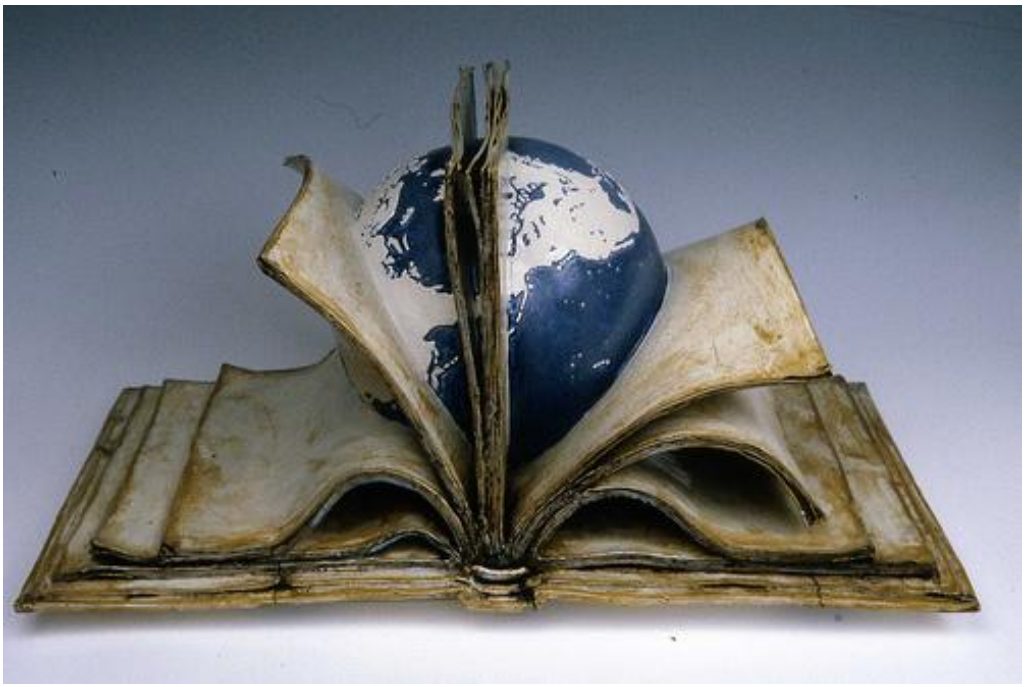
Α.Τ.Ε.Ι ΠΑΤΡΑΣ

Σχολή Διοίκησης Οικονομίας

Τμήμα Επιχειρηματικού Σχεδιασμού και Πληροφοριακών Συστημάτων

Θέμα: « Οι αλγόριθμοι της εξόρυξης δεδομένων: περιγραφή, εφαρμογές, προοπτικές»



ΜΑΛΕΣΚΟΥ ΠΑΝΑΓΙΩΤΑ Α.Μ: 2165

ΚΟΥΜΠΗ ΕΛΕΥΘΕΡΙΑ Α.Μ: 2220

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Πάτρα, Νοέμβριος 2012

Ευχαριστίες

Η παράγραφος αυτή αφιερώνεται στον επιβλέποντα καθηγητή μας κ. Μαστρογιάννη Νικόλαο που βοήθησε, ώστε να ολοκληρωθεί η πτυχιακή εργασία αυτή. Τον ευχαριστούμε για την εμπιστοσύνη που μας έδειξε, την καλή θέληση για συνεργασία μεταξύ μας και την πολύτιμη βοήθεια και καθοδήγηση που μας παρείχε.

Κουμπή Ελευθερία

Μαλέσκου Παναγιώτα

Πάτρα, Νοέμβριος 2012

Περιεχόμενα

Πρόλογος

Η παρούσα εργασία έχει σαν στόχο τη μελέτη των μεθόδων της εξόρυξης δεδομένων (Data Mining). Οι τεχνικές της Εξόρυξης Δεδομένων έχουν εφαρμογές σε ένα μεγάλο όγκο δεδομένων ο οποίος είναι διαθέσιμος στο διαδίκτυο ή σε άλλες πηγές. Σκοπός της εξόρυξης δεδομένων είναι η εξεύρεση μιας χρήσιμης πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με τη χρήση αλγορίθμων, αρχών της στατιστικής, της τεχνητής νοημοσύνης και των συστημάτων βάσεων δεδομένων. Η χρήσιμη αυτή πληροφορία θα γίνεται αντιληπτή από τον άνθρωπο έτσι ώστε να τον βοηθήσει στη λήψη ορθών αποφάσεων.

Η εργασία είναι διαρθρωμένη σε τέσσερα κεφάλαια. Στο πρώτο κεφάλαιο γίνονται κατανοητές οι έννοιες με τις οποίες ασχολείται η εξόρυξη δεδομένων, αναλύονται οι τομείς εφαρμογής της καθώς επίσης και τα δεδομένα στα οποία εφαρμόζεται. Στη συνέχεια γίνεται λόγος στις τεχνικές εξόρυξης δεδομένων, οι οποίες είναι: η κατηγοριοποίηση, η συσταδοποίηση, οι κανόνες συσχέτισης και η παλινδρόμηση.

Στο δεύτερο κεφάλαιο γίνεται μια πιο ειδική αναφορά στις τεχνικές εξόρυξής δεδομένων ,όσο αφορά τις ιδιότητές τους , τα δεδομένα που εφαρμόζονται, τα οφέλη που προκύπτουν από αυτές, καθώς επίσης και τους σημαντικότερους αλγόριθμους που πραγματεύεται.

Στο τρίτο κεφάλαιο περιγράφονται κάποιες εφαρμογές τεχνικών της εξόρυξης δεδομένων σε προβλήματα του πραγματικού κόσμου. Οι εφαρμογές αυτές διακρίνονται σε ορισμένες βασικές φάσεις. Αποτελούν μια τυπική διαδικασία της εφαρμογής εξόρυξης γνώσης που έχει ως στόχο την επίