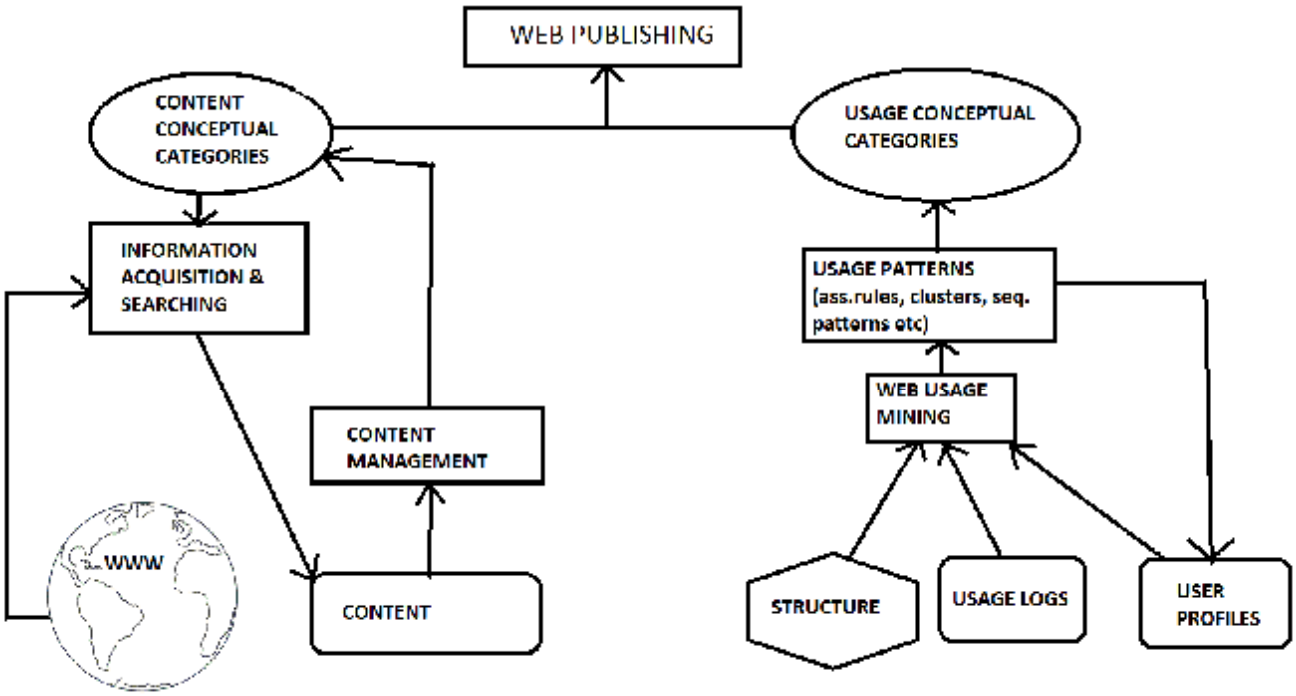
Στο φιλτράρισμα βασισμένο σε κανόνες (rule-based filtering) οι χρήστες καλούνται να απαντήσουν σε ένα σύνολο ερωτημάτων. Αυτά τα ερωτήματα παράγονται από δέντρα

αποφάσεων. Έτσι, καθώς ο χρήστης προχωρά την διαδικασία απάντησης σε κάθε ερώτημα, αυτό που τελικά λαμβάνει ως αποτέλεσμα (π.χ. μια λίστα ηλεκτρικών προϊόντων) είναι προσαρμοσμένο στις ανάγκες του. Οι διάφορες τεχνικές φιλτραρίσματος (του περιεχόμενου, των κανόνων και το συνεργατικό) συνδυάζονται επίσης για την παραγωγή και πιο έγκυρων αποτελεσμάτων.

· Μια κλασσική μέθοδος που ακολουθούν πολλοί ερευνητές για την επιτυχή εξατομίκευση του Παγκόσμιου Ιστού είναι η τεχνική της εξόρυξης δεδομένων χρήσης του Παγκόσμιου Ιστού (Web usage mining). Η διαδικασία αυτή στηρίζεται στην εφαρμογή μεθόδων εξόρυξης δεδομένων και στατιστικών μεθόδων από αρχεία πρόσβασης (Web logs), οι οποίες βοηθούν στην εξαγωγή ενός συνόλου μοτίβων (patterns) που υποδεικνύουν τη συμπεριφορά πλοήγησης του χρήστη. Οι μέθοδοι εξόρυξης δεδομένων που χρησιμοποιούνται είναι:

· η εξαγωγή κανόνων συσχέτισης (association rules mining),
· η ανακάλυψη ακολουθιακών προτύπων (sequential pattern discovery),
· η συσταδοποίηση (clustering)
· η κατηγοριοποίηση (classification).

Αυτή η γνώση χρησιμοποιείται στη συνέχεια από το σύστημα έτσι ώστε να εξατομικεύσει το site σύμφωνα με τα προφίλ και τις συμπεριφορά των χρηστών.



Σχήμα 3.1 Αναπαράσταση εξατομίκευσης στον Π.Ι.

Το παραπάνω διάγραμμα αναπαριστά τη λειτουργική αρχιτεκτονική ενός συστήματος εξατομίκευσης του παγκόσμιου ιστού αποτελούμενο από τα τμήματα και τις πηγές δεδομένων που προαναφέρθηκαν. Το τμήμα της διαχείρισης του περιεχομένου (content management) επεξεργάζεται το περιεχόμενο του site και το ταξινομεί σε παρόμοιες κατηγορίες. Το περιεχόμενο του site μπορεί να εμπλουτισθεί με επιπλέον πληροφορίες που προέρχονται από άλλες δικτυακές πηγές, χρησιμοποιώντας κι άλλες προχωρημένες τεχνικές αναζήτησης.

Με δεδομένα τη δομή του site και των αρχείων πρόσβασης, ένας αλγόριθμος εξόρυξης γνώσης παρέχει, από τη χρήση του παγκόσμιου ιστού, τα αποτελέσματα σχετικά με μοτίβα χρήσης (usage patterns), τις συμπεριφορές πλοήγησης των χρηστών, τις ομάδες χρηστών και των συνόδων (sessions) καθώς επίσης και σχετικές πληροφορίες με τη σειρά επιλογών (clickstream). Μέσω της διαδικασίας της δημιουργίας προφίλ χρήστη μπορούν να αποκτηθούν ακόμα περισσότερα δεδομένα για κάθε χρήστη. Επίσης τα δεδομένα που εξάγονται μέσω αυτής της διαδικασίας και αφορούν στη συμπεριφορά πλοήγησης του χρήστη, μπορούν στη συνέχεια να προστεθούν σαν επιπλέον χαρακτηριστικά στο προφίλ του. Όλη αυτή η πληροφορία σχετίζεται με κόμβους, συνδέσμους, τυπικές συμπεριφορές, περιεχόμενα δικτυακού τόπου και πρότυπα, μεταφέρεται σε ένα υψηλότερο επίπεδο επεξεργασίας και τέλος ταξινομείται στις ανάλογες κατηγορίες.

Οποιαδήποτε πληροφορία που εξάγεται από την σχέση μεταξύ της γνώσης που πηγάζει από τη διαχείριση του περιεχομένου και της γνώσης που προκύπτει από τη χρήση μεθόδων εξόρυξης δεδομένων δίνει στη συνέχεια το πλαίσιο για την αξιολόγηση των πιθανών εναλλακτικών προτάσεων προκειμένου να γίνει η αναδόμηση, η τροποποίηση και η εξατομίκευση του site. Ένας μηχανισμός παρουσίασης θα πραγματοποιήσει την αλλαγή του περιεχομένου ή/και της δομής του site, εξασφαλίζοντας στον κάθε χρήστη ξεχωριστά να πλοηγείται μέσω της βέλτιστης δομής για αυτόν. Το διαθέσιμο περιεχόμενο θα παρουσιαστεί σύμφωνα με τα ενδιαφέροντα του κάθε χρήστη.

3.17 Δημιουργία Προφίλ Χρήστη

Προκειμένου να επιτευχθεί η εξατομίκευση, το σύστημα θα πρέπει να μπορεί να ξεχωρίζει διαφορετικούς χρήστες ή ομάδες χρηστών. Αυτή η διαδικασία ονομάζεται δημιουργία προφίλ χρήστη και ο λόγος ύπαρξης της είναι η δημιουργία μιας βάσης δεδομένων που περιέχει τις προτιμήσεις, τα χαρακτηριστικά και τις δραστηριότητες των χρηστών. Η διαδικασία αυτή έχει αναπτυχθεί σημαντικά στην περιοχή του παγκόσμιου ιστού και ιδιαίτερα στο ηλεκτρονικό εμπόριο, αφού οι τεχνολογίες του διαδικτύου κάνουν πιο εύκολη τη διαδικασία συγκέντρωσης πληροφοριών, σχετικές με τους χρήστες ενός δικτυακού τόπου, οι οποίοι ειδικά στην περίπτωση μιας ηλεκτρονικής επιχείρησης είναι και πιθανοί πελάτες.

Ένα προφίλ χρήστη μπορεί να είναι:

· στατικό, όταν οι πληροφορίες που περιέχει αλλάζουν σπανίως ή και ποτέ (π.χ. δημογραφικά στοιχεία)

· δυναμικό, όταν τα δεδομένα του προφίλ αλλάζουν συχνά. Τέτοια δεδομένα αποκομίζονται είτε άμεσα, με τη χρήση online φορμών εγγραφής και ερωτηματολογίων,

φτάνοντας με αυτόν τον τρόπο σε στατικά προφίλ χρηστών, είτε έμμεσα, καταγράφοντας συμπεριφορά πλοήγησης και προτιμήσεις του κάθε χρήστη, δημιουργώντας έτσι τα δυναμικά προφίλ χρηστών. Σε αυτή τη περίπτωση, υπάρχουν δύο επιπλέον επιλογές:

1. ο χρήστης θεωρείται ως μέλος μιας ομάδας οπότε οι αλλαγές να απευθύνονται σε ολόκληρες τις ομάδες χρηστών.

2. ο χρήστης αναγνωρίζεται ως οντότητα, οπότε οι όποιες αλλαγές να απευθύνονται σε αυτόν μόνο.

Όταν απευθυνόμαστε στους χρήστες ως ομάδα, χρησιμοποιούμε τη μέθοδο της δημιουργίας αθροιστικών προφίλ χρηστών. Αυτά δημιουργούνται βάσει κανόνων και προτύπων που εξάγονται εφαρμόζοντας τεχνικές εξόρυξης γνώσης στο παγκόσμιο ιστό σε αρχεία πρόσβασης. Με αυτή τη γνώση, το site μπορεί να προσαρμοσθεί στις απαιτήσεις των χρηστών.

3.18 Συλλογή των Δεδομένων

Ένας τρόπος αναγνώρισης ενός επισκέπτη κατά τη διάρκεια μιας συνόδου (session), είναι η χρησιμοποίηση των cookies. Τα Cookies είναι τα δεδομένα που στέλνει ένας εξυπηρετητής (Web server) σε έναν πελάτη (Web client), αποθηκεύονται τοπικά και αποστέλλονται πίσω στον εξυπηρετητή στις επόμενες αιτήσεις. Ένα cookie είναι απλώς μια HTTP επικεφαλίδα η οποία αποτελείται από μια συμβολοσειρά κειμένου που εισάγεται στη μνήμη ενός browser. Χρησιμοποιείται για να αναγνωρίζεται μοναδικά ένας χρήστης κατά τη διάρκεια ενεργειών μέσα σε ένα site και περιέχει παραμέτρους που επιτρέπουν στον απομακρυσμένο HTTP web server να διατηρεί ένα αρχείο με την “ταυτότητα” του χρήστη και τις ενέργειες που αυτός κάνει στο απομακρυσμένο site.
Με βάση αυτά αποθηκεύονται δεδομένα σχετικά με την αναγνώριση του χρήστη (id) καθώς και δεδομένα σχετικά με τον κωδικό πρόσβασης του (password). Στο αρχείο επίσης να περιέχονται επιπλέον πληροφορίες, πχ στοιχεία πιστωτικών καρτών, λεπτομέρειες που αφορούν τις δραστηριότητες του χρήστη στο δικτυακό τόπο, πχ ποιες σελίδες επισκέφτηκε ή ποιες διαφημίσεις είδε. Συχνά τα cookies “δείχνουν” σε πιο λεπτομερή στοιχεία του πελάτη, τα οποία είναι αποθηκευμένα στον εξυπηρετητή.

Τα δεδομένα για τη δημιουργία ενός προφίλ χρήστη μπορούν να αποκτηθούν και άμεσα, χρησιμοποιώντας online φόρμες εγγραφής που ζητούν πληροφορίες σχετικές με τον επισκέπτη όπως όνομα, ηλικία, φύλλο, αρέσκειες και δυσαρέσκειες. Αυτά αποθηκεύονται σε μια βάση δεδομένων και κάθε φορά που ο χρήστης εισέρχεται στο site, ενημερώνονται σύμφωνα με την συμπεριφορά του. Σε κάθε τέλος συνεδρίας του χρήστη υπάρχει όλο και πιο σαφή εικόνα των προτιμήσεών του.

Όλες αυτές οι τεχνικές για τη δημιουργία προφίλ χρηστών έχουν μειονεκτήματα. Πχ σε περίπτωση που ο χρήστης έχει απενεργοποιήσει τα cookies στο browser του ,ενώ το σύστημα βασίζεται σε αυτά, δεν αφήνει την δυνατότητα στο σύστημα να καταγράψει την συμπεριφορά του. Άλλο πρόβλημα μπορεί να είναι ότι από τη στιγμή που ένα αρχείο cookie είναι αποθηκευμένο τοπικά στον υπολογιστή του χρήστη, ο χρήστης είναι πιθανό να το σβήσει και έτσι την επόμενη φορά που θα επισκεφθεί το δικτυακό τόπο να θεωρηθεί ως νέος. Επίσης πρόβλημα προκύπτει αν χρησιμοποιούν περισσότεροι από ένας χρήστης τον ίδιο υπολογιστή

και ταυτόχρονα δεν υπάρχει η πληροφορία του logon id (εγγεγραμμένο μέλος). Από την άλλη πλευρά, όταν τα δεδομένα συλλέγονται μέσω αιτήσεων ή ερωτηματολογίων, πολλοί χρήστες συμπληρώνουν ψευδή στοιχεία κάτι που οδηγεί στη δημιουργία λανθασμένων προφίλ για το σύστημα. Επίσης όταν πρέπει να φτιάξει ο ίδιος ο χρήστης το προφίλ του, μπορεί να το αποφύγει.

3.19 Πηγές Δεδομένων

Οι βασικές πηγές δεδομένων είναι τρεις:

- Web servers,
- Proxy servers
- Web clients.

Τα δεδομένα από τους Web servers είναι και τα περισσότερα. Είναι κυρίως (μεγάλα) logs των αιτήσεων που δέχονται από απομακρυσμένους υπολογιστές, ή και logs από τις βάσεις δεδομένων τις οποίες χρησιμοποιούν. Επίσης περιλαμβάνουν κατά κανόνα την IP του υπολογιστή του, ενδεχομένως το όνομά του, την ημερομηνία, την ώρα και την ακριβή αίτηση που δέχθηκαν από τον πελάτη. Στις περιπτώσεις μεγάλων μεγεθών, τα δεδομένα αυτά συχνά τηρούνται απευθείας σε βάσεις δεδομένων και όχι σε απλά text αρχεία. Κρίσιμο σημείο για την εκμετάλλευση των logs αυτών είναι η δυνατότητα για αναγνώριση των συνεδριών του χρήστη, δηλαδή η αναγνώριση και ομαδοποίηση όλων των αιτημάτων, κινήσεις κλπ που έκανε ένας χρήστης ώστε να γίνει εμφανές το μονοπάτι που ακολούθησε κατά την πλοήγησή του στο site.

Μια μέθοδος που διευκολύνει αυτή την αναγνώριση είναι μέσω των cookies, όταν αυτά είναι διαθέσιμα. Ακόμα όμως και με τη χρήση των cookies δεν γίνεται πάντα ακριβής ο καθορισμός του μονοπατιού που ακολούθησε ο κάθε χρήστης, καθώς η κίνηση "back" στους browsers δεν είναι ορατή στους servers και δεν καταγράφεται. Σε περιπτώσεις που τα cookies δεν είναι διαθέσιμα, επιστρατεύονται πιο εξεζητημένες τεχνικές για την αναγνώριση αυτή.

Πέρα όμως από τα logs, τα server-side δεδομένα περιλαμβάνουν και χαμηλότερου επιπέδου στοιχεία, όπως τα πακέτα TCP/IP. Η συλλογή και διαχείριση τέτοιων δεδομένων επιβαρύνει τη λειτουργία των server, αλλά υπάρχουν κάποια οφέλη τα οποία δεν μπορούν να παρέχουν τα logs: πχ μπορεί να ενσωματωθεί η δραστηριότητα σε διαφορετικούς web server σε ένα συνολικό αρχείο ή μπορούν να εντοπιστούν κινήσεις του πελάτη που δεν γίνονται ορατές στα logs όπως η επιλογή "stop".

Μειονέκτημα της εξέτασης των TCP/IP πακέτων είναι ότι αδυνατούν να καταγράψουν ωφέλιμα δεδομένα από κρυπτογραφημένα πακέτα, πράγμα που συμβαίνει συνήθως σε εμπορικές εφαρμογές.

Τα δεδομένα που συγκεντρώνονται σε κεντρικούς servers αφορούν συνήθως ομάδες χρηστών που έχουν πρόσβαση σε μεγάλες ομάδες εξυπηρετητών. Αυτοί οι servers παρέχονται από τους διαφόρους ISPs (internet service providers) για να παρέχουν βελτιωμένη ταχύτητα στους συνδρομητές τους. Τα δεδομένα αυτά παρουσιάζουν εγγενή αδυναμία για αναγνώριση των συνεδριών των χρηστών, παρόλα αυτά, αν μεταξύ του server και του πελάτη δε μεσολαβεί άλλο caching (αποθήκευση σε μνήμη cache) τότε είναι δυνατό να προκύψουν επαρκεί στοιχεία για την αναγνώριση συνεδριών.

Τέλος τα client-side δεδομένα προέρχονται είτε μέσω JavaScript, είτε μέσω java applet, είτε κι από την τροποποίηση ή επέκταση των ίδιων των browsers. Τα δεδομένα αυτά είναι ποιοτικά ανώτερα από τα προηγούμενα καθώς δεν τίθεται πρόβλημα αναγνώρισης των συνεδριών, ενώ ταυτόχρονα αντιμετωπίζεται το caching αλλά είναι και δυνατή η καταγραφή συμπεριφοράς που δεν φτάνει στον web server. Η καταγραφή και χρήση τέτοιων δεδομένων απαιτεί τη συναίνεση του χρήστη γιατί θίγονται σημαντικά ζητήματα σχετικά με την προστασία της ιδιωτικής ζωής.

3.20 Διαφύλαξη στα προσωπικά δεδομένα

Οι μέθοδοι αλλά κυρίως τα εργαλεία του Web Usage Mining συνδυάζουν διάφορες πηγές δεδομένων ώστε να μπορούν να καταγράφουν , να κατατάσσουν και να διαμορφώνονται ανάλογα με τη συμπεριφορά του κάθε χρήστη με τη περισσότερη δυνατή ακρίβεια. Και εδώ θίγεται το μεγάλο ζήτημα προστασίας των προσωπικών δεδομένων των χρηστών, ένα ζήτημα το οποίο αφορά γενικά την περιοχή του data mining. Οι χρήστες γενικά διστάζουν να μπουν σε sites με cookies ή τα απενεργοποιούν ή δίνουν τα ψευδή στοιχεία που είπαμε παραπάνω και αυτό επειδή ξέρουν ότι έτσι χάνουν την ανωνυμία που ενδεχομένως θέλουν και μετά καταγράφεται η όποια κίνηση κάνουν μέσα στο site ή οποίες μπορούν και να χρησιμοποιηθούν αλλού, ακόμα και χωρίς την συγκατάθεσή τους. Επίσης ακόμα και αν έχουν συμφωνήσει στο να δοθούν τα στοιχεία τους σε ένα site, αυτά μπορούν να δοθούν σε άλλα sites μέσω των cookies και έτσι αποκαλύπτονται σε τρίτους χωρίς την συγκατάθεσή τους. Στις ερευνητικές εργασίες δε γίνεται τόσο συχνά αναφορά σε θέματα privacy αν και τόσο στην Ε.Ε, όσο και στις ΗΠΑ έχουν ψηφιστεί ιδιαίτερα αυστηρούς νόμους ως προς την συλλογή και τον χειρισμό τέτοιων δεδομένων. Στη θέση της προστασίας αυτών το δεδομένων και γενικότερα της ιδιωτικής ζωής των χρηστών υπάρχει η Πλατφόρμα για Επιλογές Ιδιωτικότητας P3P (Platform for Privacy Preferences). Το σύστημα λειτουργεί ως εξής: στα web sites υπάρχουν πολιτικές προστασίας προσωπικών δεδομένων και τις μεταδίδουν στους πελάτες με τρόπο που μπορεί να γίνει αντιληπτός ώστε να αξιολογηθούν από τους browsers κλπ. Αν τα κριτήρια που έχει θέσει ο χρήστης πληρούνται τότε η πολιτική γίνεται αποδεκτή και η συνεδρία συνεχίζει. Παρόλα αυτά το P3P δεν λύνει ούτε τα ζητήματα της τήρησης των πολιτικής που ισχυρίζεται το κάθε site, ούτε αντιμετωπίζει το θέμα της εξόρυξης πληροφοριών από τα δεδομένα που συλλέγονται, που είναι και το βασικό για το data mining.

3.21 Σημαντικοί αλγόριθμοι Εξόρυξης Γνώσης

Οι 10 κατά σειρά καλύτεροι αλγόριθμοι που χρησιμοποιούνται για την εξόρυξη δεδομένων είναι (Πιτούρα,Ε. 2010) :

- C4.5(61 votes) – Κατηγοριοποίηση (δέντρο απόφασης)
- K-Means(60 votes)-Συσταδοποίηση
- SVM(58 votes) – Κατηγοριοποίηση (Support vector machine)
- Apriori(52 votes)-Κανόνες συσχέτισης
- EM(48 votes)Στατιστική,συσταδοποίηση(expectation maximization)
- PageRank(46 votes)-Ιστοσελίδες
- AdaBoost(45 votes)-Μετα-ταξινομητης
- kNN(45 votes)-Συσταδοποίηση
- Naïve Bayes(45 votes)-Στατιστική,ταξινόμηση
- CART(35 votes)-Κατηγοριοποιηση (δέντρο απόφασης)

ΚΕΦΑΛΑΙΟ 4.

ΕΞΟΡΥΞΗ ΓΙΑ ΕΚΠΑΙΔΕΥΤΗΚΟΥΣ ΣΚΟΠΟΥΣ-EDUCATIONAL DATA MINING(EDM)

4.1 ΕΙΣΑΓΩΓΗ

Η εξόρυξη γνώσης από εκπαιδευτικά δεδομένα είναι ένας σχετικά νέος κλάδος. Οι πρώτες δημοσιεύσεις επιστημονικών ερευνών έγιναν μετά το 2000 και το πρώτο διεθνές συνέδριο έγινε το 2008 στον Καναδά στην πόλη Μόντρεαλ. (Παπανικολάου Δ.,2010)

Ο τομέας αυτός της εξόρυξης γνώσης ασχολείται με επιστημονικά ερωτήματα που αφορούν την εκπαίδευση όπως είναι η πρόβλεψη της απόδοσης των μαθητών πριν γίνουν εξετάσεις. Αυτό μπορεί να βοηθήσει στην έγκαιρη διάγνωση προβλημάτων που υπάρχουν κατά την εκπαίδευση. Με αυτόν τον τρόπο ο εκάστοτε καθηγητής θα μπορεί να βελτίωση την εκπαίδευση και να βοηθήσει τον εκπαιδευόμενο στις δυσκολίες που τυχόν αντιμετωπίζει.

4.2 ΟΡΙΣΜΌΣ

Εξόρυξη γνώσης για εκπαιδευτικούς σκοπούς είναι η διαδικασία που μετατρέπει τα εκπαιδευτικά δεδομένα σε χρήσιμες πληροφορίες για την βελτιστοποίηση της εκπαίδευσης.

4.3 ΣΤΟΧΟΙ

Ως προς τους υπεύθυνους εκπαίδευσης είναι οι υπολογισμοί παραμέτρων που θα τους δώσουν αποτελέσματα για την οργάνωση του εκπαιδευτικού συστήματος και την αξιολόγηση της αποδοτικότητας των προγραμμάτων που εφαρμόζονται.

Ως προς τους εκπαιδευτές είναι να μπορούν να επιλέγουν τους σωστούς τρόπους εκμάθησης που θα είναι οι πιο αποτελεσματικοί, επίσης να μπορούν να αξιολογήσουν τον τρόπο δομής των μαθημάτων και των δραστηριοτήτων που προτείνονται στους εκπαιδευόμενους καθώς και την κατηγοριοποίηση αυτών ανάλογα με τις ανάγκες τους για την καλύτερη καθοδήγηση τους.

Ως προς τους μαθητές είναι η βελτίωση της μάθησης τους μέσα από την εφαρμογή πρακτικών και δραστηριοτήτων με καλύτερα αποτελέσματα τα οποία έχουν εφαρμοστεί σε άλλους εκπαιδευόμενους.

4.4 ΤΟΜΕΊΣ ΕΦΑΡΜΟΓΉΣ ΕΞΌΡΥΞΗΣ ΓΝΏΣΗΣ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ

Βελτίωση μαθητικών μοντέλων. Με την μοντελοποίηση των χαρακτηριστικών ξεχωριστά για κάθε μαθητή ή ομάδα μαθητών εξάγονται συμπεράσματα που οδηγούν στην πρόγνωση μαθητικών προβλημάτων.

Βελτίωση των μεθόδων διδασκαλίας. Ο κάθε τύπος διδασκαλίας επηρεάζει διαφορετικά την πρόοδο του μαθητή ανάλογα με διάφορους παράγοντες όπως είναι το επίπεδο γνώσης που βρίσκονται.

Βοήθεια στις εκπαιδευτικές θεωρίες. Με την εφαρμογή της εξόρυξης διακρίνονται οι παράγοντες που επηρεάζουν τη μάθηση ώστε να γίνει ο σχεδιασμός του κατάλληλου εκπαιδευτικού συστήματος.

Προσωποποιημένη εκπαίδευση**.** Εδώ η εφαρμογή αφορά τον κατάλληλο τρόπο παρουσίασης του μαθήματος ξεχωριστά για κάθε κατηγορία εκπαιδευόμενου.

4.5 ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΑΠΟ ΤΟ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ ΠΟΥ ΕΦΑΡΜΟΖΟΝΤΑΙ ΣΤΗΝ ΕΚΠΑΙΔΕΥΣΗ

Οι Δημοφιλέστεροι αλγόριθμοι που χρησιμοποιούνται για τη κατάλληλη επιλογή μαθημάτων και πρόβλεψης επίδοσης μαθητών (Παπανικολάου Δ.,2010) είναι οι Μέθοδος Chaid, Αλγόριθμος C4.5, ID3, Naïve Bayes και βασίζονται στις τεχνικές εξόρυξης γνώσης:

- Ομαδοποίηση: Εφαρμόζεται για να έχουμε κατά ομάδες τις ιστοσελίδες με παρόμοιο περιεχόμενο εκπαίδευσης και τους εκπαιδευόμενους ανάλογα με τη πλοήγηση τους στο περιεχόμενο των ιστοσελίδων.
- Ταξινόμηση: Εφαρμόζεται για να γίνεται χαρακτηρισμός των ομάδων ανάλογα με τις ιδιότητες που εμφανίζουν οι χρήστες σε κάποιο κοινό εκπαιδευτικό υλικό.
- Εύρεση ακραίων σημείων. Εφαρμόζεται για να ξεχωρίσουν οι τιμές ενός συνόλου δεδομένων που είναι ακραίες, ώστε να εντοπίσουμε τους μαθητές που έχουν μαθησιακές δυσκολίες και αυτούς που είναι σε πολύ προχωρημένο επίπεδο γνώσης.
- Κανόνες συσχέτισης: Χρησιμοποιώντας τα χαρακτηριστικά ενός συνόλου δεδομένων μπορούμε να δούμε ποια εργαλεία χρησιμοποιήθηκαν για την επίλυση μιας άσκησης ανάλογα με το εκπαιδευτικό περιεχόμενο, ώστε να μπορέσει να γίνει καλύτερη καθοδήγηση στους μαθητές.
- Ανακάλυψη σειριακών προτύπων: Εφαρμόζεται σε περιπτώσεις όπως όταν θέλουμε να δούμε ποια από τα περιεχόμενα του μαθήματος οδηγούν τους εκπαιδευόμενους να προσπελάσουν κάποια άλλα.

ΜΕΡΟΣ ΄Β

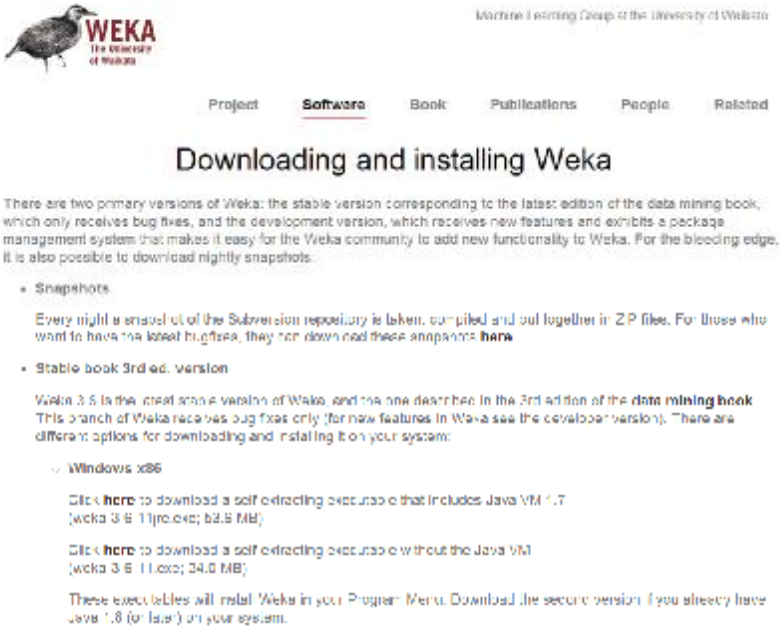
ΑΝΑΛΥΤΙΚΟ ΕΓΧΕΙΡΙΔΙΟ ΧΡΗΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

1.1 Πρόγραμμα WEKA

Το WEKA(Waikato Environment for Knowledge Learning) είναι ένα πρόγραμμα για τους υπολογιστές που αναπτύχθηκε από το Πανεπιστήμιο του Waikato της Νέας Ζηλανδίας με σκοπό την εξόρυξη γνώσης από δεδομένα. Το WEKA υποστηρίζει διάφορες εργασίες της εξόρυξης δεδομένων όπως συσταδοποίηση ,παλινδρόμηση, επεξεργασία δεδομένων και γραφήματα. Η χρήση του βασίζεται κυρίως στην βέλτιστη χρήση του υπολογιστική ισχύ με σκοπό την εκπαίδευση του ώστε να εξελιχτεί με το machine learning και να μας παρέχει πρότυπα και συσχετίσεις. Το WEKA είναι ελεύθερο λογισμικό που έχει πρόσβαση σε αυτό ο καθένας. Το πρόγραμμα είχε αρχικά γραφτεί στην γλώσσα C αλλά στην συνέχεια ξαναγράφτηκε από την αρχή σε γλώσσα Java.

Εγκατάσταση

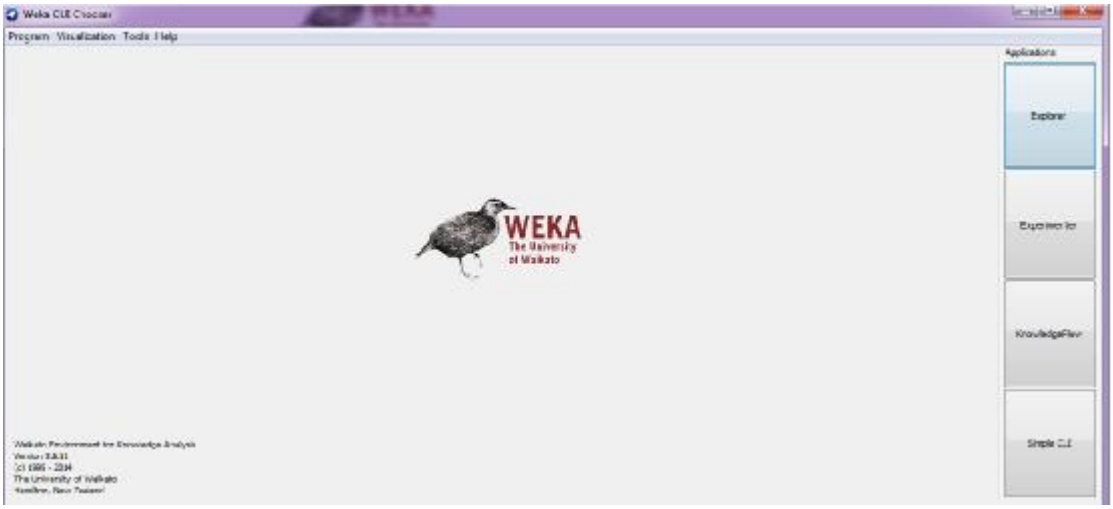
Το πρόγραμμα είναι δωρεάν και ο οποιοσδήποτε μπορεί να το κατεβάσει από την διεύθυνση http://www.cs.waikato.ac.nz/ml/weka. Στο site εκτός από το πρόγραμμα είναι διαθέσιμες και άλλες πληροφορίες όπως :Διάφορα αρχεία που είναι για χρήση για οποιονδήποτε θέλει να ασχοληθεί με την εξόρυξη δεδομένων και συγκεκριμένα το WEKA.



1.2 Επιλογές προγράμματος WEKA

Το WEKA διατίθεται για κάθε λογισμικό (Windows,Mac,Linux) και υπάρχει η επιλογή που περιλαμβάνει το Java Virtual Machine που δίνει την επιλογή στον χρήστη να γράψει κώδικα και να το ενσωματώσει στο WEKA για τις δικές του ανάγκες κ την δικιά του χρήση.

Όταν κατεβάσουμε κ εγκαταστήσουμε το πρόγραμμα υπάρχει περίπτωση που θα ζητηθεί να γίνει και κάποια αναβάθμιση στο λογισμικό για την χρήση της java.Το περιβάλλον της αρχικής οθόνης του WEKA φαίνεται παρακάτω όπου θα εξηγηθούν αρχικά οι επιλογές που μας παρέχονται και αργότερα πιο λεπτομερειακά και με παραδείγματα.



1.3 Αρχικό Μενού WEKA

Υπάρχουν τέσσερεις επιλογές:

Explorer: Είναι το γραφικό περιβάλλον που χρησιμοποιούμε κυρίως για να επεξεργαστούμε τα δεδομένα τα οποία δεν έχουν υποστεί κάποια άλλη επεξεργασία.

Experimenter: Σε αυτό το περιβάλλον παίρνουν μέρος συνήθως διάφορες εργασίες που αφορούν στατιστική χρήση.

KnowledgeFlow: Η χρήση του είναι ακριβώς η ίδια με του Explorer απλά με δύο σημαντικές διαφορές. Για αρχή σε αυτή την επιλογή έχουμε την δυνατότητα να κάνουμε drag & drop δηλαδή να μεταφέρουμε δεδομένα από άλλα αρχεία στο αρχείο που χρησιμοποιούμε με μια κίνηση του ποντικιού. Δεύτερη μεγάλη διαφορά είναι ότι η συγκεκριμένη επιλογή υποστηρίζει την incremental learning(είναι μια μέθοδος εκμάθησης που συνδέει διάφορες τεχνικές) κ έτσι μπορούμε να αντλήσουμε γνώσεις από προηγούμενες εργασίες μας που είχαν παρόμοια μορφή.

Simple CLI:Παρέχει στους χρήστες που δεν έχουν την δυνατότητα να χρησιμοποιήσουν το πρόγραμμα με το γραφικό περιβάλλον του, τη δυνατότητα χρήσης μέσα από την γραμμή εντολών.
Το περιβάλλον της τελευταίας επιλογής έχει συγκεκριμένες εντολές που παρέχουν τα ανάλογα αποτελέσματα. Για παράδειγμα η εντολή :java weka.classifiers.trees.J48-t temp.arff

Θα ανοίξει το αρχείο μας temp.arff και θα εφαρμόσει τον αλγόριθμο J48 το οποίο αφορά δέντρα αποφάσεων. Επίσης θα πρέπει να γνωρίζουμε ότι και για την αποθήκευση σε αυτήν την επιλογή θα πρέπει να την κάνουμε πάλι με τον ίδιο τρόπο δηλαδή java weka.classifiers.TYPE.CLASSIFIER_NAME -t PATH/temp.arff -d PATH/temp_J48.model

1.4 Υποστηριζόμενα αρχεία WEKA

Το WEKA υποστηρίζει αρχεία της μορφής .arff(Attribute-Relation File Format) το οποίο είναι ένα αρχείο κειμένου ASCII το οποίο περιγράφει μια λίστα με δεδομένα που έχουν κοινά χαρακτηριστικά. Τα αρχεία .arff αποτελούνται από δύο μέρη:

Κύριο μέρος: Σε αυτό το κομμάτι περιέχονται

Όνομα αρχείου και ο συγγραφέας του αρχείου(δεν είναι απαραίτητο),καθώς και διάφορες πληροφορίες. Πριν από αυτά τα στοιχεία αναγράφουμε το % το οποίο συμβολίζει ότι η πληροφορία που εμφανίζεται μετά είναι κάποια σημείωση και δεν επηρεάζει το πρόγραμμα.

Υπάρχει επίσης το @RELATION το οποίο συσχετίζει ένα όνομα με το δεδομένα μας.

Και τελευταίο χαρακτηριστικό του κύριου μέρους είναι τα @ATTRIBUTES δηλαδή τα χαρακτηριστικά του αρχείου μας. Εδώ δηλώνουμε επίσης και το τι μορφές δεδομένων μπορούν να αποθηκευτούν στην κάθε μεταβλητή μας(numeric,string,date, <nominal-specification>

Δεδομένα: Τα δεδομένα μας έχουν στην αρχή τους το σύμβολο @DATA και ακριβώς μετά από αυτή την δήλωση αρχίζουμε και γράφουμε τις τιμές ανάλογα με την μεταβλητή μας.

Ένα παράδειγμα αρχείου .arff είναι το εξής:

%1. Title :Temperature of summer

%

% 2.Source

(a) www.meteo.gr

(b)www.poseidon.com

(c) 31/8/2014

@RELATION Weather

@ATTRIBUTE temp numeric

@ATTRIBUTE wind {N,W,E,S,NW,NE,SW,SE}

@ATTRIBUTE city {Athens,Patra,Thessaloniki}

@ATTRIBUTE rain{yes,no}

@DATA

13,N,Athens,yes

34,NE,Patra,no

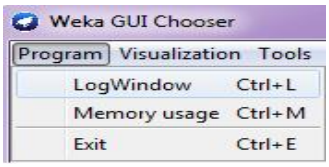
23,SW,Athens,no

39,SE,Thessaloniko,yes

Πρέπει να αναφέρουμε ότι οι ετικέτες @DATA,@ATTRIBUTE,@RELATION πρέπει να αναγράφονται πάντα με κεφαλαία γράμματα.

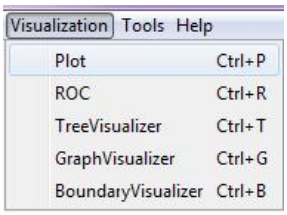
1.5 Μενού WEKA:

Στο ***Progra***m υπάρχουν οι επιλογές:



· LogWIndow :Το οποίο μας ανοίγει ένα νέο παράθυρο που καταγράφει διάφορες διαδικάσίες.
· Memory usage:Η μνήμη που απασχολείται από το πρόγραμμα μας
· Exit:Κλείσιμο του προγράμματος μας.

Στο ***Visualization*** υπάρχουν οι επιλογές:



· Plot: Για σχεδιασμό δυσδιάστατων μοντέλων από το αρχείο μας.
· ROC: Επιλέγουμε την πιο πρόσφατη αποθηκευμένη καμπύλη ROC.
· TreeVisualizer: Για απεικόνιση γραφημάτων , πχ Δέντρα αποφάσεων.
· GraphVisualizer: Απεικονίζει XML BIF ή DOT γραφήματα για δίκτυα Bayesian.
· BoundaryVisualizer: Μας επιτρέπει την απεικόνιση της ταξινόμησης των δεδομένων μας ανάμεσα στα όρια που έχουμε θέσει.

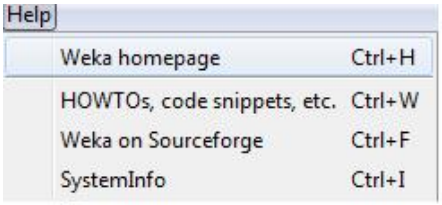
Στο ***Tools*** βρίσκουμε τις επιλογές:

ArffViewer: Αυτή η επιλογή μας επιτρέπει να βλέπουμε τα αρχεία .arff σε μορφή υπολογιστικού φύλλου.

SqlViewer: Αναπαριστά ένα φύλλο εργασίας sql για επεξεργασία μέσω JDBC(Java Database Connectivity, Συνδετικότητα Βάσης Δεδομένων JAVA)

Bayes net editor : Μας επιτρέπει την επεξεργασία, απεικόνιση, και εκμάθηση από τα δίκτυα Bayes.

Στο ***Help*** βρίσκουμε:



· Weka homepage: Μας μεταφέρει στο σάιτ του προγράμματος μας.
· HOWTOs, code snippets,etc: Είναι ένας οδηγός για το Weka που περιλαμβάνει πολλά παραδείγματα, καθώς και το πώς λειτουργεί η κάθε επιλογή του προγράμματος.
· Weka on Sourceforge: Η ιστοσελίδα του Weka στο Sourcforge
· SystemInfo: Διάφορες πληροφορίες που αφορούν το πρόγραμμα μας και τον υπολογιστή μας.

2.1 Simple CLI

Το Simple CLI μας παρέχει όλες τις δυνατότητες του Weka όπως κατηγοριοποίηση, συσταδοποίηση, φίλτρα κτλ.

2.1.1 Εντολές Simple CLI:

· Java<classname>[<args>]: Καλεί μια κλάση με τις επιμέρους παραμέτρους της(αν υπάρχουν)
· Break: Σταματάει οποιαδήποτε εργασία κάνουμε.Για παράδειγμα μια κατηγοριοποίηση.
· Kill: Τερματίζει απότομα οποιαδήποτε εργασία.
· Cls: Καθαρίζει την οθόνη εξόδου μας.
· Exit: Έξοδος από το περιβάλλον Simple CLI
· Help[<command>]: Μας παρέχει γενικές πληροφορίες για της εντολές αν δεν έχουμε βάλει κάποια παράμετρο ,αλλιώς αν έχουμε βάλει μας δείχνει πληροφορίες για την συγκεκριμένη εντολή

Το αρχείο που θα χρησιμοποιήσουμε περιλαμβάνει μια μελέτη που έγινε για τον διαβήτη.Τα αποτελέσματα έχουν μετρηθεί σε γυναίκες ασθενείς ηλικίας από 21 και πάνω.

Για παράδειγμα αν θέλουμε να εμφανίσουμε το δέντρο αποφάσεων για το αρχείο μας θα πρέπει να γράψουμε : java weka.classifiers.trees.J48 -t [την τοποθεσία και το όνομα του αρχείου μας] C:/temp/diabetes.arff.Και θα έχουμε σαν αποτέλεσμα:

```
J48 pruned tree
------------------

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.561: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.561
|   |   |   |   |   preg <= 6
|   |   |   |   |   |   age <= 30: tested_positive (4.0)
|   |   |   |   |   |   age > 30
|   |   |   |   |   |   |   age <= 34: tested_negative (7.0/1.0)
|   |   |   |   |   |   |   age > 34
|   |   |   |   |   |   |   |   mass <= 33.1: tested_positive (6.0)
|   |   |   |   |   |   |   |   mass > 33.1: tested_negative (4.0/1.0)
|   |   |   |   |   preg > 6: tested_positive (13.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pres <= 82
|   |   |   |   |   |   |   pedi <= 0.396: tested_positive (8.0/1.0)
|   |   |   |   |   |   |   pedi > 0.396: tested_negative (3.0)
|   |   |   |   |   |   pres > 82: tested_negative (4.0)
|   |   |   |   age > 61: tested_negative (4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pres <= 61: tested_positive (15.0/1.0)
|   |   |   pres > 61
|   |   |   |   age <= 30: tested_negative (40.0/13.0)
|   |   |   |   age > 30: tested_positive (60.0/17.0)
|   |   plas > 157: tested_positive (92.0/12.0)

Number of Leaves  :     20

Size of the tree :      39
=== Error on training data ===

Correctly Classified Instances         646               84.1146 %
Incorrectly Classified Instances       122               15.8854 %
Kappa statistic                          0.6319
Mean absolute error                      0.2383
Root mean squared error                  0.3452
Relative absolute error                 52.4339 %
Root relative squared error             72.4207 %
Total Number of Instances              768


=== Confusion Matrix ===

   a   b   <-- classified as
 468  32 |   a = tested_negative
  90 178 |   b = tested_positive


=== Stratified cross-validation ===

Correctly Classified Instances         567               73.8281 %
Incorrectly Classified Instances       201               26.1719 %
Kappa statistic                          0.4164
Mean absolute error                      0.3158
Root mean squared error                  0.4463
Relative absolute error                 69.4841 %
Root relative squared error             93.6293 %
Total Number of Instances              768


=== Confusion Matrix ===

   a   b   <-- classified as
 407  93 |   a = tested_negative
 108 160 |   b = tested positive
```

Στην αριστερή εικόνα βλέπουμε τις περιπτώσεις

Α)Η περιεκτικότητα της γλουτώζης να είναι πάνω από 127

Β) Η περιεκτικότητα της γλουτώζης να είναι κάτω από 127

Και ανάλογα η κάθε περίπτωση μετά χωρίζεται ανάλογα με την μάζα κ ύστερα με την ηλικία. Και καταλήγουμε στα πόσα άτομα βρέθηκαν θετικά στο τεστ και πόσα όχι.

Αυτό που μας ενδιαφέρει περισσότερο είναι η δεξιά εικόνα που αναφέρει το:Correctly Classified Instances.Από εδώ μπορούμε να έχουμε μια πρώτη εκτίμηση για το αν το μοντέλο μας είναι ακριβές. Μπορούμε να θεωρήσουμε με το ποσοστό να αγγίζει το 73 % ότι είναι ακριβές.

Στο κάτω μέρος της δεξιάς φωτογραφίας μπορούμε να δούμε ένα πινακάκι που ονομάζεται Confusion Matrix.Το 93 συμβολίζει το «false positives» και το 160 το «false negative». Τι σημαίνουν όμως αυτά.

False Positive: Είναι ένα λάθος το οποίο βγάζει σαν αποτέλεσμα ότι μια συγκεκριμένη κατάσταση υπάρχει στο πρόβλημα μας ενώ στην πραγματικότητα δεν υπάρχει.

Παράδειγμα: Όταν κάποιος φωνάξει “σεισμός” και αρχίσουν οι άνθρωποι να τρέχουν με τον έλεγχο να γίνεται αν όντος έγινε σεισμός, θα είχε σαν αποτέλεσμα ότι δεν έγινε. Ό λόγος που ο κόσμος όμως έτρεξε είναι επειδή κάποιος φώναξε ότι γίνεται.

False Negative: Είναι ένα λάθος το οποίο γίνεται σε ένα πρόβλημα μας όταν υποθέτει ότι μια συγκεκριμένη κατάσταση δεν υπάρχει ενώ στην πραγματικότητα υπάρχει κ επηρεάζει την λύση του. Παράδειγμα: Στην περίπτωση που αναρωτιόμαστε αν ένας κρατούμενος είναι ένοχος και το πρόγραμμα υπέθεσε ότι δεν είναι ενώ στην πραγματικότητα ήταν.

Ο λόγος που υπάρχουν αυτά τα δύο στον έλεγχο μας είναι :

Το False Positive παρατηρείται όταν ελέγχουμε μια κατάσταση και τα αποτελέσματα έχουν την μορφή “true”or “false”

Το False Negative παρατηρείται όταν ελέγχεται μια κατάσταση και τα αποτελέσματα μπορεί να είναι «ναι» ή «όχι».

2.1.2 Διάφοροι Παράμετροι του Simple CLI:

-t	Προσδιορίζει το αρχείο που θα επεξεργαστούμε training set(αρχείο που το χρησιμοποιούμε και για το machine learning) (ARFF format)
-T	Προσδιορίζει το αρχείο που θα επεξεργαστούμε test set(αρχείο το οποίο κάνουμε διάφορα τέστ) (ARFF format). Αν μας λείπει κάποια παράμετρος τότε θα κάνουμε διασταυρωμένη επικύρωση (Cross-validation)(χωρίζεται σε σε 10 μέρη)
-x	Αυτή η παράμετρος καθορίζει τα μέρη που θα γίνει το cv.
-c	Αυτή η παράμετρος θέτει την μεταβλητή κλάσης με την μέθοδο one-based index.
-d	Το μοντέλο μας μετά το training αποθηκεύεται με αυτή την παράμετρο. Αποθηκεύεται το μοντέλο που βρίσκεται στο training και όχι όλα τα μοντέλα που κάνουν το cv.
-l	Καλεί ένα προηγούμενο αρχείο .
-p <attrib_range>	Αν έχουμε επιλέξει ένα συγκεκριμένο αρχείο ,αυτή η παράμετρος μας δείχνει την πρόβλεψη και μία μεταβλητή για όλα τα τεστ που έχουν γίνει μέσω του cv. Αν δεν έχουμε επιλέξει κάποιο αρχείο τότε αυτή η παράμετρος δεν βγάζει αποτέλεσμα και θα πρέπει να χρησιμοποιήσουμε το callClassifier .
-i	Μια λεπτομερείς περιγραφή βάση ακρίβειας παρέχεται από αυτή την παράμετρο. Αυτά τα αποτελέσματα μπορούμε να τα βρούμε στο

	confusion matrix
-o	Με αυτή την παράμετρο αφαιρούμε την ικανότητα παρουσίασης με μορφή που ο χρήστης μπορεί να διαβάσει.

Ας δούμε και άλλα παραδείγματα κατηγοριοποίησης:
Αλγόριθμος SVM support vector machines
java weka.classifiers.functions.SMO –t [<path>][<file name>]

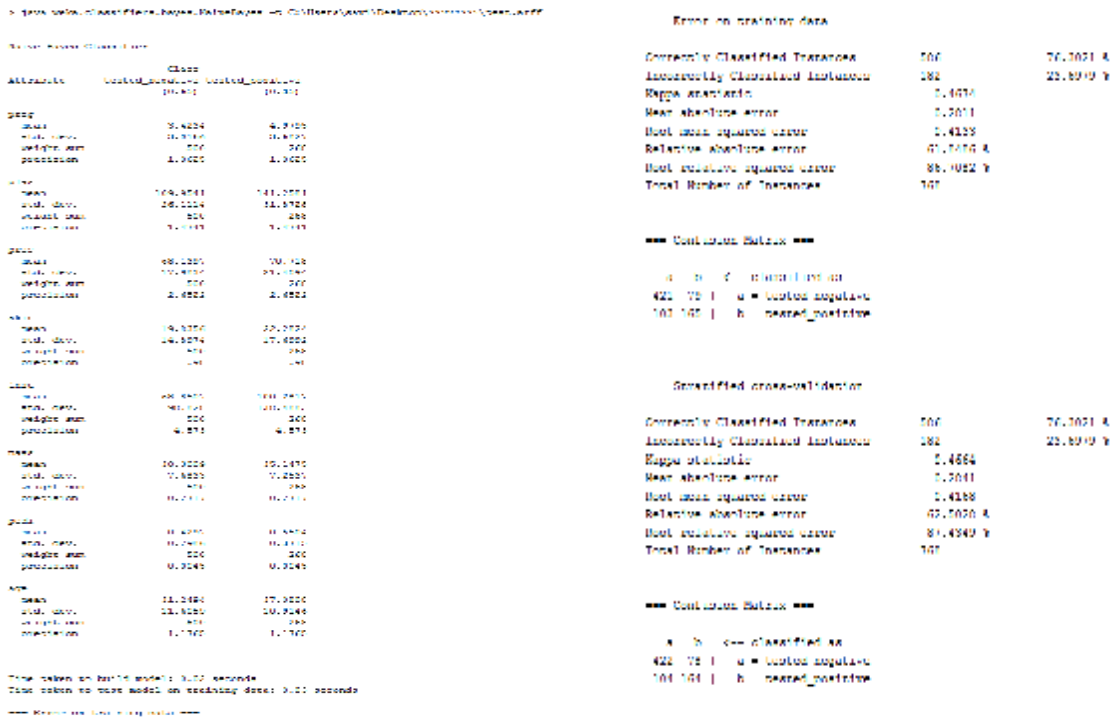
```
> java weka.classifiers.functions.SMO -t C:\Users\savi\Desktop\????????\test.arff

SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: tested_negative, tested_positive

BinarySMO

Machine linear: showing attribute weights, not support vectors.

          1.3614 * (normalized) preg
 +        4.8764 * (normalized) plas
 +       -0.8118 * (normalized) pres
 +       -0.1158 * (normalized) skin
 +       -0.1776 * (normalized) insu
 +        3.0745 * (normalized) mass
 +        1.4242 * (normalized) pedi
 +        0.2601 * (normalized) age
 -        5.1761

Number of kernel evaluations: 19131 (69.279% cached)


Time taken to build model: 0.07 seconds
Time taken to test model on training data: 0.01 seconds

=== Error on training data ===

Correctly Classified Instances         595               77.474  %
Incorrectly Classified Instances       173               22.526  %
Kappa statistic                          0.4688
Mean absolute error                      0.2253
Root mean squared error                  0.4746
Relative absolute error                 49.5632 %
Root relative squared error             99.5752 %
Total Number of Instances              768

=== Confusion Matrix ===

   a   b   <-- classified as
 452  48 |   a = tested_negative
 125 143 |   b = tested_positive


=== Stratified cross-validation ===

Correctly Classified Instances         594               77.3438 %
Incorrectly Classified Instances       174               22.6563 %
Kappa statistic                          0.4682
Mean absolute error                      0.2266
Root mean squared error                  0.476
Relative absolute error                 49.848  %
Root relative squared error             99.862  %
Total Number of Instances              768


=== Confusion Matrix ===

   a   b   <-- classified as
 449  51 |   a = tested_negative
 123 145 |   b = tested_positive
```

Η εξήγηση είναι ακριβώς η ίδια με την προηγούμενη κατηγοριοποίηση. Αλλά μπορούμε να παρατηρήσουμε ότι το μοντέλο μας με αυτή είναι πιο ακριβές.
java weka.classifiers.functions.VotedPerceptron –t [<path>][<file name>]

```
> java weka.classifiers.functions.VotedPerceptron -t C:\Users\savi\Desktop\????????\test.arff

VotedPerceptron: Number of perceptrons=330

Time taken to build model: 0.03 seconds
Time taken to test model on training data: 0.02 seconds

=== Error on training data ===

Correctly Classified Instances         520               67.7083 %
Incorrectly Classified Instances       248               32.2917 %
Kappa statistic                          0.1751
Mean absolute error                      0.3237
Root mean squared error                  0.5684
Relative absolute error                 71.2228 %
Root relative squared error            119.2491 %
Total Number of Instances              768

=== Confusion Matrix ===

   a   b   <-- classified as
 456  44 |   a = tested_negative
 204  64 |   b = tested_positive


=== Stratified cross-validation ===

Correctly Classified Instances         513               66.7969 %
Incorrectly Classified Instances       255               33.2031 %
Kappa statistic                          0.1353
Mean absolute error                      0.3319
Root mean squared error                  0.5752
Relative absolute error                 73.0209 %
Root relative squared error            120.6751 %
Total Number of Instances              768

=== Confusion Matrix ===

   a   b   <-- classified as
 462  38 |   a = tested_negative
 217  51 |   b = tested_positive
```

Εδώ χρησιμοποιούμε κατηγοριοποίηση με Νευρωνικά δίκτυα. Παρατηρούμε ότι το μοντέλο μας δεν είναι τόσο ακριβές με αυτό όσο με τα δύο προηγούμενα.

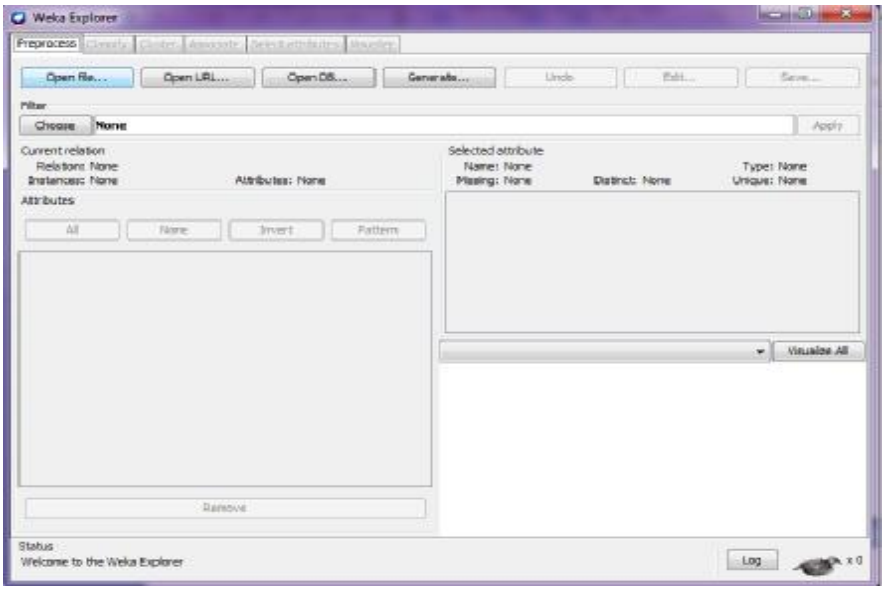
Και τέλος java weka.classifiers.bayes.NaiveBayes –t (Κατηγοριοποίηση σύμφωνα με το θεώρημα του Bayes.)



Στην μέθοδο Bayes παρατηρούμε σύμφωνα με την αριστερή φωτογραφία ότι κάνει έλεγχο και σε κάθε μια μεταβλητή μας ώστε με το machine learning να μπορεί να ανακαλύψει παρόμοια πρότυπα μελλοντικά.

3.1 Explorer

Στην αρχική σελίδα του explorer παρατηρούμε τις εξής επιλογές:



Preprocess: Αυτή η επιλογή μας επιτρέπει να επιλέξουμε το αρχείο που θα χρησιμοποιήσουμε στο πρόγραμμα μας.

Classify: Σε αυτή την επιλογή παίρνουμε το αρχείο που έχουμε επιλέξει κ κάνουμε διάφορα τεστ.

Cluster: Εδώ εφαρμόζουμε διάφορες μεθόδους που μας επιτρέπουν αν ανακαλύψουμε αν υπάρχουν διάφορα πρότυπα.

Associate: Εδώ εφαρμόζουμε διάφορους κανόνες που μας επιτρέπουν να δούμε αν τα δεδομένα μας σχετίζονται μεταξύ τους.

Select attributes: Εδώ εφαρμόζοντας διάφορους κανόνες και αφού επιλέξουμε / απεπιλέξουμε κάποιες παραμέτρους ανακαλύπτουμε διάφορες αλλαγές(αν υπάρχουν) στο μοντέλο μας.

Visualize: Σε αυτήν την επιλογή έχουμε την δυνατότητα να αναπαραστήσουμε το μοντέλο μας σε δυο διαστάσεις σε διάφορα γραφήματα.

3.1.1 Preprocess:

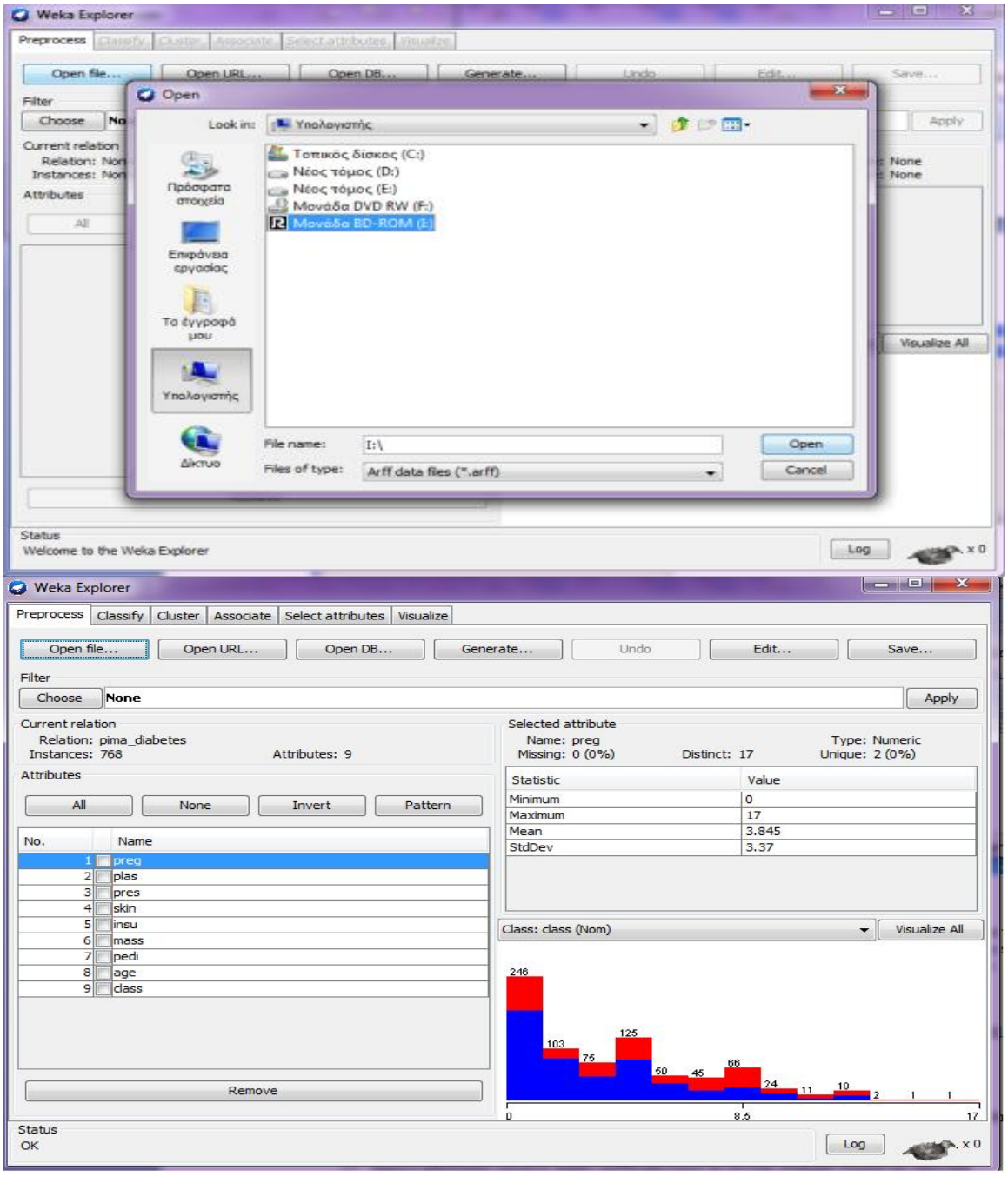
Υπάρχουν οι επιλογές :

Open file: Επιλέγουμε το αρχείο που θα χρησιμοποιήσουμε από μια τοποθεσία στον υπολογιστή μας.

Open Url : Επιλέγουμε ένα αρχείο το οποίο είναι αποθηκευμένο σε διαφορετική τοποθεσία από του υπολογιστή μας.

Open DB: Εδώ επιλέγουμε μια βάση δεδομένων από την οποία θα αντλήσουμε τα δεδομένα μας.

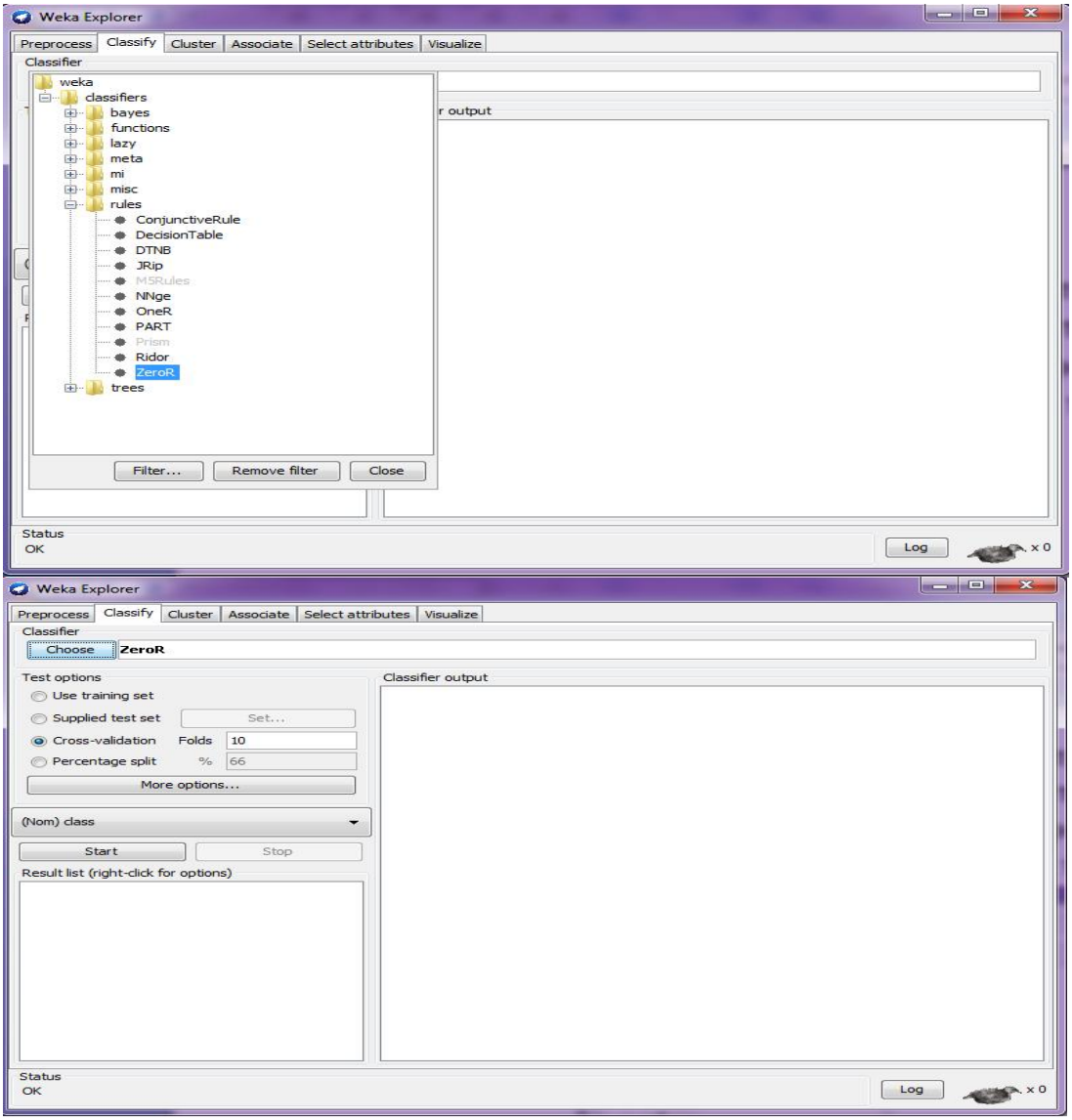
Τα αρχεία σε παλιές εκδόσεις του WEKA θα πρέπει να έχουν την μορφή .arff όπως έχουμε αναφέρει, αλλά σε νέες εκδόσεις υποστηρίζονται αρχεία κ των μορφών csv, c4.5 με τις καταλήξεις .csv,.bsi,.names,.data.



Με το που επιλέξουμε το αρχείο μας ,εμφανίζεται στην οθόνη μας η αριστερή εικόνα. Εδώ μπορούμε να επιλέξουμε τις διάφορες μεταβλητές μας και να τους εφαρμόσουμε κάποια φίλτρα και να επεξεργαστούμε το αρχείο μας.

3.1.2 Classify :

Ο χρήστης έχει την δυνατότητα να εφαρμόσει στα δεδομένα του διάφορους αλγόριθμους για να συγκεντρώσει διάφορες πληροφορίες. Ο καλύτερος τρόπος για να γίνει αυτό είναι να εφαρμόσει ξεχωριστά τις διάφορες επιλογές μέχρι να βρει αυτή που του παρέχει τα καλύτερα αποτελέσματα.



Εδώ έχουμε άλλες τέσσερεις επιλογές που αφορούν το αρχείο μας:

¨ Use training set : Χρησιμοποιούμε το αρχείο που έχουμε για το machine learning

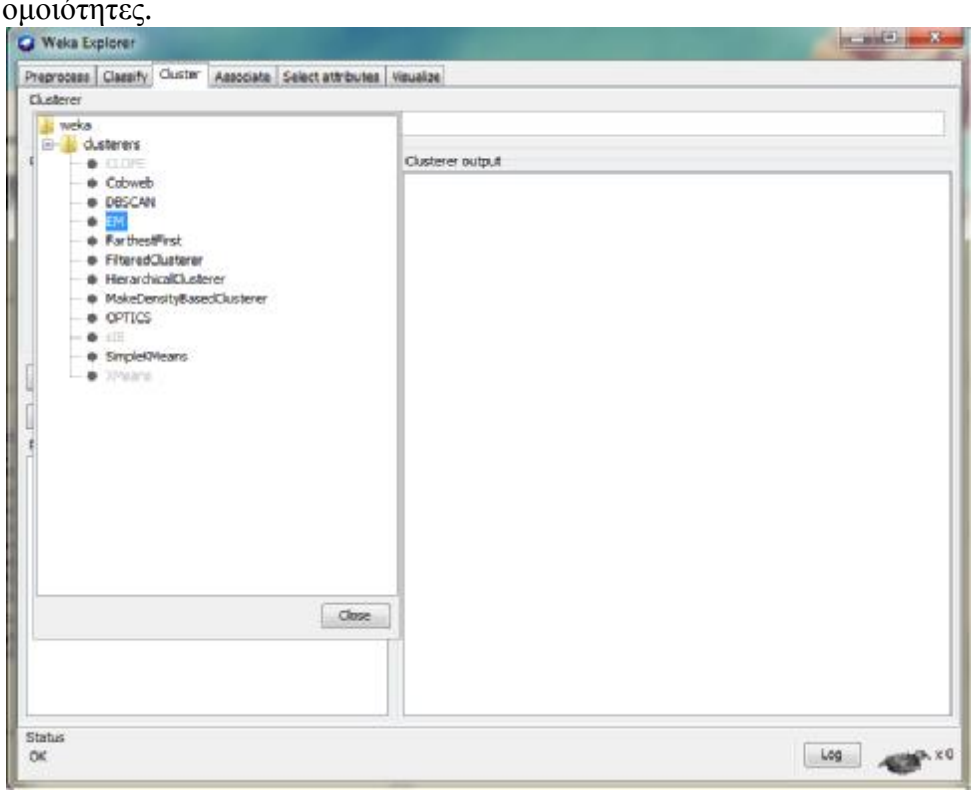
¨ Supplied test set: Χρησιμοποιούμε το αρχείο που έχουμε για να εξάγουμε διάφορα αποτελέσματα.

¨ Cross-validation: Πραγματοποιούμε διασταυρωμένη επικύρωση ανάμεσα στα δύο αρχεία μας (training –test set)

¨ Percentage split: Το χρησιμοποιούμε όταν δεν έχουμε δύο αρχεία(test – training)και θέλουμε να χωρίσουμε τα δεδομένα μας για τους δύο αυτούς σκοπούς. Επιλέγουμε το ποσοστό που θέλουμε να έχει το καθένα και επιλέγουμε το “Start”.

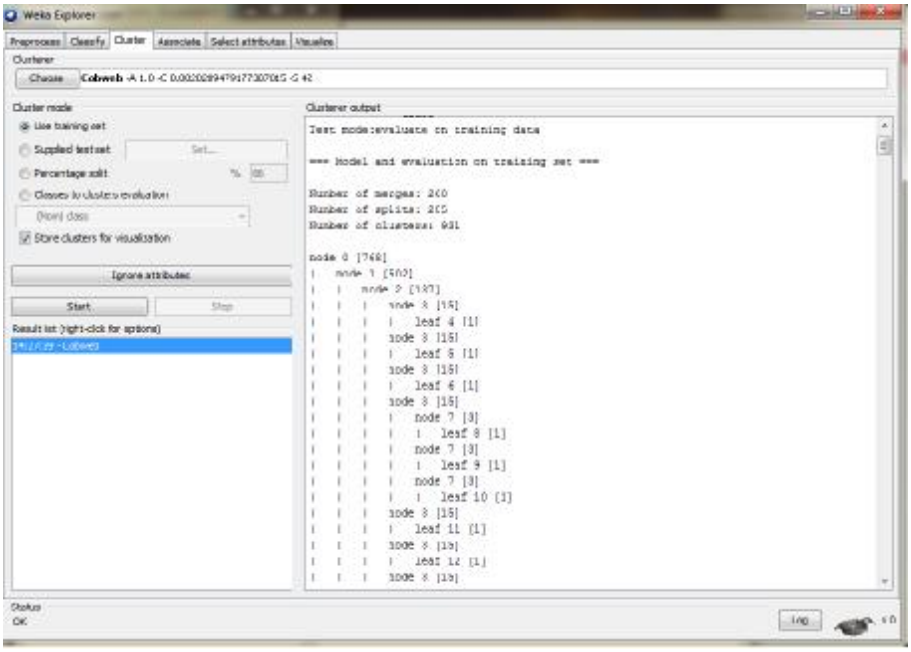
3.1.3 CLUSTER:

Η ανάλυση σε ομάδες (clustering ή cluster analysis) είναι μια μέθοδος που ο σκοπός της είναι η κατάταξη των παρατηρήσεων μας σε ομάδες χρησιμοποιώντας τις πληροφορίες που υπάρχουν. Με απλά λόγια ελέγχει πόσο όμοιες είναι οι παρατηρήσεις μας σύμφωνα με κάποιον αριθμό μεταβλητών. Έπειτα δημιουργεί ομάδες από τις παρατηρήσεις που έχουν ομοιότητες.



Εδώ βλέπουμε τις διάφορες προσεγγίσεις που έχουμε, με πιο ευρέως χρησιμοποιούμενες την SimpleKMeans. Ο κάθε αλγόριθμος έχει τα πλεονεκτήματα του και τα μειονεκτήματα, ας αναλύσουμε τους ποιους βασικούς αλγόριθμους.

Αλγόριθμος Cobweb: Ο αλγόριθμος αυτός αναπτύχθηκε από ερευνητές για την χρήση σε αντικειμενοστραφή αρχεία δεδομένων. Δημιουργεί ένα δενδρόγραμμα το οποίο χαρακτηρίζει την κάθε συστάδα με πιθανολογική περιγραφή. Χρησιμοποιεί ιεραρχικές μεθόδους όπου οι ομάδες περιγράφονται πιθανοκρατικά.

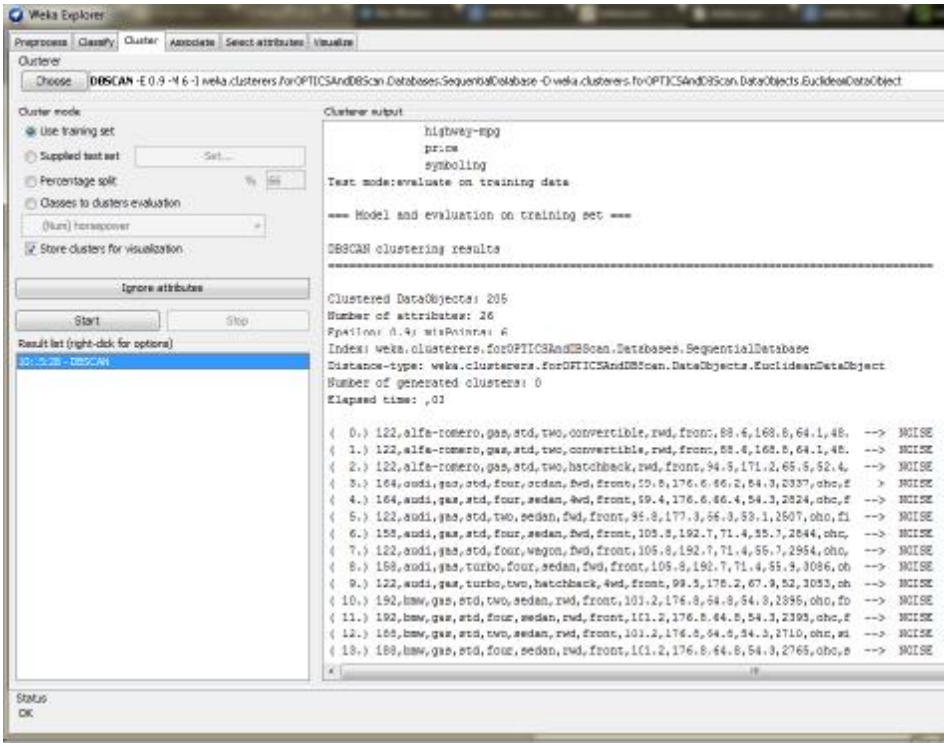


Πλεονεκτήματα του αλγορίθμου Cobweb:

· Ο αλγόριθμος χρησιμοποιεί μια ιεραρχική μέθοδο αξιολόγησης που βοηθάει την δημιουργία του δέντρου.
· Ενσωματώνει σταδιακά τα αντικείμενα μας στο δέντρο ταξινόμησης με σκοπό την βελτιστοποίηση του. Εδώ εμφανίζεται η μεγάλη διαφορά με τον αλγόριθμο K-means καθώς στον αλγόριθμο Cobweb μπορούμε να δημιουργήσουμε άμεσα νέα κλάση.
· Μια ακόμα λειτουργία του είναι η διάσπαση και ένωση των ομάδων μας προσφέροντας αμφίδρομη αναζήτηση καθώς μια ένωση μπορεί να «ακυρώσει» μια προηγούμενη διάσπαση.
Όμως ο Cobweb έχει αρκετά μειονεκτήματα

· Συγκεκριμένα βασίζεται στην υπόθεση ότι η κατανομή πιθανοτήτων σε ξεχωριστές μεταβλητές είναι στατιστικά ανεξάρτητες.
· Άλλο μεγάλο πρόβλημα του είναι ότι το δέντρο αποφάσεων δεν βγάζει σωστά αποτελέσματα όταν υπάρχει η πιθανότητα τα δεδομένα μας να είναι διαστρεβλωμένα.

Αλγόριθμος DBSCAN: Χρησιμοποιείτε πιο πολύ σε αρχεία δεδομένων που υπάρχει «θόρυβος». Βασίζεται στην πυκνότητα των ομάδων(clusters) αφού έχει την ιδιότητα να βρίσκει διάφορες ομάδες αρχίζοντας από την εκτιμώμενη κατανομή πυκνότητας των αντίστοιχων κόμβων.



Πλεονεκτήματα του αλγορίθμου DBSCAN**:**

· Σε αντίθεση με το k-means δεν χρειάζεται να γνωρίζουμε το σύνολο των ομάδων που υπάρχουν στο δείγμα μας.

· Επίσης μπορεί να ανακαλύψει τις αυθαίρετα διαμορφωμένες συστάδες

· Όπως αναφέραμε είναι ιδανικός για δείγμα με θόρυβο.

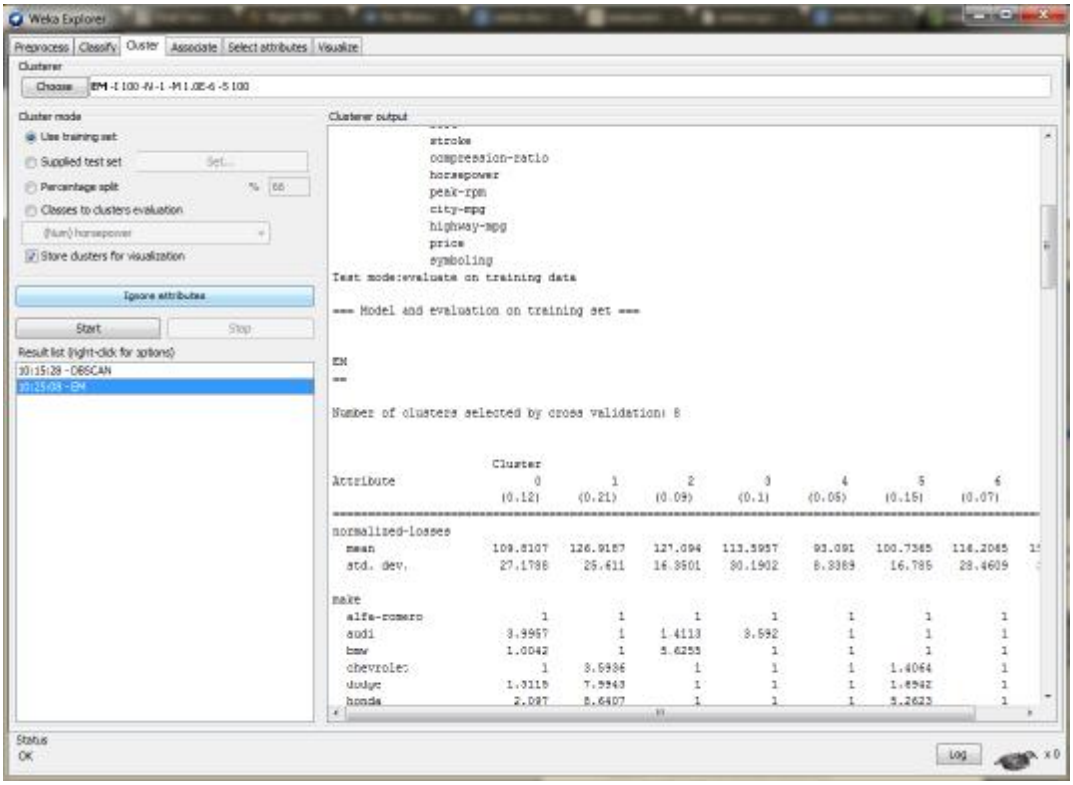
· Και τέλος χρειάζεται μόνο δύο παραμέτρους.

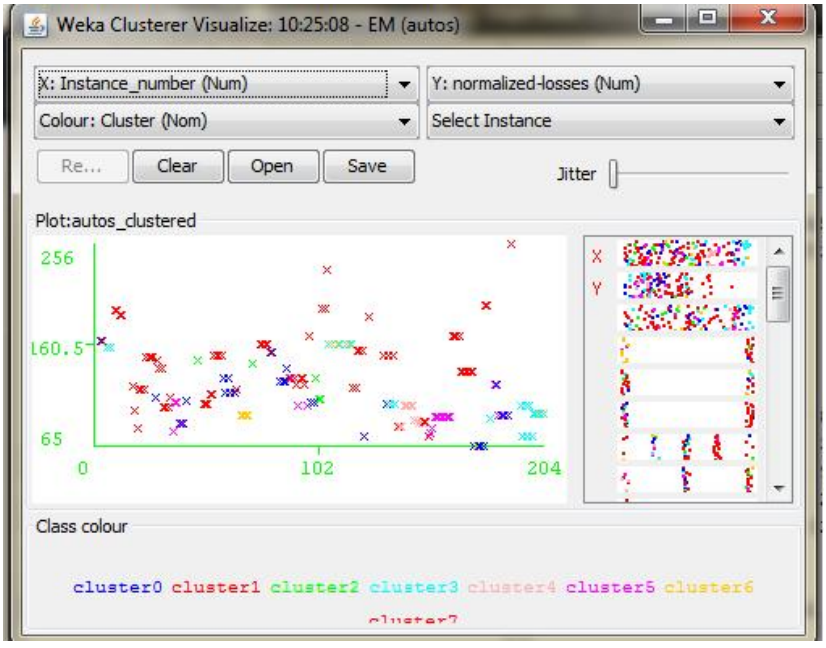
Μειονεκτήματα DBSCAN:

· Ο αλγόριθμος δεν μπορεί να ομαδοποιήσει δεδομένα τα οποία έχουν μεγάλη διαφορά στην πυκνότητα τους λόγω της μη δυνατότητας επιλογής του σωστού συνδυασμού minipts-e

· Η ποιότητα του DBSCAN εξαρτάται από την απόσταση που έχει χρησιμοποιηθεί στην συνάρτηση regionQuery(P,ε)

Αλγόριθμος EM:





Ο Αλγόριθμος ΕΜ χρησιμοποιείται όταν τα αποτελέσματα από την μέθοδο k-means δεν μας ικανοποιούν και θέλουμε να βρούμε το μέγιστο a posteriori .Δηλαδή ελέγχει το κάθε στιγμιότυπο του δείγματος μας με μια πιθανότητα που συμβολίζει την πιθανότητα να ανήκει στην κάθε συστάδα.

Πλεονεκτήματα του αλγορίθμου ΕΜ:

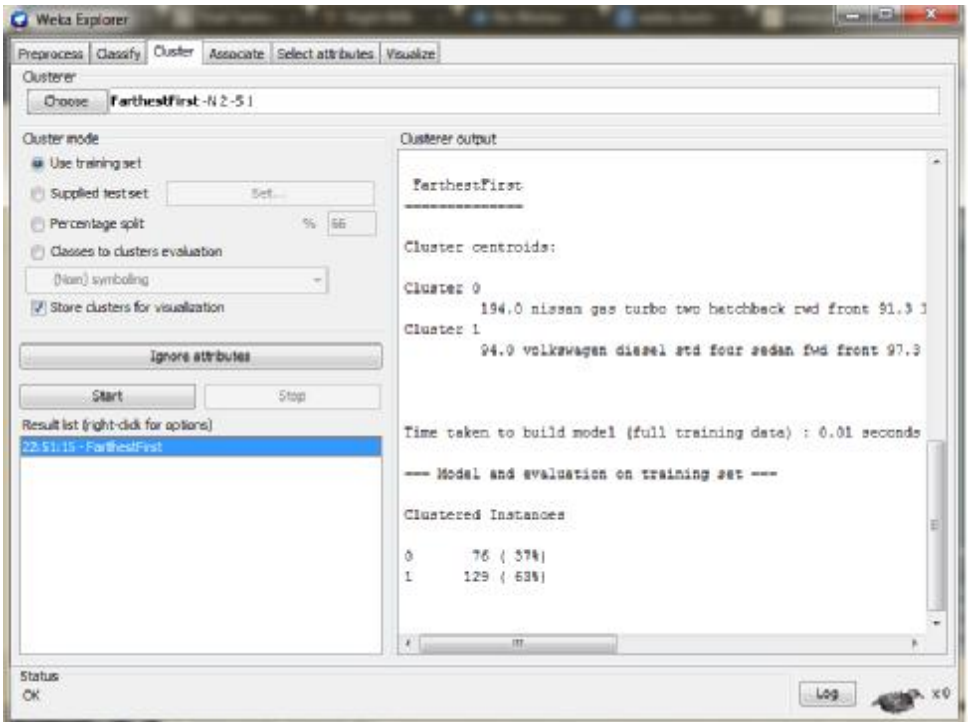
· Μας δίνει ακριβή και χρήσιμα αποτελέσματα.

· Έχουμε την δυνατότητα να κάνουμε ομαδοποίηση σε ένα μικρότερο μέρος του δείγματος.

Μειονεκτήματα του αλγορίθμου EM:
· Πολυπλοκότητα.

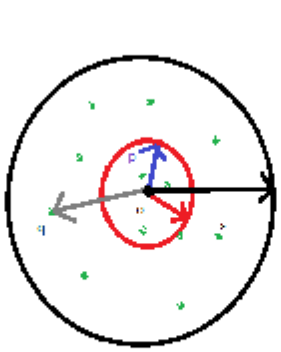
Αλγόριθμος Farthest First:

Ο αλγόριθμος Farthest First είναι μια παραλλαγή του K-means ο οποίος τοποθετεί το κέντρο της κάθε συστάδας στην πιο μακρινή απόσταση από τους άλλους κάθε φορά. Το σημείο που τοποθετείτε πρέπει να βρίσκεται μέσα στο εύρος των δεδομένων μας. Αυτό μας βοηθάει γιατί είναι πιο γρήγορος ο υπολογισμός της ομαδοποίησης διότι υπολογίζονται λιγότερες μεταβλητές. Ο αλγόριθμος χωρίζει το αρχείο μας σε δύο συστάδες και είναι κατάλληλος για πολύ μεγάλα αρχεία εξόρυξης.

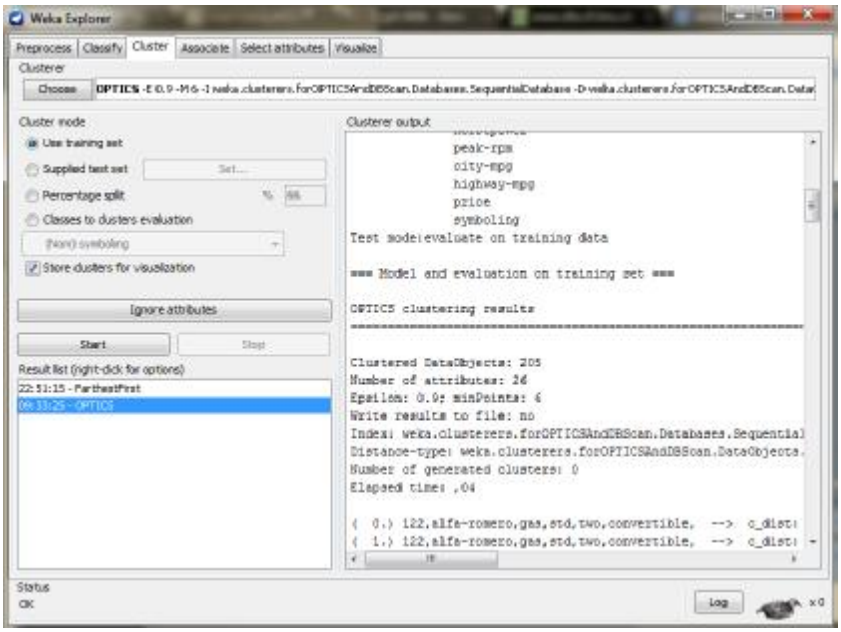


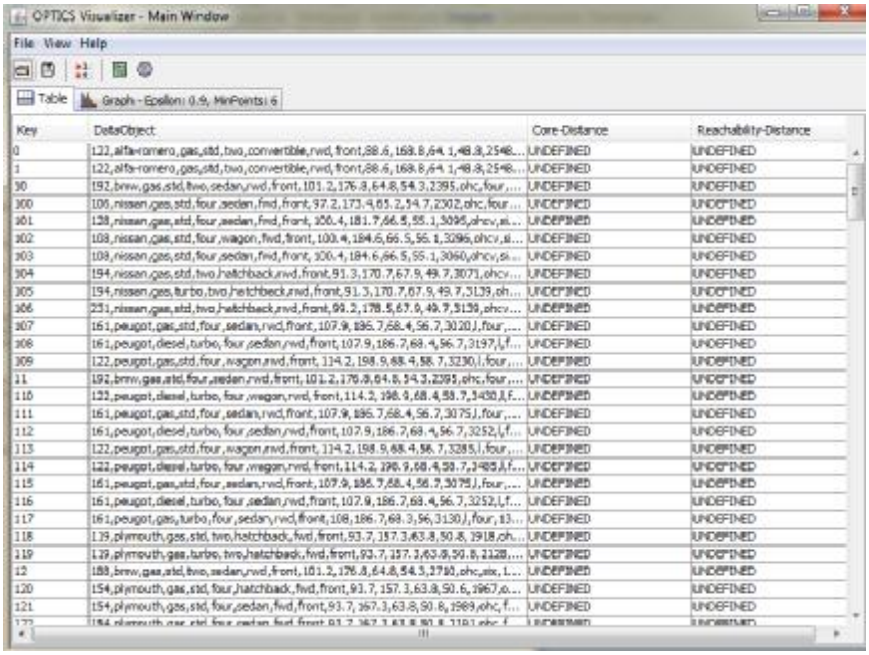
Ο αλγόριθμος Optics:
Ο αλγόριθμος Optics είναι διαδικαστικά ίδιος με τον προαναφερθέντα DBSCAN.
Ο αλγόριθμος πραγματοποιείτε από ένα άπειρο αριθμό i όπου 0<=i<=ε. Αντίθετα δεν δημιουργεί συστάδες αλλά τα αποτελέσματα του χωρίζονται σε δύο ομάδες ,core distance και reachability –distance.
Core distance από ένα σημείο ο είναι η ελάχιστη απόσταση i ώστε το σημείο ο είναι core point.
Reachability –distance από ένα σημείο p είναι η μικρότερη απόσταση η οποία το σημείο ο είναι core point και το p ανήκει στο ε –εύρος του.

Στο παρακάτω σχήμα θα εξηγήσουμε τα σημεία p,q,o:

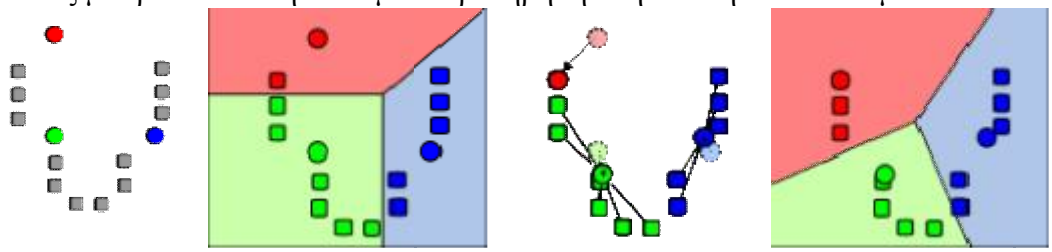


Στο σχήμα μας παρατηρούμε ότι το core –distance ενός σημείου ο είναι η απόσταση στο πιο μακρινό MinPts.Δηλαδή η μικρότερη τιμή του ει όπου το σημείο ο είναι core point. Η Reachability –distance μεταξύ δύο σημείων μπορεί να μετρηθεί με διάφορους τρόπους για παράδειγμα αν ένα σημείο p βρίσκεται μέσα στην ει "γειτονιά" ενός σημείου ο η reachability –distance του σημείου p είναι ίση με την core –distance από το σημείο διότι η απόσταση ει είναι η μικρότερη δυνατή απόσταση για να θεωρηθεί το σημείο ο core point.Αλλιώς αν ένα άλλο σημείο q βρίσκεται εκτός του εύρους της ει-"γειτονιάς" ενός σημείου ο η reachability –distance είναι η απόσταση από ένα σημείο ο.





Ο αλγόριθμος K-means:
Ο αλγόριθμος K-means είναι μια μέθοδος που έχει σκοπό να χωρίσει n παρατηρήσεις σε k συστάδες με τρόπο όπου η κάθε μια παρατήρηση ανήκει στην συστάδα με τον κοντινό μέσο.



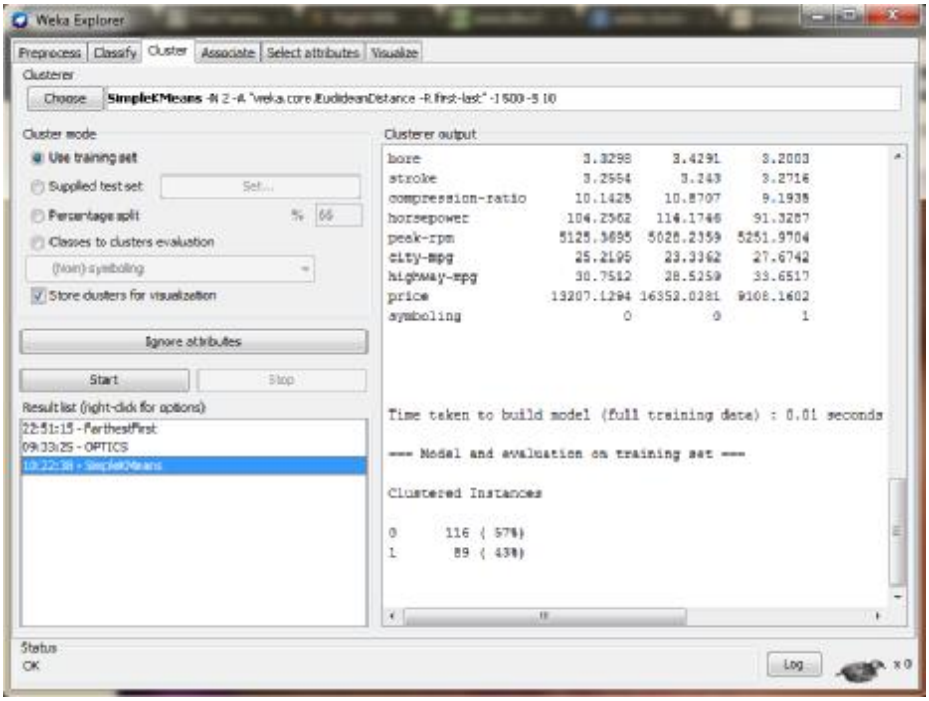
Στο πρώτο σχήμα βλέπουμε τους τρείς μέσους μας(κόκκινο, μπλε, πράσινο). Στο επόμενο σχήμα βλέπουμε τον αρχικό διαχωρισμό με τον πιο κοντινό μέσο για κάθε παρατήρηση. Στο τρίτο σχήμα είναι η συνέχεια της διεργασίας μας όπου ο μέσος μεταφέρεται στο κέντρο της κάθε συστάδας και επαναλαμβάνεται το βήμα δύο κ τρία μέχρι να φτάσουμε στη σύγκλιση που φαίνεται στο σχήμα τέσσερα.

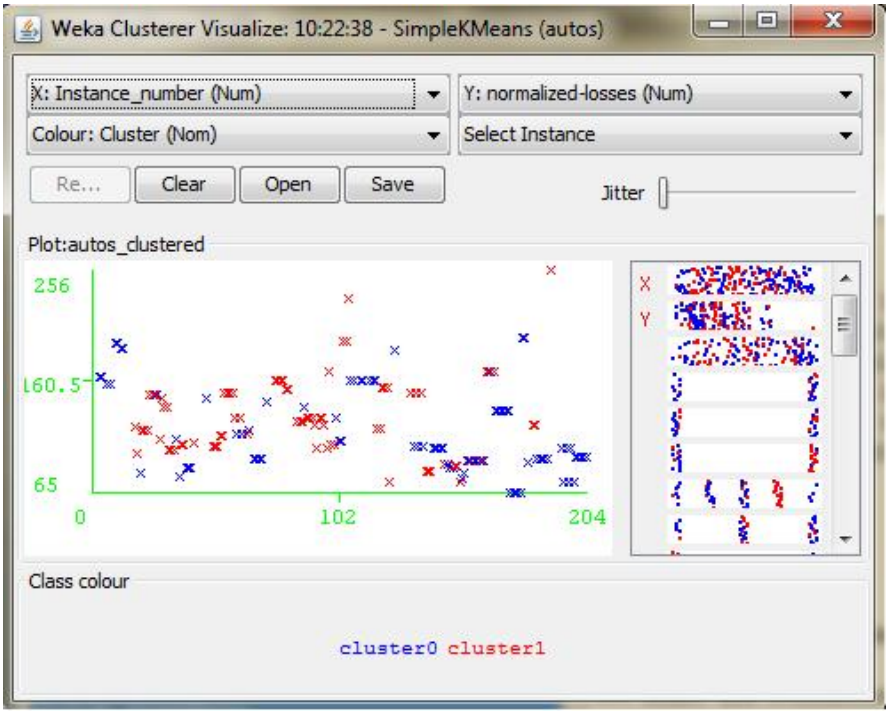
Πλεονεκτήματα του αλγορίθμου k-means:

· Σε μεγάλο όγκο μεταβλητών είναι πιο γρήγορος από την ιεραρχική ομαδοποίηση.
· Δημιουργεί πιο δομημένες συστάδες από την ιεραρχική.

Μειονεκτήματα του αλγορίθμου k-means:

· Δύσκολη σύγκριση της ακρίβειας των συστάδων
· Δεν βγάζει ακριβή αποτελέσματα όταν συνδυάζεται με μη σφαιρικές συστάδες



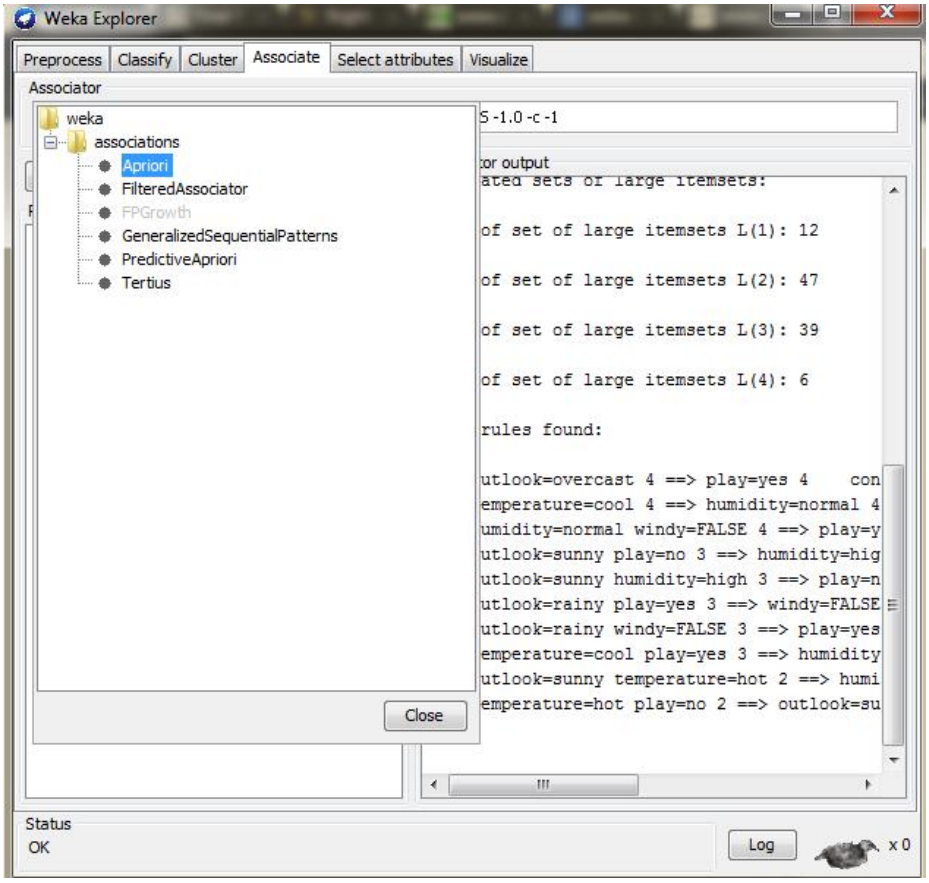


Ο κάθε αλγόριθμος έχει την δικιά του χρησιμότητα καθώς και τα δικά του πλεονεκτήματα μειονεκτήματα.

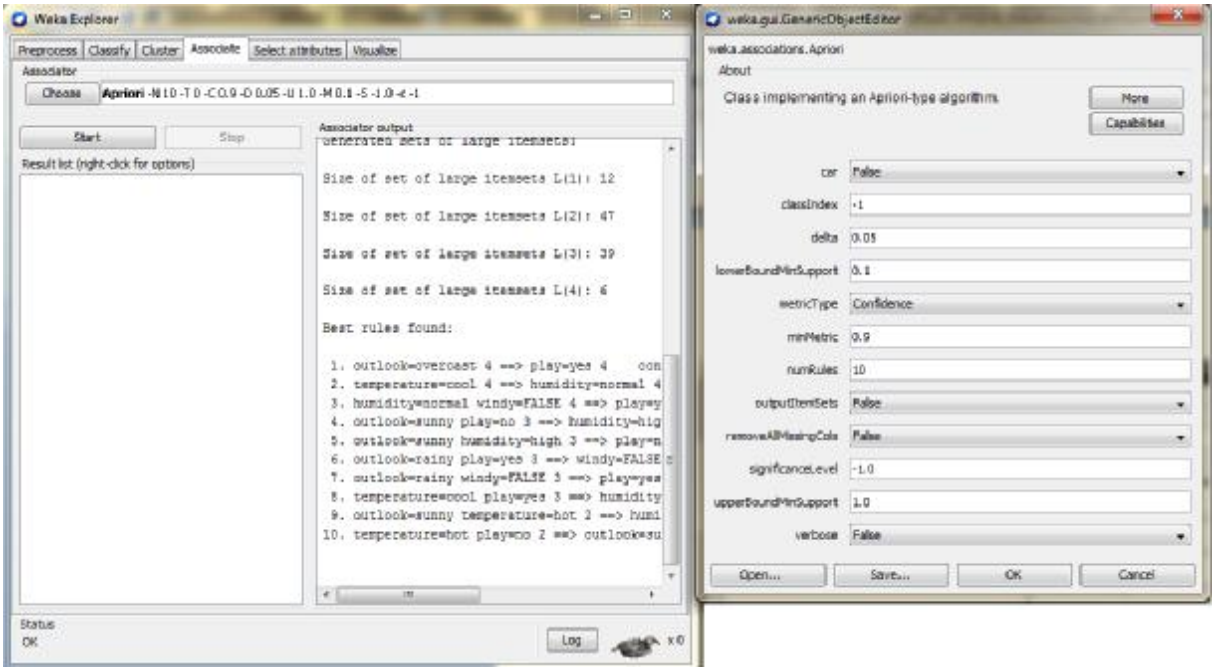
3.1.4 Associate

Τα μενού αυτό ψάχνει να βρεί σχέσεις ανάμεσα στα δεδομένα μας, αλλά οι μόνες τιμές που αναγνωρίζει είναι ονομαστικές και όχι αριθμητικές. Χρησιμοποιείται συνήθως με basket-market analysis.Ένα γνωστό παράδειγμα είναι οι πελάτες που αγοράζουν βούτυρο και ψωμί στο σουπερμάρκετ με confidence 0,2/0,2=1.0.Δηλαδή στο 100% των συναλλαγών που ο πελάτης αγόραζε βούτυρο και ψωμί αγοράζει και γάλα.

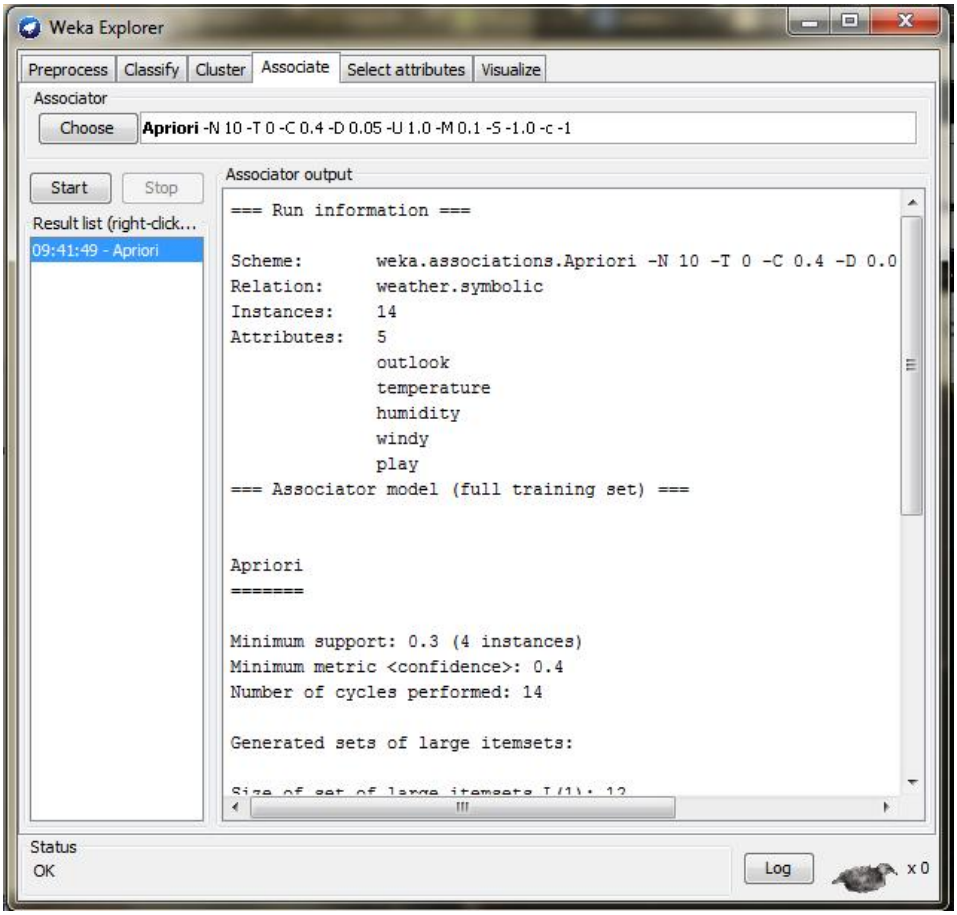
	A	B	C	D	E	F
1	Βάση δεδομένων από αγορές σε ΣουπερΜαρκετ.					
2		Γάλα	Ψωμί	Βούτυρο	Σοκολάτα	Αυγά
3	1	1	1	0	0	0
4	2	0	0	1	0	0
5	3	0	0	0	1	1
6	4	1	1	1	0	0
7	5	0	1	0	0	0



Εδώ βλέπουμε τις διάφορες επιλογές μας για τον πιο κανόνα θα επιλέξουμε με πιο γνωστό τον Apriori. Αφού επιλέξουμε τον κανόνα θα πρέπει να περάσουμε κάποιες παραμέτρους όπως φαίνεται και στην αποκάτω εικόνα. Για να ανοίξουμε την επιλογή των παραμέτρων κάνουμε κλικ στο όνομα του κανόνα. Πρέπει να αλλάξουμε κάποιες επιλογές όπως για παράδειγμα αν θέλουμε σαν μέτρο το confidence ή το lift ή το conviction. Στο minMetric βάζουμε την τιμή που παρουσιάζεται στο δείγμα μας σύμφωνα με τον κανόνα του confidence conf(X→Y)=supp(X U Y)/supp(X).To upperBoundMinSuppport πρέπει να το θέσουμε 1.0 και το lowerBoundMinSupport 0.1 γιατί ο Apriori ξεκινάει από το upperBound και μειώνεται κατά 5% (προεπιλεγμένη επιλογή) ή μέχρι να δημιουργηθούν όλοι οι κανόνες ή μέχρι να φτάσει το lowerBound.



Όταν τρέξουμε τον αλγόριθμο θα πάρουμε τα ακόλουθα :



Ας εξηγήσουμε λίγο τα παραπάνω :

· Το όνομα της σχέσης “weather”(relation name)

· αριθμός των περιπτώσεων στην συγκεκριμένη σχέση(number of instances in the relation)

· αριθμός των χαρακτηριστικών στην σχέση μας καθώς και η αναφορά τους(number of attributes in the relation)

3.1.5 Select attributes

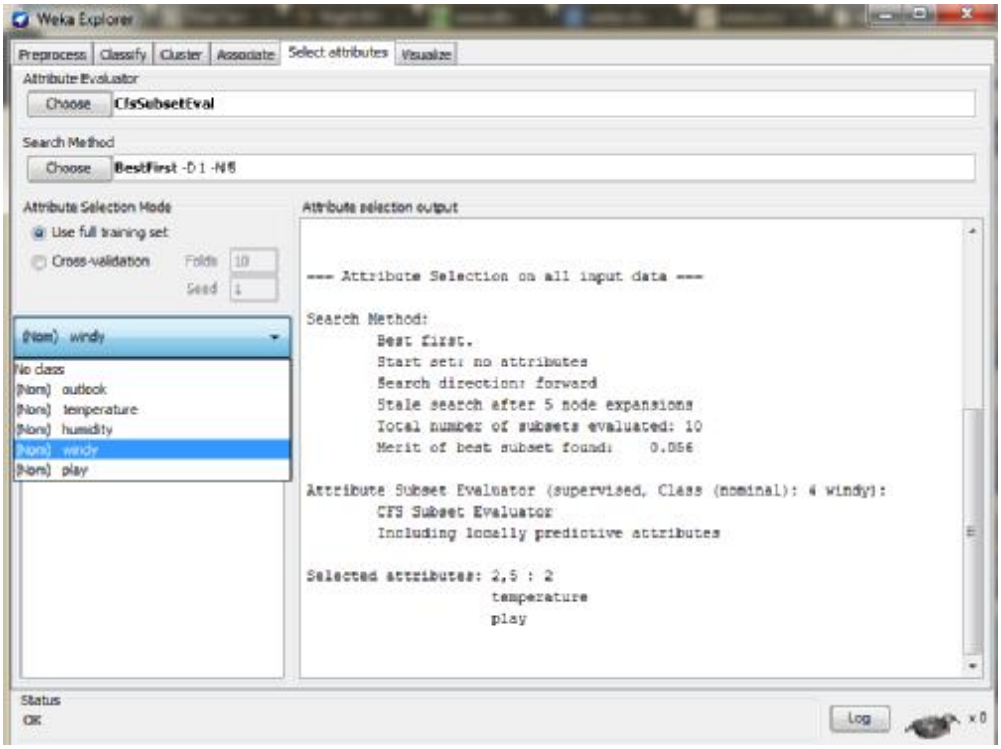
Εδώ μας δίνεται η επιλογή να ψάξουμε όλους τους δυνατούς συνδυασμούς των γνωρισμάτων μας στο αρχείο και να βρει πιο υποσύνολο είναι το καλύτερο για να κάνουμε κάποια πρόγνωση. Αυτό αποτελείται από δύο μέρη :

Πρώτο μέρος : Μέθοδος αναζήτησης
· Best-first
· Forward selection
· Random
· Exhaustive
· Genetic algorithm
· Ranking

Δεύτερο μέρος :Μέθοδος αξιολόγησης
· Correlation-based
· Wrapper
· Information gain
· Chi-squared

Μπορούμε να επιλέξουμε να χρησιμοποιήσουμε όλο το αρχείο ώστε η αξία του υποσυνόλου να βασίζεται σε αυτό(use full training set)ή να χρησιμοποιήσουμε την μέθοδο cross-validation.

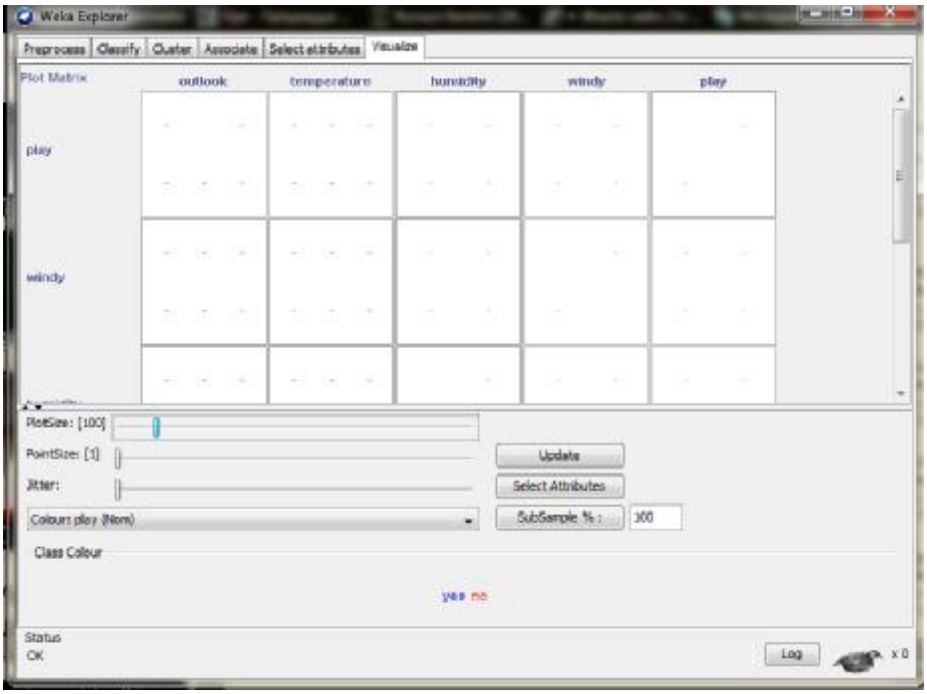
Και τέλος πρέπει να επιλέξουμε ποια θα είναι η κλάση μας.



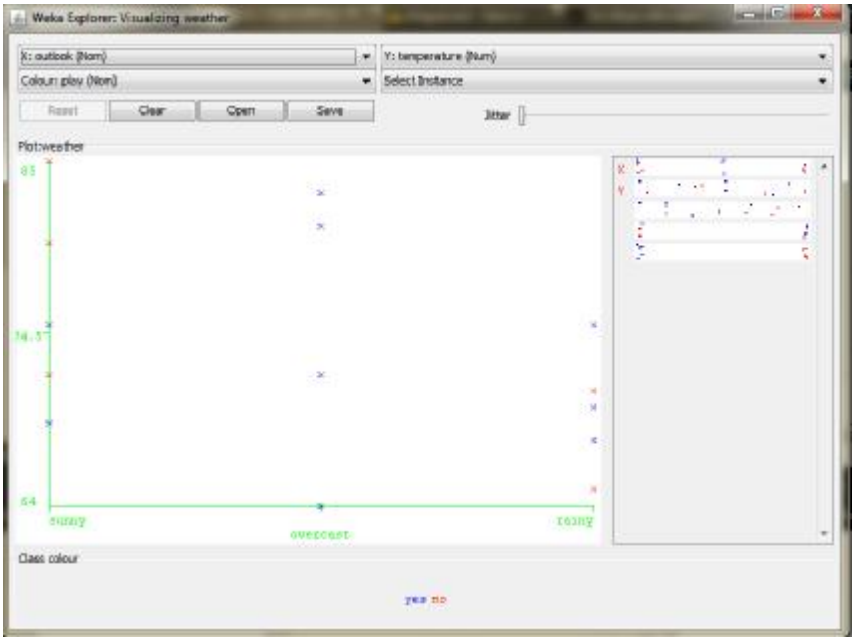
Σαν αποτέλεσμα μπορούμε να δούμε ότι έχουμε την δυνατότητα να επιλέξουμε δύο μεταβλητές για την πρόγνωση μας,την “temperature” και “play”.

3.1.6 Data Visualization

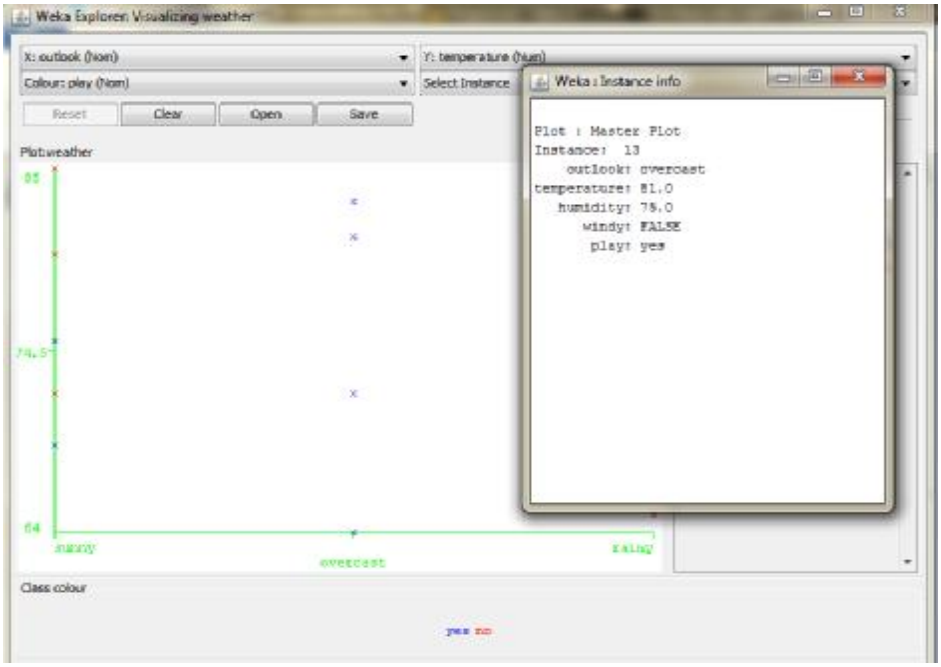
Το πρόγραμμα Weka μας παρέχει την δυνατότητα να απεικονίσουμε 2-D διαγράμματα καθώς έτσι μπορούμε να πειραματιστούμε και να μπορέσουμε να καταλάβουμε καλύτερα το πρόβλημα μας.Έτσι έχουμε την δυνατότητα απεικόνισης 1-d,2-d και να χρησιμοποιήσουμε το (Xgobi style) για 3-d μοντέλα.Για να χρησιμοποιήσουμε αυτό το μενού απλά πατάμε την καρτέλα visualize και μας εμφανίζεται η παρακάτω εικόνα.



Ας διαλέξουμε στον άξονα του X την μεταβλητή outlook και την temperature στον άξονα του Y. Και μετά μας εμφανίζεται το παράθυρο αυτό :



Στα αριστερά πάνω και δεξιά πάνω είναι οι επιλογές του άξονα Χ και Υ που μπορούν να τα αλλάξουμε. Ακριβώς από κάτω έχουμε την επιλογή να επιλέξουμε τον χρωματισμό των μεταβλητών που θέλουμε να δούμε. Στο κάτω μέρος αναφέρεται σε τι τιμές αναφέρεται το κάθε χρώμα στην εικόνα μας το μπλε σημαίνει yes και το κόκκινο no.Αν πατήσουμε πάνω σε αυτά θα εμφανιστεί ένα άλλο παράθυρο το οποίο μας δίνει την δυνατότητα να αλλάξουμε τα χρώματα. Στα δεξιά μας είναι κάποιες γραμμές .Η κάθε γραμμή αναπαριστά ένα χαρακτηριστικό και οι τελείες μέσα σε αυτές είναι οι τιμές που έχει το κάθε ένα. Επίσης έχουμε την δυνατότητα όταν κλικάρουμε κάποιο στοιχείο του διαγράμματος μας να δούμε πληροφορίες για αυτό.



4.1 EXPERIMENTER

Η Επιλογή του experimenter μας δίνει την δυνατότητα να τρέξουμε,να δημιουργήσουμε ,να αναλύσουμε τα πειράματα μας με ένα πιο βολικό τρόπο όταν επεξεργαζόμαστε τα σχέδια μας.Για παράδειγμα έχουμε την δυνατότητα να δημιουργήσουμε ένα πείραμα το οποίπ θα τρέχει διάφορα σχέδια σε μια σειρά από δεδομένα και να αναλύσει μέχρι να βρει πιο είναι στατιστικά καλύτερο από τα άλλα .

Έχουμε την επιλογή να τρέξουμε τον experimenter και από την επιλογή του Simple CLI.

Java weka.experiment.Experiment-r –T data/breastcanser.arff
-D weka.experiment.InstancesResultListener
-P weka.experiment.RandomSplitResultProducer—
-W weka.experiment.ClassifierSplitEvaluator—
-W weka.classifiers.rules.OneR

Οι προηγούμενες εντολές πρέπει να δωθούν σε μια γραμμή και το αποτέλεσμα που θα έχουμε είναι να τρέξει αλγόριθμο OneR στο αρχείο δεδομένων breastcanser.

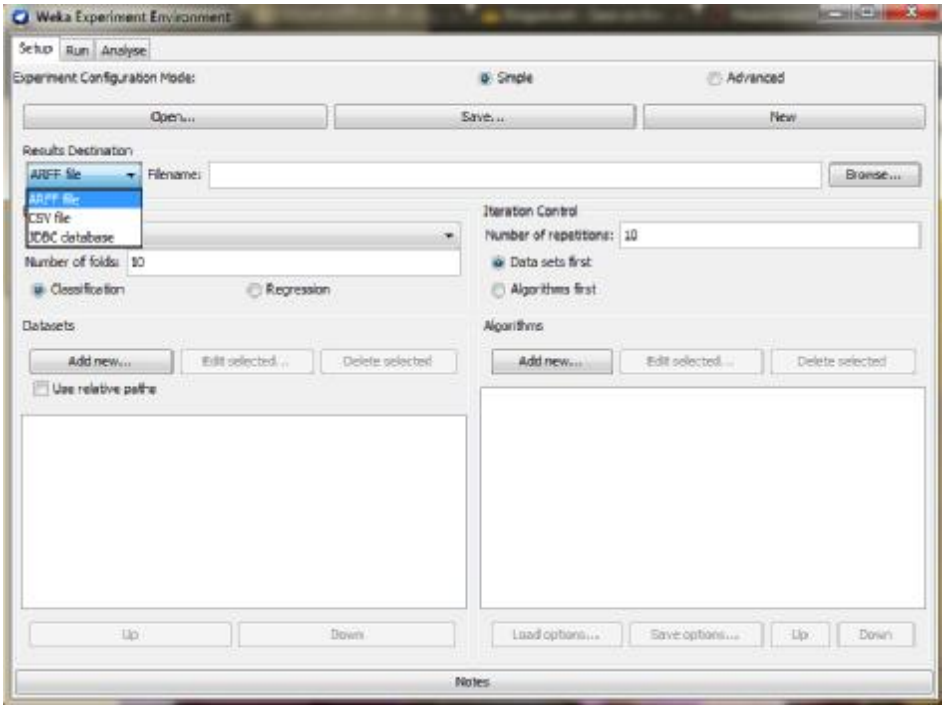
Δεν είναι όμως ότι πιο εύκολο να κάνουμε καθώς ένα τυπογραφικό λάθος μπορεί να μας οδηγήσεις σε λάθος αποτελέσματα.

Στον Experimenter έχουμε δύο επιλογές την Simple και την Advanced.

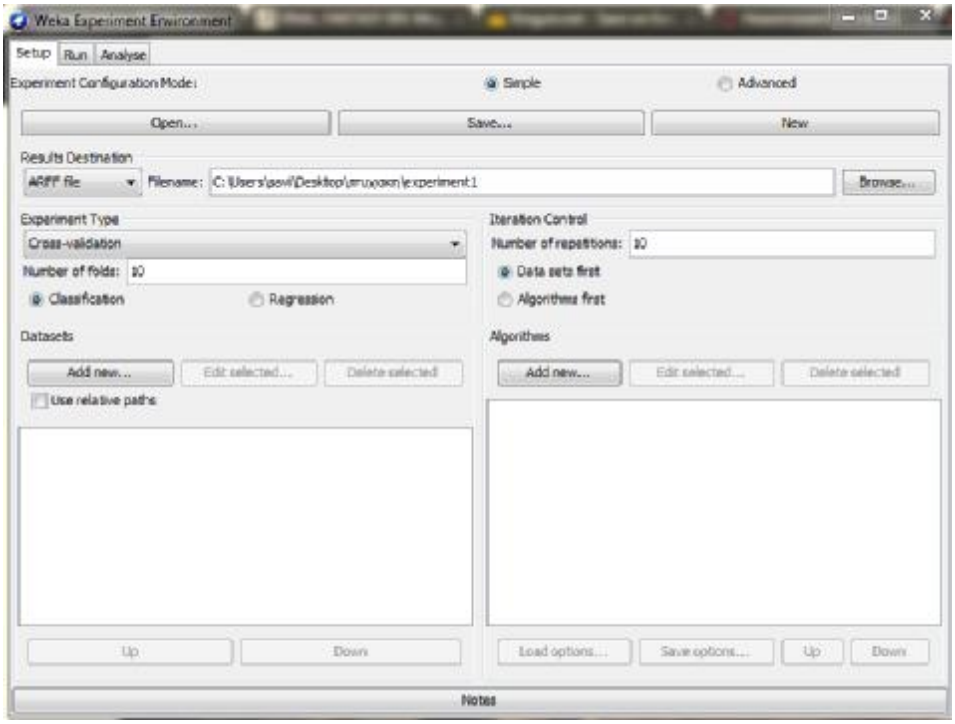
Πρώτα επιλέγουμε το αρχείο μας που θα χρησιμοποιηθεί.

4.1.2 Simple

4.1.2.1 New experiment



Εδώ φαίνεται η επιλογή που έχουμε για να αποθηκεύσουμε το πείραμα μας. Μπορούμε σε arff , csv αρχεία και jdbc database. Αν δεν επιλέξουμε ένα όνομα για το αρχείο μας τότε θα πάρει ένα τυχαίο και θα αποθηκευτεί στο φάκελο temp.Όταν θα αποθηκεύσουμε το νέο μας αρχείο το όνομα του θα εμφανιστεί δίπλα στην μορφή που επιλέξαμε.



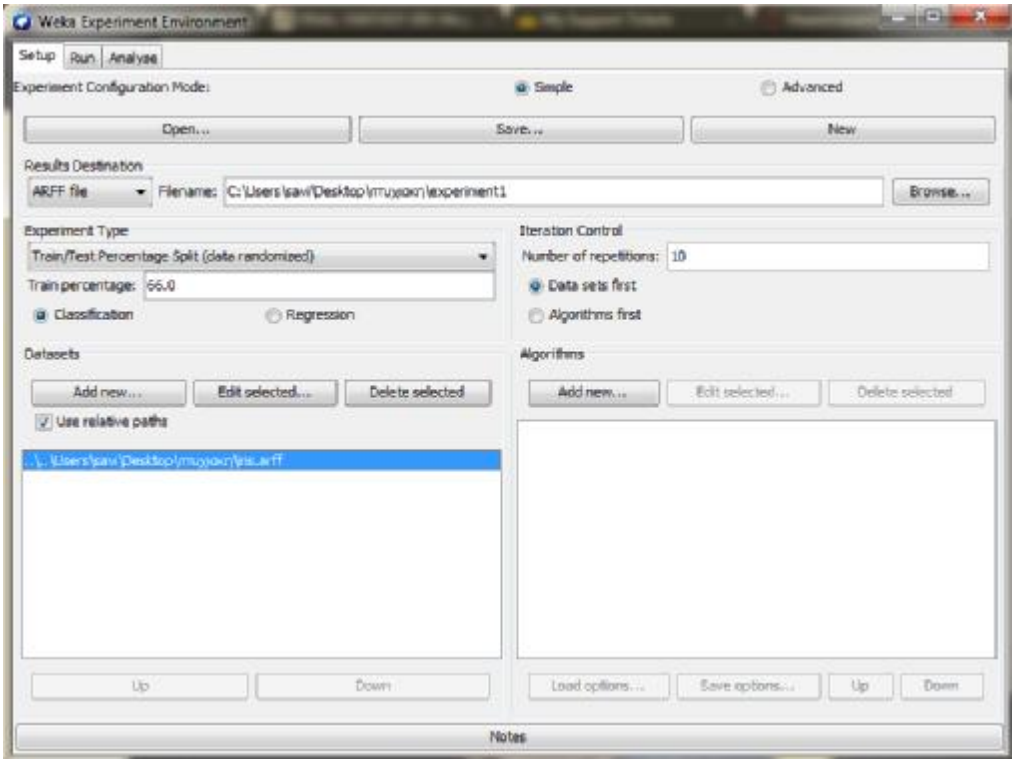
Τα αρχεία arff και csv έχουν σαν κύριο πλεονέκτημα ότι δεν χρειάζεται να δημιουργήσουμε έξτρα κλάσεις εκτός από αυτές που δημιουργήθηκαν από το πρόγραμμα μας. Σαν κύριο μειονέκτημα είναι ότι δεν έχουμε την δυνατότητα να συνεχίσουμε κάποιο πείραμα το οποίο διακόπηκε στην μέση. Ιδιαίτερα όταν κάποια πειράματα χρειάζονται αρκετό χρόνο για την επεξεργασία τους αυτό το μειονέκτημα είναι αρκετά ενοχλητικό. Ενώ το αρχείο σε jdbc σε περίπτωση διακοπής έχει αποθηκεύσει ότι διεργασία έχει γίνει κ απλά συνεχίζει από εκεί που

σταμάτησε. Ανάλογα λοιπόν το μέγεθος και την δυσκολία πρέπει να κάνουμε την ανάλογη επιλογή.

Αμέσως μετά έχουμε την επιλογή του Experiment type που είναι :

· Cross-validarion (προεπιλεγμένο): Κάνει διασταυρωμένο έλεγχο ανάλογα με τις φορές που θα επιλέξουμε.
· Train/Test Percentage Split: Διαχωρίζει ένα αρχείο ανάλογα με το ποσοστό που έχουμε επιλέξει σε train και test .
Ακόμα έχουμε την επιλογή να επιλέξουμε το αν θα χρησιμοποιήσουμε ομαδοποίηση (classification) ή παλινδρόμηση(regression) ανάλογα με τα αρχεία και τον αλγόριθμο που θα χρησιμοποιήσουμε.

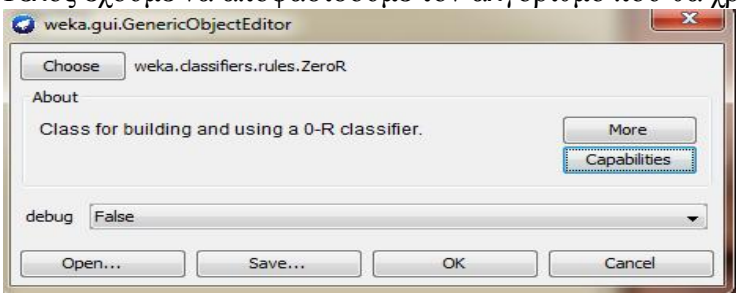
Αμέσως μετά πρέπει να επιλέξουμε το/α αρχείο/α το/α οποίο/α θα χρησιμοποιήσουμε για να τρέξουμε το πείραμα μας.



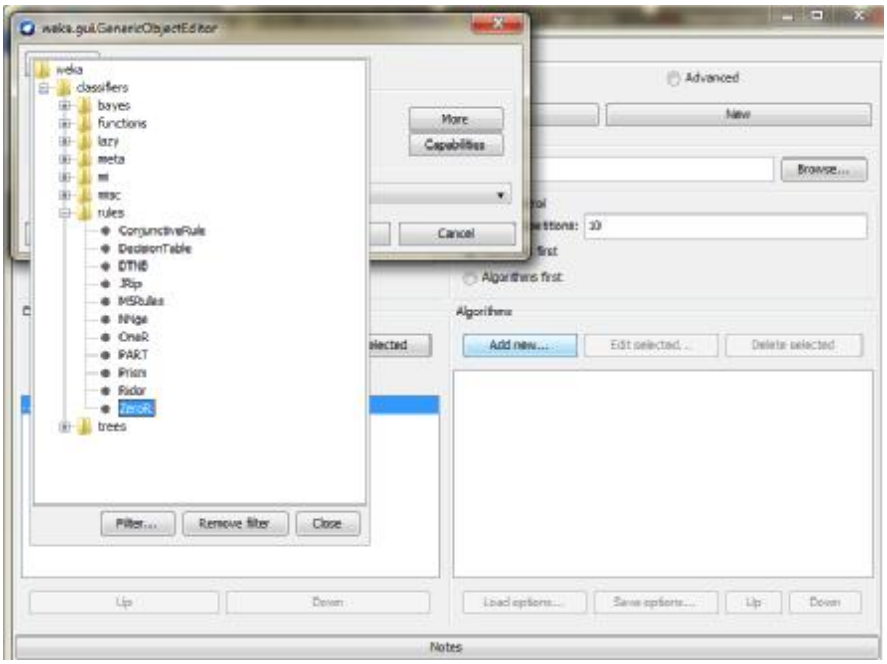
Αν δεν είμαστε σίγουροι αν το αρχείο που επιλέξαμε είναι το σωστό με ένα διπλό κλικ πάνω στο όνομα του αρχείου μπορούμε να δούμε το περιεχόμενο του.

Ύστερα έχουμε να επιλέξουμε τις επαναλήψεις που θα γίνουν (Iteration control).
· Αριθμός επαναλήψεων (number of repetition):Για να έχουμε στατιστικά ακριβή αποτελέσματα θα πρέπει να κάνουμε τουλάχιστον 10 επαναλήψεις.
· Αρχεία δεδομένων πρώτα/Αλγόριθμοι πρώτα (data sets first/algorithms fist):Αν χρησιμοποιούμε παραπάνω από ένα αλγόριθμο και αρχείο έχουμε την επιλογή για το ποιο από τα δύο θα κάνει πρώτα τις επαναλήψεις του. Αυτό μας βοηθάει αν αποθηκεύουμε σε βάση δεδομένων και θέλουμε να συγκρίνουμε κάποια αποτελέσματα όσο το δυνατόν πιο γρήγορα.

Τέλος έχουμε να αποφασίσουμε τον αλγόριθμο που θα χρησιμοποιηθεί:



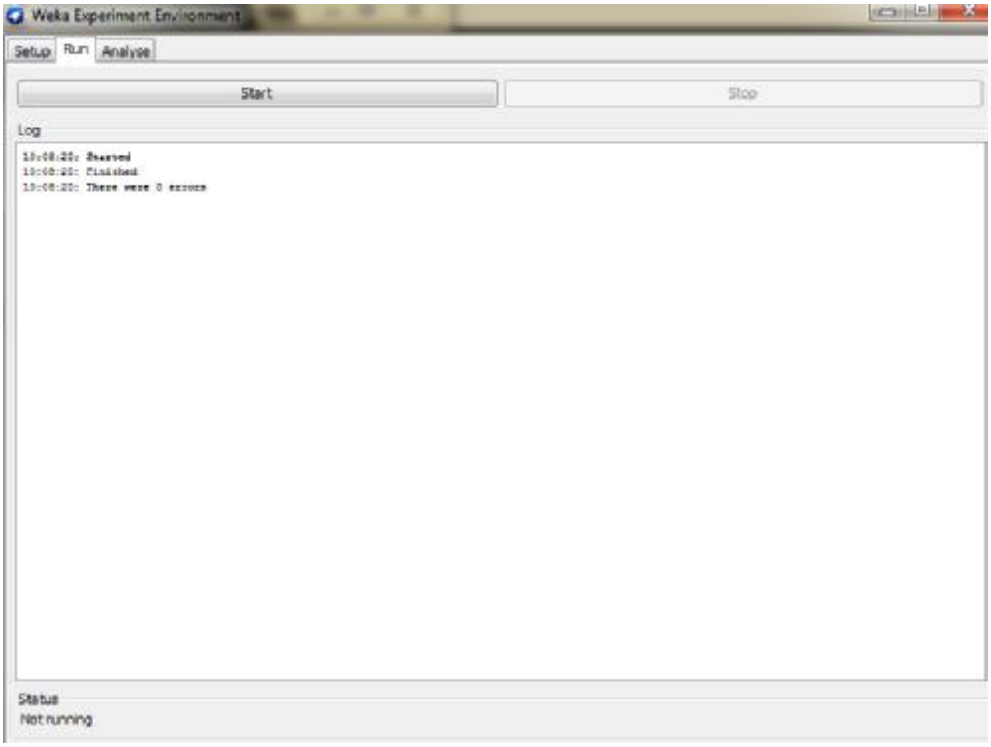
Προεπιλεγμένη επιλογή είναι ο ZeroR αλλιώς είναι αυτός που χρησιμοποιήσαμε τελευταία φορά. Έχουμε πολλές επιλογές για ποιον αλγόριθμο θα επιλέξουμε πατώντας το choose.



Με τις επιλογές load και save μπορούμε να σώσουμε ή να καλέσουμε ένα αρχείο μας που σε κάποιες περιπτώσεις που οι ρυθμίσεις που κάνουμε είναι χρονοβόρες μας βοηθάει αρκετά. Πρέπει απλά να προσέξουμε ότι τα αρχεία του experiment αποθηκεύονται σε δυαδική μορφή η οποία υποστηρίζεται από την Java.Κάποιες εκδόσεις του προγράμματος Weka δεν υποστηρίζουν συμβατικότητα ανάμεσα σε αρχεία *.exp διαφορετικών εκδόσεων. Με ένα διαφορετικό τρόπο αποθήκευσης να είναι το *.xml.

4.1.2.2 Run Experiment

Αφού έχουμε τελειώσει το setup μας πάμε να το τρέξουμε. Πατώντας στην καρτέλα run το κουμπί start έχουμε το ακόλουθο αποτέλεσμα:

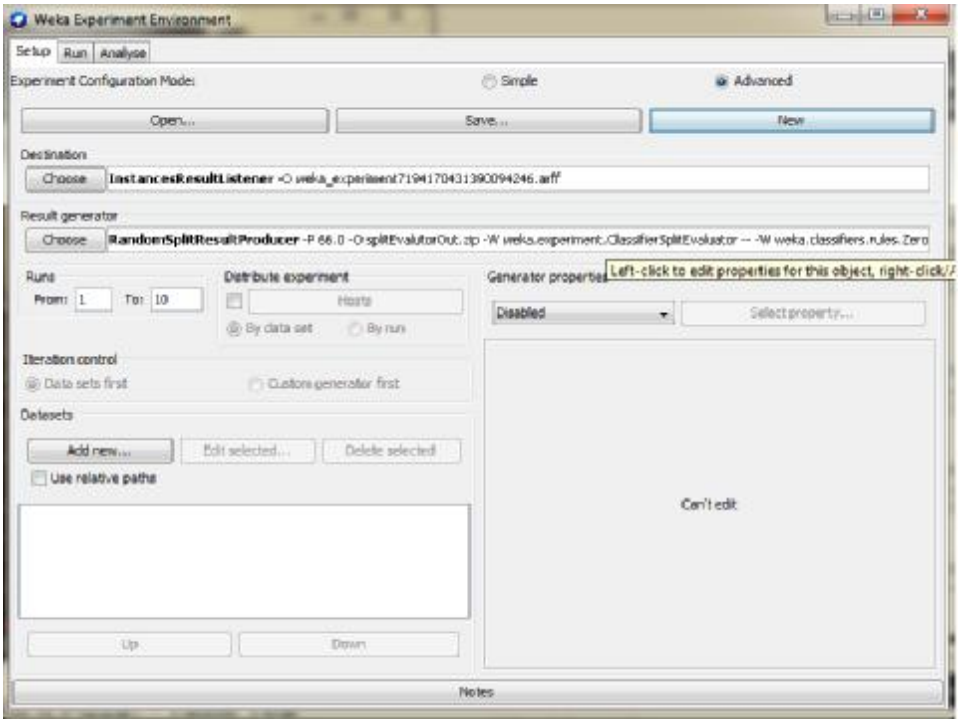


Αν οι ρυθμίσεις μας ήταν σωστές θα πρέπει να μας εμφανιστούν αυτά τα τρία μηνύματα.

4.1.3 Advanced

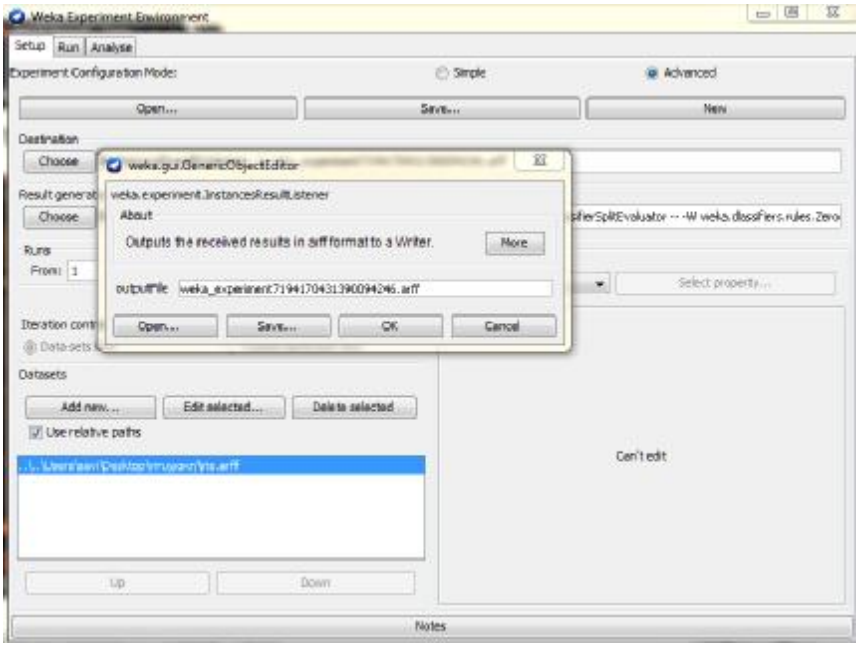
4.1.3.1 New Experiment.

Όπως και στην simple επιλογή πάλι επιλέγουμε να ξεκινήσουμε ένα νέο πείραμα.



Για να επιλέξουμε ποια δεδομένα θα επιλέξουμε για το πείραμα μας πρώτα θα επιλέξουμε το “use relative paths” στο κομμάτι του Datasets και μετά πατάμε το Add new.Για να

αποθηκεύσουμε σε ποιο αρχείο θα αποθηκευτεί το πρόγραμμα μας επιλέγουμε το Instances-ResultListener στην επιλογή Destination. Αφού την επιλέξουμε κάνουμε κλίκ πάνω στην επιλογή που εμφανίζεται στο διπλανό κενό και μας εμφανίζεται το ακόλουθο παράθυρο όπου πληκτρολογούμε το πώς θέλουμε να ονομάσουμε το αρχείο μας.

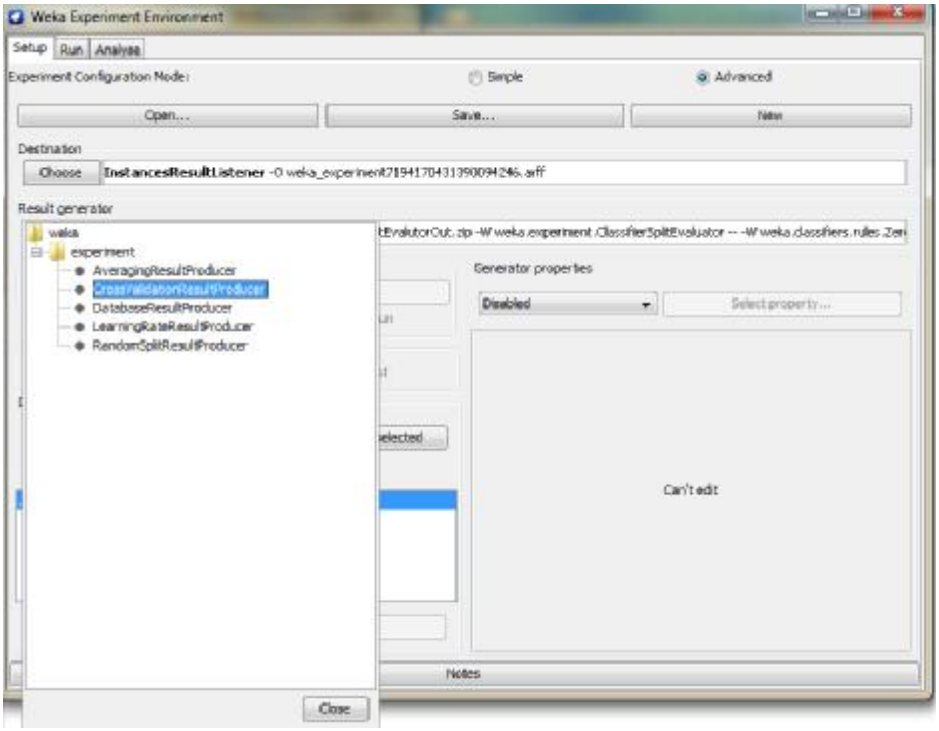


Για να τρέξουμε το πείραμα μας στην καρτέλα run επιλέγουμε το start και περιμένουμε να δούμε τα τρία μηνύματα ώστε να σιγουρευτούμε ότι το πείραμα μας έχει υπολογιστεί σωστά.Αν η διαδικασία μας ήταν σωστή στο αρχείο μας θα δούμε ότι έχουν αποθηκευτεί κάποιες γραμμές:

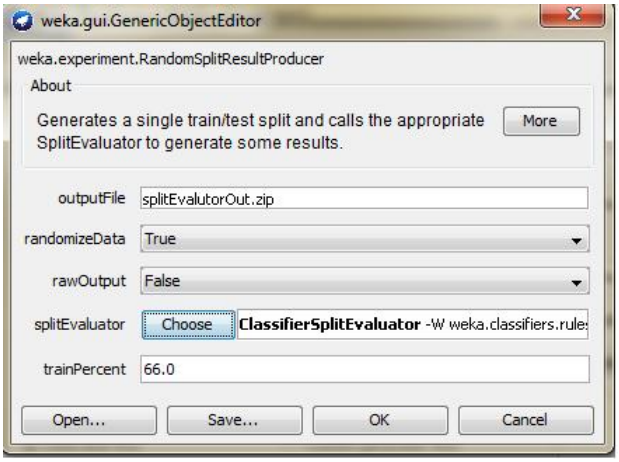
@relation InstanceResultListener @attribute Key_Dataset {iris} @attribute Key_Run {1,2,3,4,5,6,7,8,9,10} @attribute Key_Scheme {weka.classifiers.rules.ZeroR,weka.classifiers.trees.J48} @attribute Key_Scheme_options {,'-C 0.25 -M 2'} @attribute Key_Scheme_version_ID {48055541465867954,-217733168393644444} @attribute Date_time numeric @attribute Number_of_training_instances numeric @attribute Number_of_testing_instances numeric @attribute Number_correct numeric @attribute Number_incorrect numeric @attribute Number_unclassified numeric @attribute Percent_correct numeric @attribute Percent_incorrect numeric @attribute Percent_unclassified numeric @attribute Kappa_statistic numeric @attribute Mean_absolute_error numeric @attribute Root_mean_squared_error numeric @attribute Relative_absolute_error numeric @attribute Root_relative_squared_error numeric @attribute SF_prior_entropy numeric @attribute SF_scheme_entropy numeric @attribute SF_entropy_gain numeric @attribute SF_mean_prior_entropy numeric @attribute SF_mean_scheme_entropy numeric @attribute SF_mean_entropy_gain numeric @attribute KB_information numeric 16 @attribute KB_mean_information numeric @attribute KB_relative_information numeric @attribute True_positive_rate numeric @attribute Num_true_positives numeric @attribute False_positive_rate numeric @attribute Num_false_positives numeric @attribute True_negative_rate numeric @attribute Num_true_negatives numeric @attribute False_negative_rate numeric @attribute Num_false_negatives numeric @attribute IR_precision numeric @attribute IR_recall numeric @attribute F_measure numeric @attribute Area_under_ROC numeric @attribute Time_training numeric @attribute

Time_testing numeric @attribute Summary {'Number of leaves: 3\nSize of the tree: 5\n', 'Number of leaves: 5\nSize of the tree: 9\n', 'Number of leaves: 4\nSize of the tree: 7\n'} @attribute measureTreeSize numeric @attribute measureNumLeaves numeric @attribute measureNumRules numeric @data
iris,1,weka.classifiers.rules.ZeroR,,48055541465867954,20051221.033,99,51,
17,34,0,33.333333,66.666667,0,0,0.444444,0.471405,100,100,80.833088,80.833088,
0,1.584963,1.584963,0,0,0,0,1,17,1,34,0,0,0,0,0.333333,1,0.5,0.5,0,0,?,?,?,?

Μπορούμε να αλλάξουμε τις παραμέτρους του πειράματος απλά πατώντας την επιλογή result generator panel.

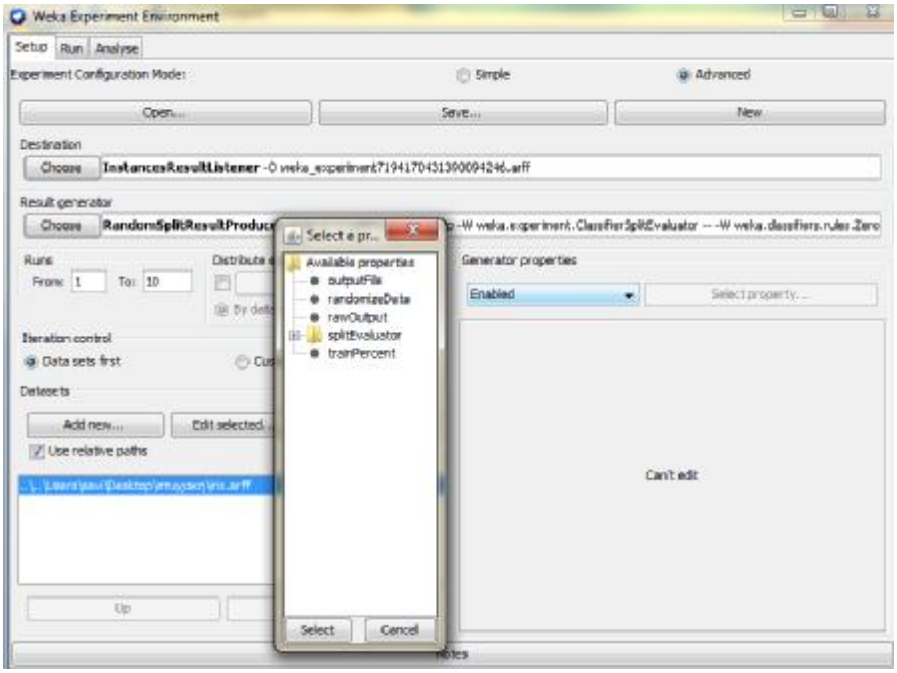


Η κάθε παράμετρος που επιλέγουμε αναφέρει ποιος είναι ο σκοπός της :

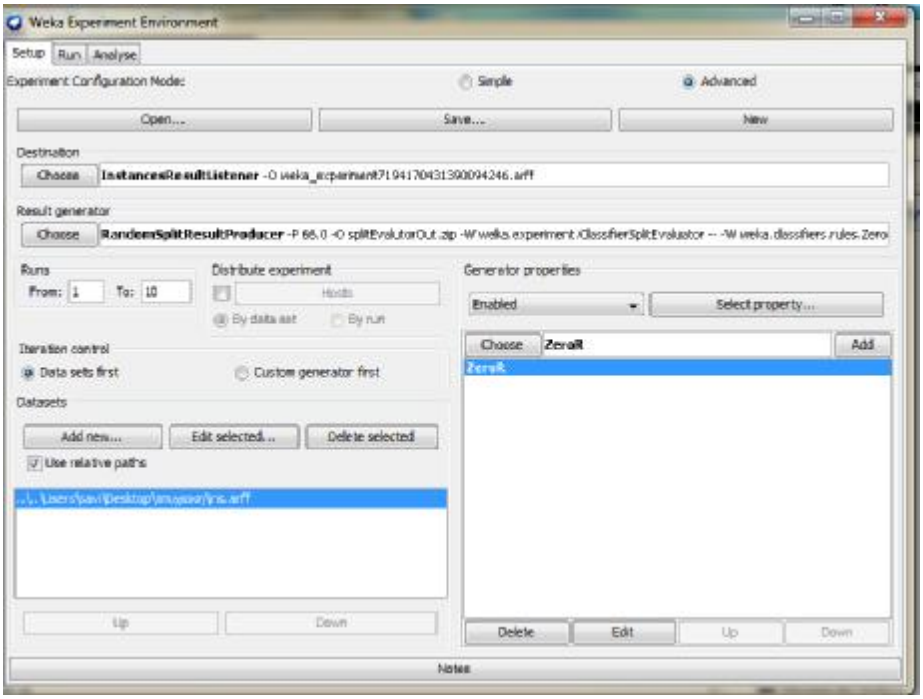


Αν πατήσουμε την επιλογή choose στο splitEvaluator βλέπουμε ποιόν αλγόριθμο έχουμε επιλέξει και τις ρυθμίσεις που μπορούμε να κάνουμε. Για ακόμα μια φορά ο προεπιλεγμένος αλγόριθμος είναι ο ZeroR .Ο συγκεκριμένος αλγόριθμος όταν χρησιμοποιείται στον experimenter δεν έχει έξτρα ρυθμίσεις .Η επιλογή capabilities ανοίγει ένα παράθυρο και

εμφανίζονται οι μορφές των αρχείων που μπορεί να επεξεργαστεί ο αλγόριθμος. Στην περίπτωση που θέλουμε να χρησιμοποιήσουμε στο πείραμα μας παραπάνω από έναν αλγόριθμο τότε στο generator properties επιλέγουμε το enable.



Κάνοντας κλικ και ανοίγοντας το splitEvaluator βλέπουμε τις διαθέσιμες επιλογές.Επιλέγουμε την επιλογή classifier και πατάμε το select και μας εμφανίζεται η παρακάτω εικόνα.



Για να διαλέξουμε ποιόν αλγόριθμο θέλουμε να χρησιμοποιήσουμε πατάμε το κουμπί choose.Η επιλογή filter μας επιτρέπει να ξεχωρίσουμε ποιοι αλγόριθμοι μπορούν να επεξεργαστούν συγκεκριμένες μεταβλητές. Με το remove filter ότι είχαμε επιλέξει

επανέρχεται στην αρχική του επιλογή. Ακόμα μπορούμε να επιλέξουμε να τρέξουμε το πείραμα μας σε παραπάνω από ένα αρχεία επιλέγοντας το στο Add new στο ταμπλό datasets.

Στον experimenter έχουμε την δυνατότητα να αποθηκεύσουμε τα αποτελέσματα από το πείραμα μας χωρίς να τα έχουμε αναλύσει. Για να ενεργοποιήσουμε αυτή την επιλογή ανοίγουμε το resultProducer πατώντας στο ResultGenerator και κάνοντας κλικ στο rawOutput επιλέγουμε το true.Τα δεδομένα μας θα αποθηκευτούν αυτόματα στο συμπιεσμένο αρχείο splitEvaluator.zip.

Με τα αποτελέσματα να είναι :

ClassifierSplitEvaluator: weka.classifiers.trees.J48 -C 0.25 -M 2(version -217733168393644444)Classifier model:

J48 pruned tree ------------------

petalwidth <= 0.6: Iris-setosa (33.0)

petalwidth > 0.6 |

petalwidth <= 1.5: Iris-versicolor (31.0/1.0) |

petalwidth > 1.5: Iris-virginica (35.0/3.0)

Number of Leaves : 3

Size of the tree : 5

Correctly Classified Instances 47 92.1569 %

Incorrectly Classified Instances 4 7.8431 %

Kappa statistic 0.8824

Mean absolute error 0.0723

Root mean squared error 0.2191

Relative absolute error 16.2754 %

Root relative squared error 46.4676 %

Total Number of Instances 51

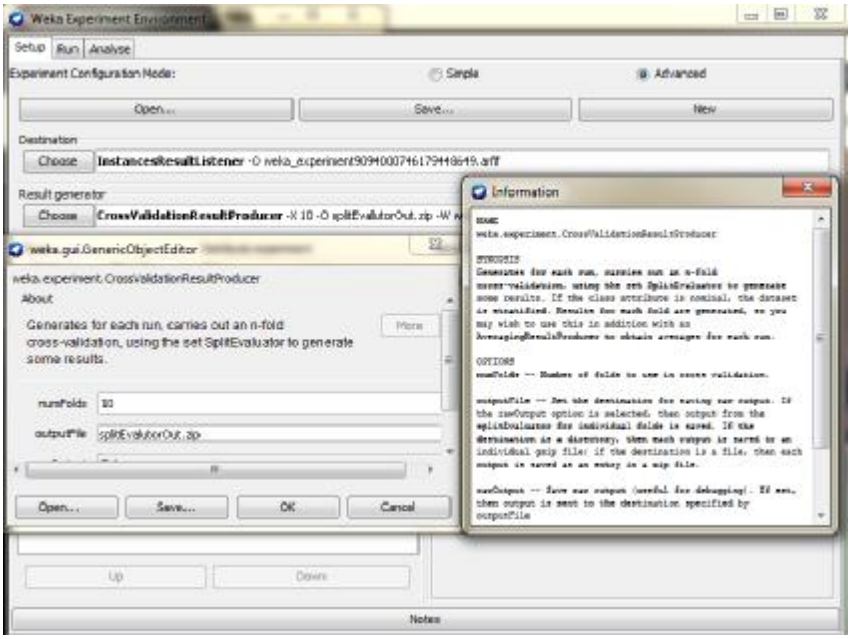
measureTreeSize : 5.0

measureNumLeaves : 3.0

measureNumRules : 3.0

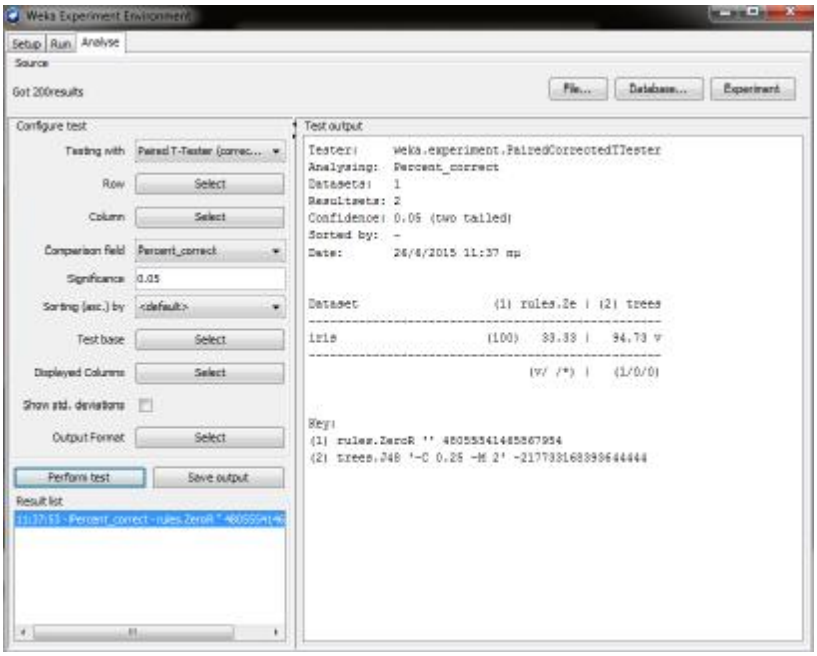
4.1.3.2 Cross-Validation Result Producer

Στο result generator όταν επιλέξουμε το cross-validation result producer και εμφανίζεται ένα παράθυρο με παραμέτρους .Σε περίπτωση που δεν ξέρουμε ακριβώς τον σκοπό του μπορούμε να βρούμε περισσότερες πληροφορίες στο more.



Όπως και προηγουμένως πάλι έχουμε την δυνατότητα να επιλέξουμε και άλλους αλγόριθμους και άλλα αρχεία ώστε να τρέξουμε το πείραμα μας με τα βήματα που αναφέραμε προηγουμένως .

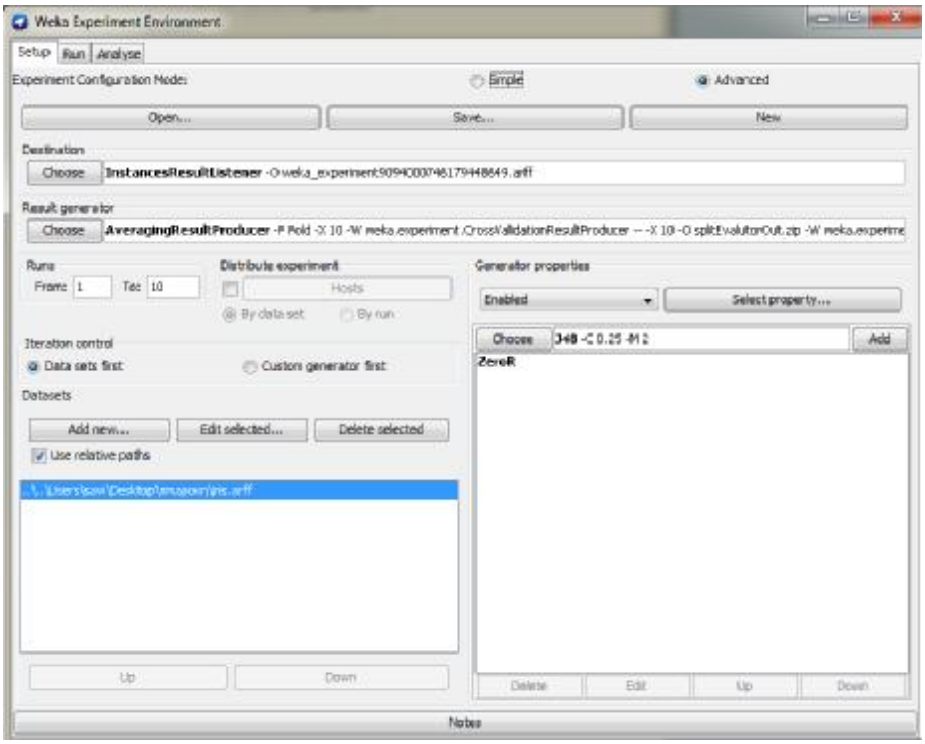
Για να είμαστε σίγουροι ότι τα αποτελέσματα που έχουμε πάρει είναι σωστά μπορούμε να ελέγξουμε στην καρτέλα analyse και αφού πατήσουμε το perform test το test output.Βλέπουμε ότι έχει τρέξει το πείραμα μας με τον ZeroR και τον j48 στο αρχείο iris.arff.



4.1.3.3 Averaging Result Producer

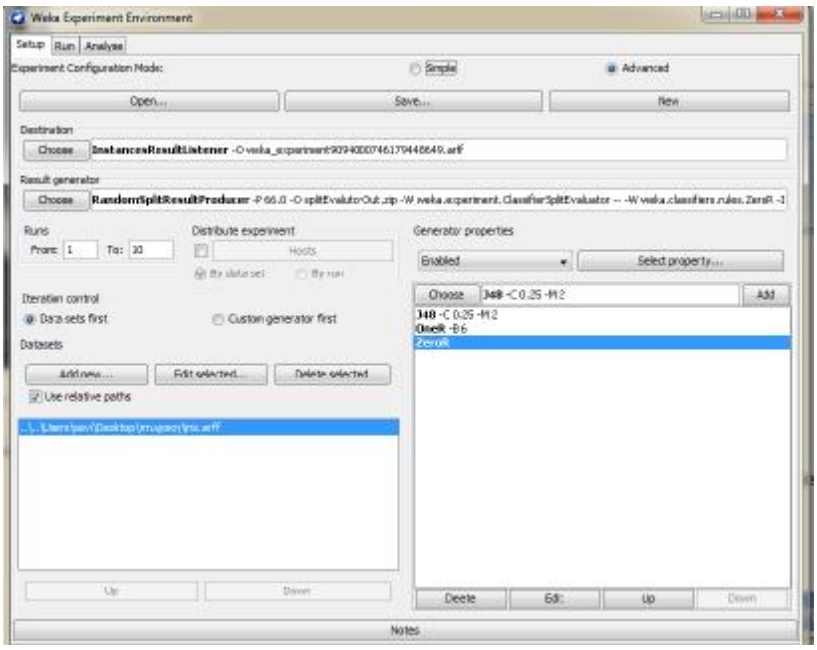
Ακόμα μια επιλογή με παρόμοια αποτελέσματα με την CrossValidationResultProducer. Η κύρια διαφορά είναι ότι με αυτήν την επιλογή βγάζουμε μέσο όρο από τα σετ που τρέξαμε το πείραμα μας(συνήθως το συνδυάζουμε με cross-validation για πιο ακριβή αποτελέσματα). Πατώντας στο GenericObjectEditor και επιλέγοντας το AveragingResultProducer είμαστε έτοιμοι να αρχίσουμε τις ρυθμίσεις του. Οι ρυθμίσεις γίνονται με τον ίδιο ακριβώς τρόπο

όπως και στην cross-validation result producer δίνοντας μας και πάλι την δυνατότητα να επιλέξουμε παραπάνω από ένα αρχεία και αλγόριθμους ώστε να πειραματιστούμε.

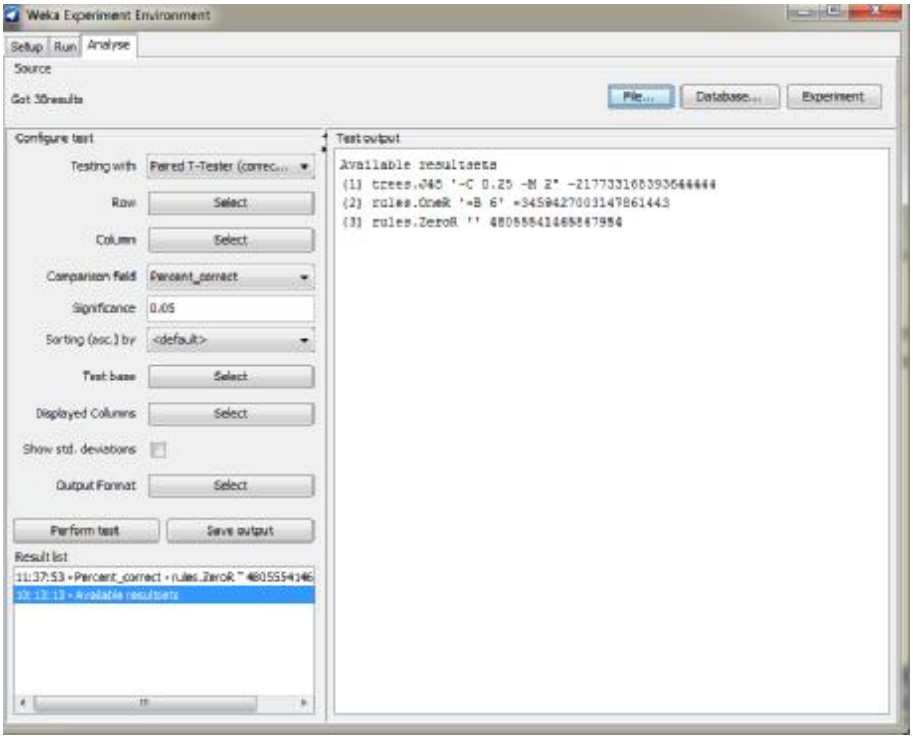


4.1.3.4 Analyze

Αφού έχουμε επιλέξει τις τεχνικές, τα αρχεία και τους αλγόριθμους που θα χρησιμοποιήσουμε είναι ώρα να αναλύσουμε τα αποτελέσματα μας. Στο setup που χρησιμοποιήσαμε είναι για ακόμα μια φορά το αρχείο iris.arff και επιλέξαμε τους αλγόριθμους J48 ,ZeroR,OneR το οποίο θα τρέξει 10 φορές με αναλογία στο training,testing 34% και 66%.

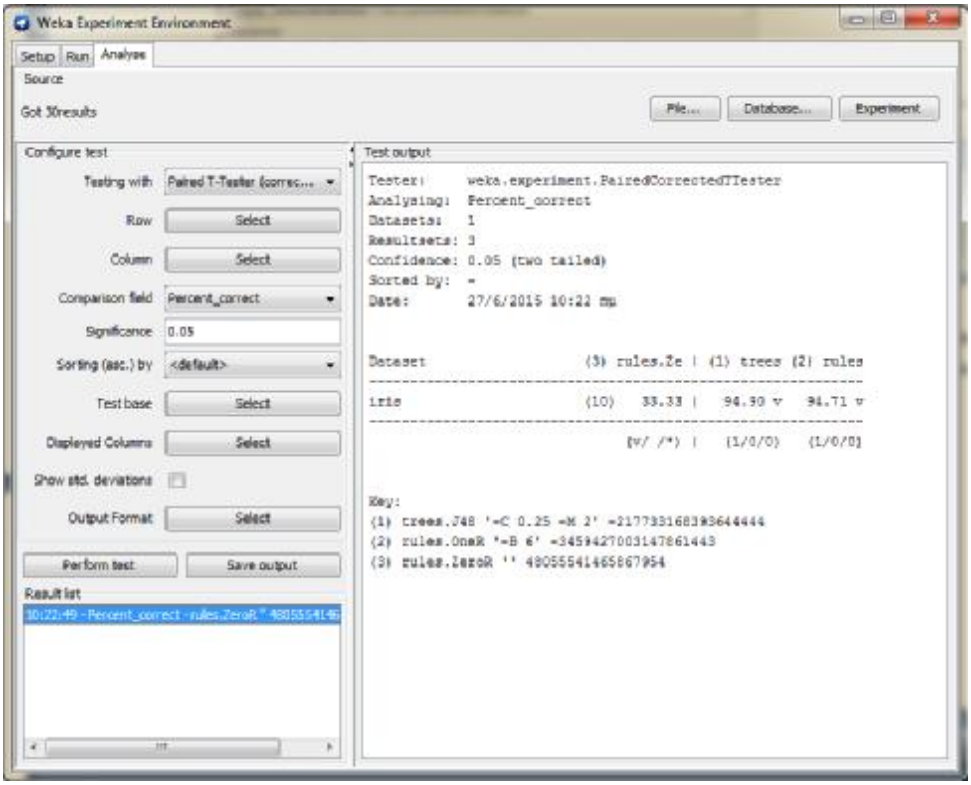


Τρέχουμε το πείραμα για να δούμε ότι οι ρυθμίσεις μας ήταν σωστές και μετά επιλέγουμε την καρτέλα analyse.Πατάμε στο κουμπί experiment για να αναλύσουμε το πείραμα που μόλις τρέξαμε.



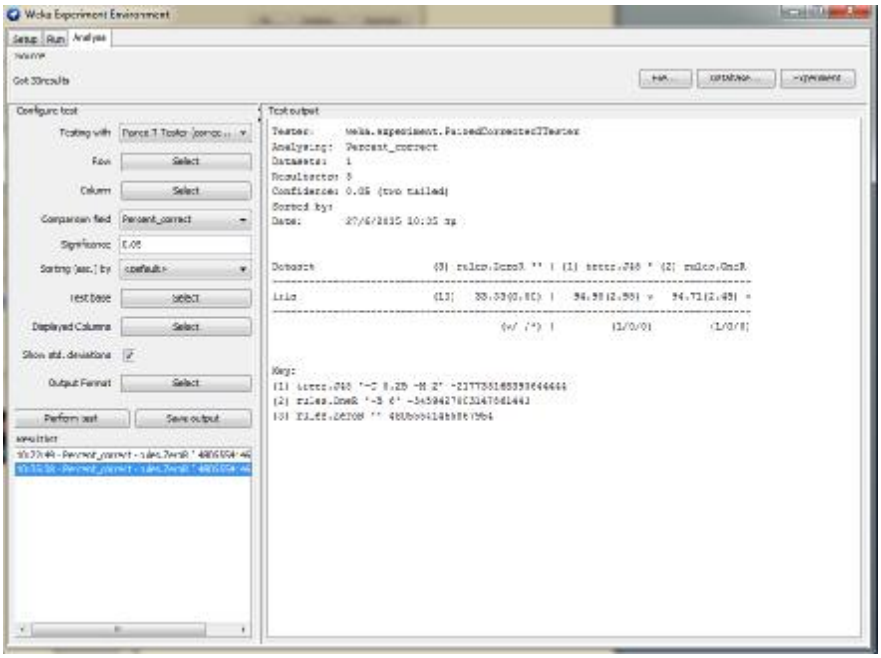
Αν θέλουμε να αναλύσουμε κάποιο άλλο αρχείο μπορούμε να επιλέξουμε την επιλογή file για arff αρχεία ή csv και την επιλογή database για αρχεία Jdbc.

Στο ταμπλό source βλέπουμε επίσης τον αριθμό των αποτελεσμάτων που έχουμε δηλαδή το 30.Αυτό σημαίνει τρείς αλγόριθμοι (OneR,ZeroR,J48) επί δέκα επαναλήψεις. Στην αριστερή στήλη στο comparison field επιλέγουμε το percent_correct και μετά στο test base διαλέγουμε το ZeroR (για να θέσουμε με ποιον αλγόριθμο θα συγκρίνουμε τους άλλους δύο. Τέλος πατάμε στο perform test συγκρίνοντας έτσι τους τρείς αλγόριθμους μαζί.



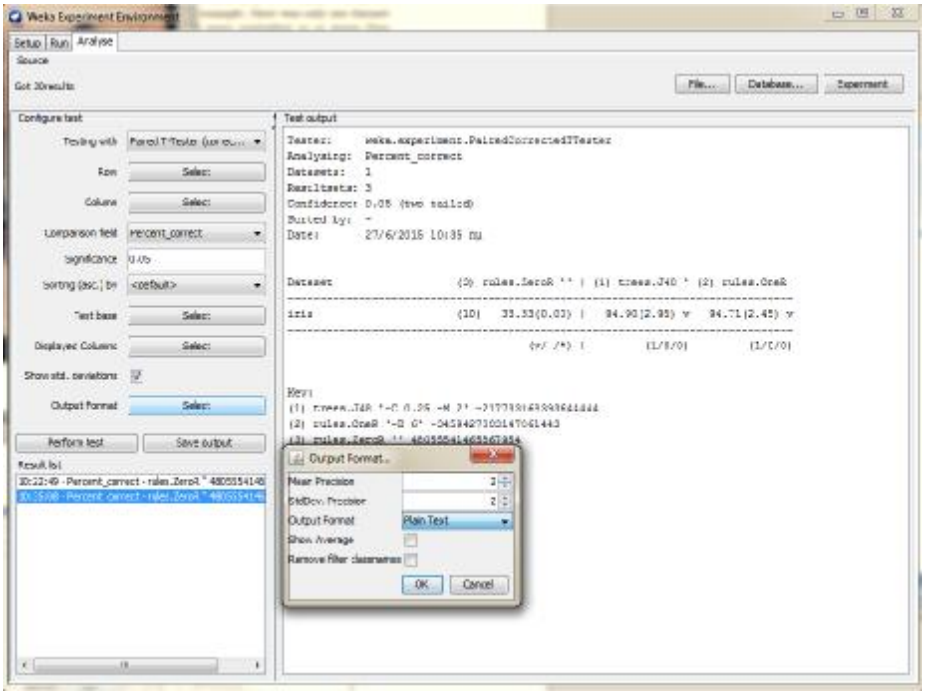
Τα σύμβολα v ή * σημαίνουν ότι κάποιος αλγόριθμος είναι στατιστικά καλύτερο(v) ή χειρότερο (*) από τον αλγόριθμο που θέσαμε σαν μέτρο σύγκρισης. Βλέπουμε ότι ο OneR και ο J48 είναι στατιστικά καλύτεροι από τον ZeroR (33.33 vs 94.9 vs 94.71).Στην αμέσως επόμενη γραμμή παρατηρείτε ότι υπάρχει ακόμα ένα σετ με νούμερα στους αλγόριθμους OneR και J48.Τα νούμερα αυτά ακολουθούν το φορμάτ (xx,yy,zz) που σημαίνει πόσες φορές στο πείραμα που τρέξαμε ο αλγόριθμος ήταν καλύτερος από την βάση(xx) ίδιος με την βάση(yy) και χειρότερος(zz).

Επίσης μπορούμε να υπολογίσουμε και άλλα στατιστικά μέτρα όπως η τυπική απόκλιση. Για να το κάνουμε αυτό επιλέγουμε το κουτάκι από την αριστερή στήλη στο show Std. Deviation.



Στην επιλογή output Format μπορούμε να επιλέξουμε την ακρίβεια της τυπικής απόκλισης (MAP) και του μέσου καθώς και την μορφή του αποτελέσματος. Αν επιλέξουμε και την επιλογή show average θα εμφανιστεί μια νέα γραμμή στην οποί θα εμφανίζεται ο μέσος της κάθε στήλης. Οι μορφές που υποστηρίζονται είναι :

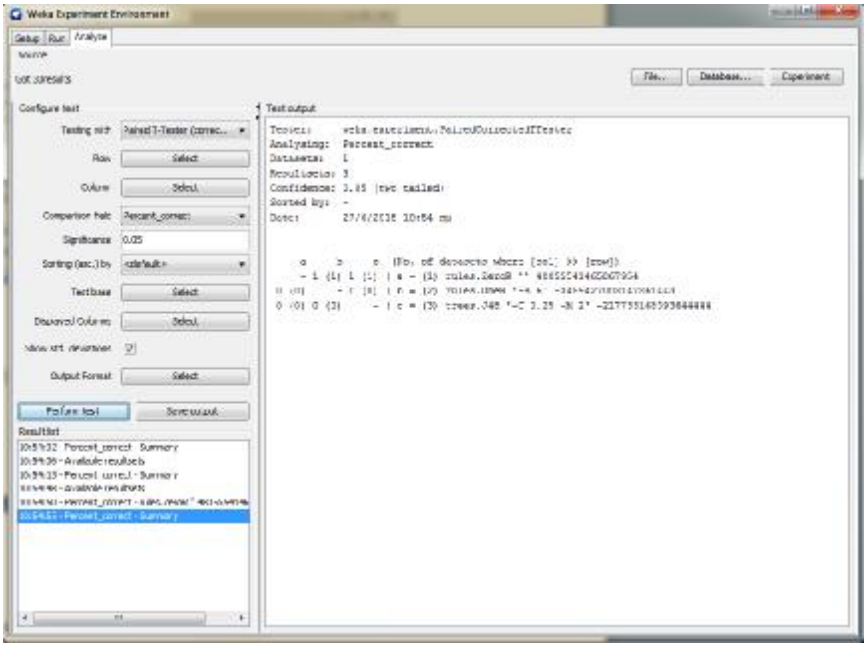
- Csv
- GNplot
- Html
- Latex
- Plain text
- Significanse only.



Για να αποθηκεύσουμε τα αποτελέσματα μας επιλέγουμε το save output.

4.1.3.4.1 Summary Test

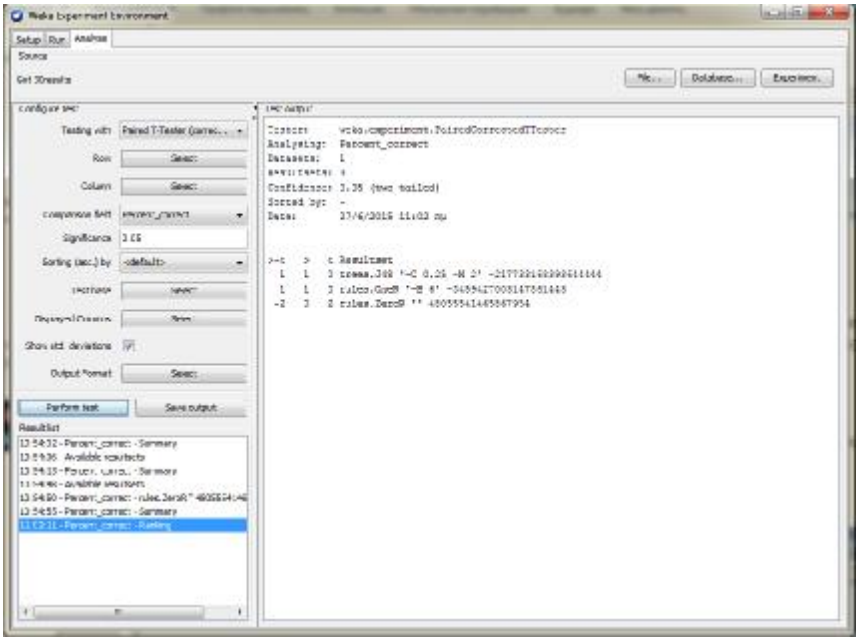
Επιλέγοντας το summary test από το test base και πατώντας το perform test εμφανίζονται τα παρακάτω αποτελέσματα.



Στο πείραμα μας η πρώτη γραμμή (-1 1)σημαίνει ότι η στήλη b(OneR) είναι καλύτερη από την γραμμή a (ZeroR) και ότι η στήλη c (J48) είναι επίσης καλύτερη από την γραμμή a.Το 0 σημαίνει ότι καμία φορά δεν έχει υπάρξει κάποια γραμμή καλύτερη από την άλλη.

4.1.3.4.2 Ranking Test

Επιλέγοντας το Ranking test από το test base και τρέχοντας το πείραμα μας εμφανίζονται τα κάτωθι αποτελέσματα.



Το Ranking Test κατατάσσει τους αλγόριθμους σύμφωνα με τις πόσες φορές σύμφωνα με τις νίκες του απέναντι στους άλλους. Συγκεκριμένα βλέπουμε ότι ο ZeroR έχει χάσει δύο φορές(-

2) ενώ οι άλλοι δύο έχουν βρεθεί μία φορά καλύτεροι από τον ZeroR (1) και ισόπαλοι με τον άλλον αντίστοιχα(1).

5.1 Knowledge Flow

Το knowledge flow είναι μια διαφορετική έκδοση του explorer ως μέσο γραφικής αναπαράστασης του weka.Πρέπει να επισημάνουμε ότι ΔΕΝ υποστηρίζονται όλες οι επιλογές του explorer καθώς δεν είναι 100 % ολοκληρωμένο project.Από την άλλη υπάρχουν επιλογές που δεν υποστηρίζει ο explorer στο knowledge flow.

Ο χρήστης μπορεί να επιλέξει τις λειτουργίες του Weka από μια μπάρα εργαλείων και να το τοποθετήσει σε ένα πλέγμα και να τα συνδέσει ώστε να δημιουργήσει μια ροή δεδομένων(knowledge flow) για επεξεργασία και ανάλυση.Μέχρι στιγμής όλοι οι αλγόριθμοι ομαδοποίησης,συσταδοποίησης,φίλτρα είναι διαθέσιμοι.

Στο knowledge flow μπορούμε να δώσουμε αρχεία σε παρτίδες ή σταδιακά(Αξίζει να αναφέρουμε ότι στον explorer μπορούμε μόνο σε παρτίδες –datasets-) και για αυτό τον λόγο έχουμε ειδικούς αλγόριθμους για αυτή την επεξεργασία :

- AOBE
- IB1
- IBk
- KSTar
- NaiveBayesMultinomialUpdateable
- NaiveBayesUpdateable
- NNge
- Winnow

Το knowledge flow μας παρέχει:

- Επεξεργασία δεδομένων σε παρτίδες ή σταδιακά
- Μπορούμε να επεξεργαστούμε ταυτόχρονα διαφορετικές παρτίδες παράλληλα χωρίς να σταματάει η προηγούμενη
- Μπορούμε να συνδιάσουμε διαφορετικά φίλτρα μαζί
- Μπορούμε να δούμε την διαδικασία cross validation βήμα βήμα.
- Μπορούμε να δούμε την επεξεργασία τον αλγορίθμων ομαδοποίησης παράλληλα της επεξεργασίας
- Έχουμε την δυνατότητα να προσθέσουμε έξτρα επιλογές με plugins.

5.1.2 Επιλογές knowledge flow:

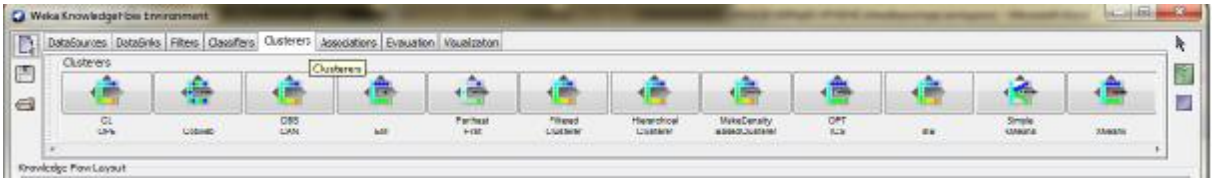
5.1.2.1 Datasources

5.1.2.2 Datasinks

5.1.2.3 Filters

5.1.2.4 Classifiers

5.1.2.5 Clusters

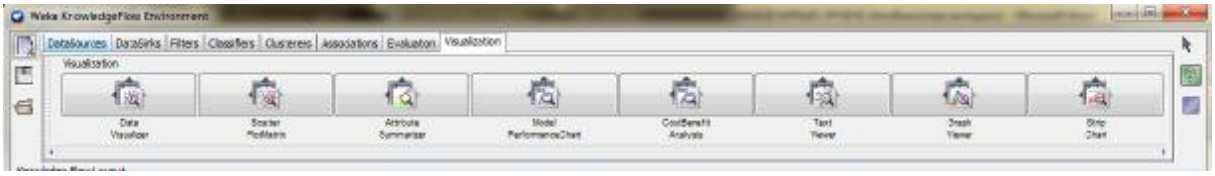


5.1.2.6 Associations

5.1.2.7 Evaluation

· TrainingsetMaker:Δημιουργεί ένα training set από ένα αρχείο
· TestSetMaker:Δημιουργεί ένα test set από ένα αρχείο
· CrossValidationFoldMaker:Διαχωρίζει ένα αρχείο,training ή test set σε μέρη
· TrainTestSplitMaker: Διαχωρίζει ένα αρχείο,training ή test set σε training ή test set
· ClassAssigner:Ορίζει μια στήλη σε κλάση για οποιαδήποτε αρχείο ή training ή test set
· ClassifierPerformanceEvaluator:Αξιολογεί την απόδοση ενός αλγόριθμου ομαδοποίησης
· IncrementalClassifierEvaluator :Αξιολογεί την απόδοση ενός αλγόριθμου ομαδοποίησης με σταδιακά δεδομένα.
· ClustererPerformanceEvaluator:Αξιολογεί την απόδοση ενός αλγόριθμου συσταδοποίησης
· PredictionAppender:Τοποθετεί τις προβλέψεις ενός αλγόριθμου ομαδοποίησης σε ένα test set.

5.1.2.8 Visualization



· Data Visualizer:Είναι ένα εξάρτημα το οποίο ανοίγει ένα ταμπλό για να κάνουμε αναπαράσταση δεδομένων σε διάγραμμα διασποράς δύοδιαστάσεων
· ScatterPlotMatrix: Ένα εξάρτημα το οποίο ανοίγει ένα ταμπλό που περιέχει μια μήτρα από μικρού μεγέθους διαγράμματων διασποράς.
· AttributeSummarizer: Ένα εξάρτημα το οποίο ανοίγει ένα ταμπλό που περιέχει ιστογράμματα.
· ModelPerformanceChart: Ένα εξάρτημα το οποίο ανοίγει ένα ταμπλό για αναπαράστασει καμπυλών.
· TestViewer:Εξάρτημα που απεικονίζει δεδομένα κειμένου
· GraphViewer: Ένα εξάρτημα το οποίο ανοίγει ένα ταμπλό για απεικόνιση μοντέλα δέντρου.
· StripChart: Ένα εξάρτημα το οποίο ανοίγει ένα ταμπλό που χρησιμοποιείται κυρίως για να βλέπουμε κατευθείαν την απόδοση τον αλγορίθμων με σταδιακά αρχεία δεδομένων.

5.1.3 Χρήση Knowledge Flow

Ας κάνουμε με την χρήση του knowledge flow το δεντρόγραμμα j48 για το αρχείο iris.arff που χρησιμοποιήσαμε προηγουμένως.

Στην καρτέλα DataSources επιλέγουμε το ArffLoader και μετά κάνουμε κλικ στο σημείο που θέλουμε να το τοποθετήσουμε. Κάνοντας δεξί κλικ εμφανίζεται ένα παράθυρο και πατώντας το configure επιλέγουμε το αρχείο που θα χρησιμοποιήσουμε.

Μετά στην καρτέλα Evaluation επιλέγουμε το ClassAssigner και το τοποθετούμε.

Συνδέουμε τα δύο εξαρτήματα κάνοντας δεξί κλικ στο ArFFLoader και επιλέγοντας το dataset .Εμφανίζεται μια κορδέλα την οποία την ενώνουμε με το ClassAssigner.Τώρα αν κάνουμε δεξί κλικ στο ClassAssigner και μετά configure μπορούμε να επιλέξουμε την κλάση μας.

Επιλέγουμε το CrossValidationFoldMaker και το τοποθετουμε. Συνδέουμε το ClassAssigner με το CrossValidationFoldMaker επιλέγοντας το dataset.

Στην καρτέλα Classifiers βρίσκουμε το J48 και το τοποθετούμε.Συνδέουμε το CrossValidationFoldMaker δύο φορές (μία για το training set και μια για το test set)

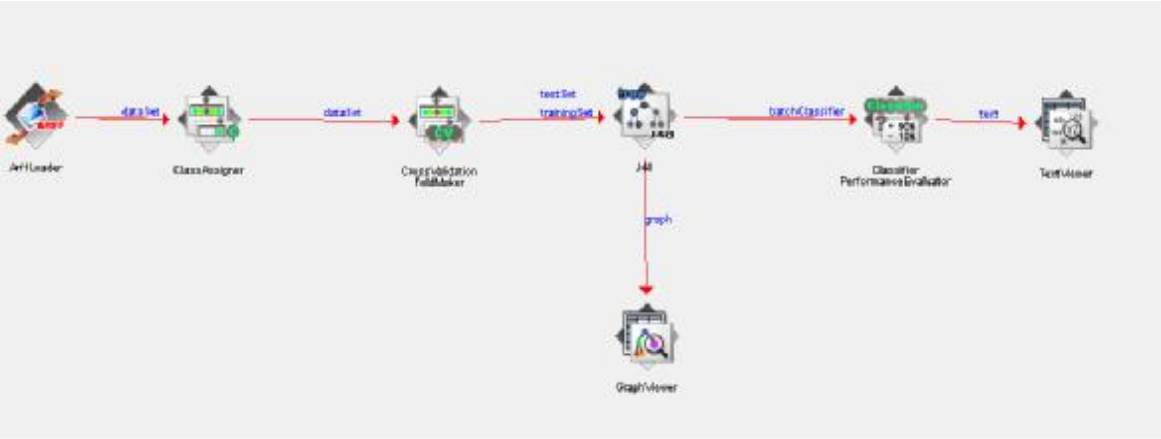
Μετά πάλι στην καρτέλα Evaluation επιλέγουμε το ClassifierPerformanceEvaluator και το συνδέουμε με το J48 επιλέγοντας το batchClassifier.

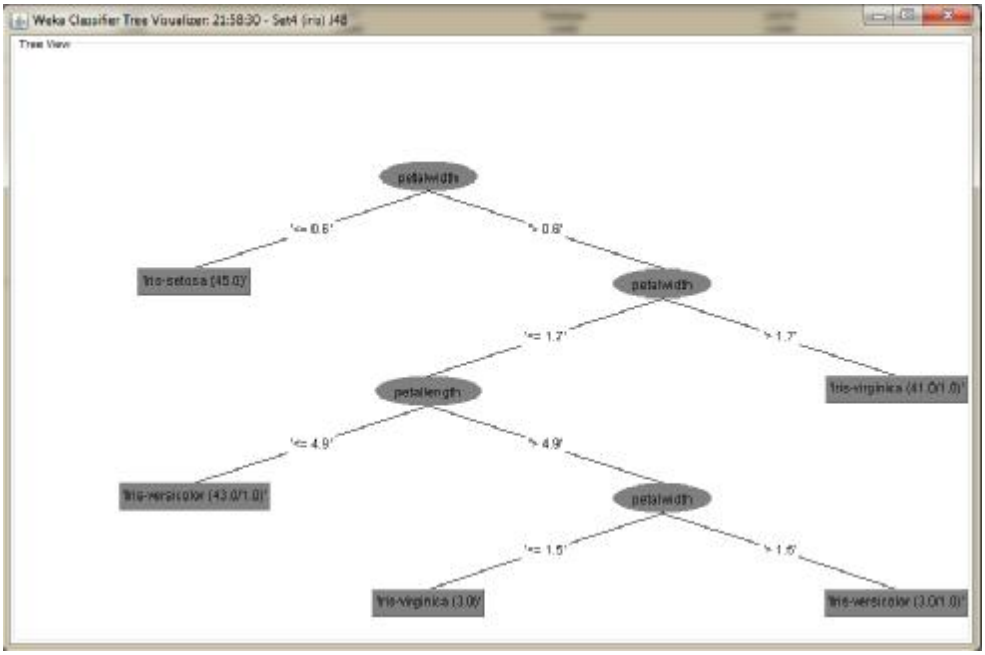
Στην καρτέλα Visualization επιλέγουμε το TextViewer και το συνδέουμε με το ClassifierPerformanceEvaluator επιλέγοντας το text.

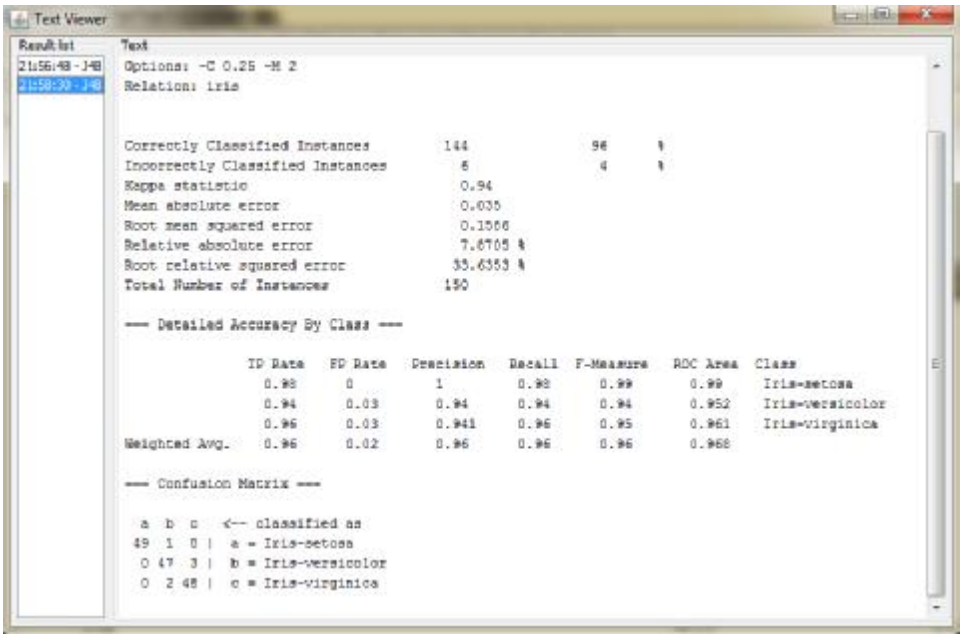
Ακόμα επιλέγουμε το GraphViewer στο J48 επιλέγοντας το graph .

Για να το τρέξουμε κάνουμε κλικ στο ArffLoader και start loading.

Για να δούμε τα αποτελέσματα Πατάμε Show Results ή στο Text Viewer ή στο Graph Viewer (εφόσον έχουμε ανοίξει το μενού με δεξί κλικ)







ΚΕΦΑΛΑΙΟ 6

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία είχαμε ως γενικό σκοπό να μελετήσουμε την έννοια της εξόρυξης δεδομένων με στόχο την εισαγωγή και κατανόηση της συγκεκριμένης εργασίας.
Μέσα από τη συγκεκριμένη εργασία επετεύχθη η προσέγγιση της εξόρυξης δεδομένων. Η πρώτος στόχος της τρέχουσας εργασίας, μέσα από την ανάλυση των διάφορων ορισμών και εργασιών σκοπεύει στην περιγραφή και στην κατανόηση των ωφελειών που προκύπτουν καταρχάς για τους επιχειρηματίες και τους διάφορους χρήστες που θέλουν να ασχοληθούν.

Πρέπει να αναφέρουμε ότι η εξόρυξη δεδομένων είναι ένα εργαλείο που χρησιμοποιείται από τις επιχειρήσεις και μερικές φορές από τις κυβερνήσεις για την πρόβλεψη και την παρακολούθηση των τάσεων με συγκεκριμένους σκοπούς. Όταν την χρησιμοποιούμε το κάνουμε για να βρούμε κάποια συγκεκριμένα αποτελέσματα. Για παράδειγμά η Αμερικάνικη κυβέρνηση το χρησιμοποιεί για να εντοπίσει τρομοκράτες πριν πραγματοποιήσουν τρομοκρατικές ενέργειες. Επιχειρήσεις όπως το Amazon.com χρησιμοποιεί την εξόρυξη για να αυξήσει τις πωλήσεις του ενώ αρκετές βιβλιοθήκες την χρησιμοποιούν για την βελτίωση των λειτουργιών της.
Υπάρχουν κίνδυνοι για την ιδιωτικότητα στην εξόρυξη δεδομένων (. Οι κυβερνήσεις αναγνωρίζουν την προστασία της ιδιωτικής ζωής και τους κίνδυνους που ενέχονται με την εξόρυξη δεδομένων, και έχουν αναπτύξει νομοθεσία που προβλέπει την προσφυγή έναντι πιθανών παραβιάσεων της ιδιωτικής ζωής. Νομοθεσία όπως η PIPEDA στον Καναδά καθορίζει τις κατευθυντήριες γραμμές που εταιρείες και κρατικούς φορείς πρέπει να τηρούν κατά τη συλλογή προσωπικών πληροφοριών. Ομοίως, οι εταιρείες που αναπτύσσουν λογισμικό εξόρυξης δεδομένων είναι ευαίσθητοι για την προστασία της ιδιωτικής ζωής, και αναπτύσσουν το λογισμικό τους με μέτρα για να περιορίσουν τον όγκο των προσωπικών πληροφοριών που συλλέγονται.
Δεύτερος στόχος ήταν η σχεδίαση ενός εγχειριδίου ενός προγράμματος (WEKA) το οποίο μας παρέχει εύκολη πρόσβαση στην εξόρυξη δεδομένων με την απλή μορφή του, χωρίς να υπολείπεται τις εργασίες για τους χρήστες με μεγαλύτερη εμπειρία.
Θα θέλαμε να αναφέρουμε ότι ολοένα και περισσότερος κόσμος ασχολείται με την εξόρυξη δεδομένων από επιστήμονες μέχρι και κυβερνήσεις. Αυτό οφείλεται στον μεγάλο όγκου πληροφοριών που υπάρχει ανεκμετάλλευτος και περιμένει να εμφανίσει χρήσιμες πληροφορίες σε όποιον γνωρίζει.
Διάφοροι οργανισμοί έχουν την δυνατότητα να συλλέγουν πιο εύκολα πληροφορίες για της δραστηριότητες τους ενώ το κόστος αποθήκευσης δεδομένων έχει μειωθεί. Σε συνδυασμό με τις πιέσεις που υφίστανται οι οργανισμοί λόγο ανταγωνισμού που ολοένα και αυξάνεται ψάχνουν να βρούν τον έλεγχο των επενδύσεων τους καθώς και την κατάλληλη αναλογία του δείκτη κόστους/ απόδοσης. Οι εφαρμογές τους είναι ποικίλες από την ανάλυση και πρόβλεψη μέχρι και τον σχεδιασμό και την προώθηση προϊόντων μιας εταιρίας
Οι σημαντικότερες τεχνικές εξόρυξης δεδομένων όπως έχουν περιγραφεί παραπάνω είναι τα δέντρα απόφασης, οι νευρωνικοί αλγόριθμοι, η προσέγγιση του κοντινότερου γείτονα, η συσταδοποίηση, και οι γενετικοί αλγόριθμοι.
Κλείνοντας να αναφέρουμε ότι η εξόρυξη δεδομένων αποτελεί ένα ισχυρό εργαλείο που η χρήση του θα πρέπει να γίνεται με μεγάλη προσοχή και με σκοπούς την αύξηση της ικανοποίησης των πελατών και την παροχή καλύτερων και ασφαλέστερων προϊόντων (εφόσον αναφερόμαστε για επιχειρήσεις). Πιστεύουμε σύμφωνα με την έρευνα που πραγματοποιήσαμε ότι το μέλλον της εξόρυξης είναι υποσχόμενο και εξελισσόμενο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Βικιπαίδεια.(2015).Ανακτήθηκε 18 Μαρτίου,2015 από Βικιπαίδεια wiki: http://el.wikipedia.org/wiki/Εξόρυξη_δεδομένων.

Γολέμη, Ε.(2010).Κρυπτογραφία & Εξόρυξη Δεδομένων. Κεφάλαιο 2 Εξόρυξη Δεδομένων. Μεταπτυχιακή Εργασία, Πανεπιστήμιο Πατρών.

Καψωμενάκης,Νίκος (2015).Google PageRank: Οσα πρεπεί να γνωρίζετε Ανακτήθηκε 28Μαρτίου, 2015: http://www.searchenginemarketing.gr/blog/archives/131

Κωνσταντίνος, Δ. (2007). Τεχνητά Νευρωνικά Δίκτυα Αθήνα: Εκδόσεις Κλειδάριθμος

Μαστρογιάννης, Ν.(2009).Μεθοδολογικό Πλαίσιο Υποστήριξης της Εξόρυξης Γνώσης από δεδομένα με την χρήση αρχών της πολυκριτήριας ανάλυσης αποφάσεων, Τμήμα Διοίκησης Επιχειρήσεων, Πανεπιστήμιο Πατρών.

Μεγαλοοικονόμου,Β. & Μακρής,Χ.(2013).Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης. Ανακτήθηκε 19 Μαρτίου, 2015: https://www.ceid.upatras.gr/webpages/courses/cplusplus/dm/2013_2014_data_mining_introduction.ppt

Μακρής, Αριστομένης (2015).Τεχνολογίες υποστήριξης διοικητικών συστημάτων -Εξόρυξη δεδομένων -Datamining. Ανακτήθηκε 25 Μαρτίου,2015 :http://amacris.ode.unipi.gr/DST/07_DST_DM.pdf

Νανόπουλος Α. & Μανωλόπουλος, Γ. (2008).Εισαγωγή στην Εξόρυξη γνώσης δεδομένων και τις αποθήκες δεδομένων.Αθήνα:Νέων Τεχνολογιών.

Ντούση, Ε.(2003). Εξόρυξη γνώσης από ειδησεογραφικά δεδομένα και συσχετισμός με πραγματικά γεγονότα. Μεταπτυχιακή Εργασία, Τμήμα Μηχανικών Η/Υ και Πληροφορικής.

Παπανικολάου, Δ. (2010).Εφαρμογή τεχνικών εξόρυξης γνώσης στην εκπαίδευση. Μεταπτυχιακή Εργασία, Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών.

Παπαρριζος, Κ. Ιωάννης (2009). Αλγόριθμος εξόρυξης από δεδομένα δομής, περιεχομένου και χρήσης του Παγκόσμιου Ιστού. Διπλωματική Εργασία, Τμήμα Πληροφορικής, Α.Π.Θ.

Πιτούρα Ευαγγελεια (2010).Εξόρυξη δεδομένων, Πανεπιστήμιο Ιωαννίνων Τμήμα Μηχανικών Η/Υ και Πληροφορικής , Ανακτήθηκε 5 Μαίου, 2015: http://www.cs.uoi.gr/~pitoura/courses/dm/introspring11.pdf

Τσιράκης Ν.(2006).Αλγόριθμοι και τεχνικές εξόρυξης δεδομένων από ροές δεδομένων στον παγκόσμιο ιστό. Μεταπτυχιακή Εργασία, Τμήμα Μηχανικών Η/Υ και Πληροφορικής.

Χαλκίδη,Μ & Βαζιργιαννης,Μ.(2005).Εξόρυξη γνωσης απο βάσεις δεδομένων και τον παγκόσμιο ιστό.Αθήνα:Τυποθήτο-Δαρδανός Γ.

Ahmed, S.R.(2004) Application of data mining in Retail Business. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04).IEEE Computer Society Washington DC, USA

Agrawal,R.,Gehrke,J.,Gunopulos,D.,and Raghavan,P.(1998) Automatic Subspace Clustering of High Dimensional Data for Data Mining Application,in Proceedings of the AMC SIGNOD Conference on Management of Data.

Bengio, Y. and Nadeau, C. (1999) Inference for the Generalization Error

Bing Liu, Springe(2011).Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, ISBN: 978-3642194597, 2011

Brin,S & Page,L.(1998)The anatomy of a large-scale hypertextual Web engine,Combuters Networks,30(1-7):107-117,Proceedings of the 7th InternationalWord Wide Workshop

Chen,M.S.,Han,J.,Yu,P.S.(1996) Data Mining: An overwiew from database prespective.IEEE Transactions on knowledge and data engineering,Vol 8

Danham,Margaret (2004).Εισαγωγικά και προιγμένα θέματα εξόρυξης γνώσης απο δεδομένα.Επιμέλεια Ελληνικης έκδοσης Βερύκιος Β. και Θεοδωρίδης Γ.Αθήνα: Εκδόσης Νεων Τεχνολογιών.

Ellen Monk & Bret Wagner (2006). Concepts in Enterprise Resource Planning, Second Edition. Thomson Course Technology, Boston, MA. ISBN 0-619-21663-8.

Etzioni,O.(1996).The world wide web:Quagmire or gold mine,Communications of the ACM.

History of Data Mining.(2015) Ανακτήθηκε 12 Μαρτίου, 2015:http://www.sqldatamining.com/data-mining-basics/history-of-data-mining

Han,J.& Kamber,M.(2006).Data Mining-Concepts and Techniques

Kaufmann,Morgan & Chakrabarti, Soumen (2002). Mining the Web: Discovering Knowledge from Hypertext Data, ISBN: 978-1558607545

Leskovec, Jure Rajaraman, Anand Ullman, Jeffrey David(2014). Mining of Massive Datasets, 2nd Edition cambridge university press,ISBN: 978-1107077232

Murthy S. (1998): Automatic construction of decision trees from data, A multidisciplinary survey. Data Mining and Knowledge Discovery

Ross Quinlan (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

Tan,P.N.,Steinbach, M. & Kumar, V. (2010) Εισαγωγή στην Εξόρυξη Δεδομένων, Θεσ/νίκη: Εκδόσεις Α. Τζιόλα & Υιοί Ο.Ε.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996).From Data Mining to Knowledge Discovery in Databases..Ανακτήθηκε 25

Μαρτίου,2015:http:http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

http://weka.wikispaces.com/Frequently+Asked+Questions

http://www.cs.waikato.ac.nz/~ml/weka/datasets.html

http://www.cs.waikato.ac.nz/~ml/weka/documentation.html (weka manual)

https://en.wikipedia.org/wiki/Weka_(machine_learning)

http://hsqldb.org/

Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)by Ian H. Witten and Mark A. Hall

E.Frank, Machine Learning With WEKA, University of Waikato, New Zealand.

B. Mobasher, Data Preparation and Mining with WEKA, http://maya.cs.depaul.edu/~classes/ect584/WEKA/association_rules.html, DePaul University, 2003.

M. H. Dunham, Data Mining, Introductory and Advanced Topics, Prentice Hall, 2002.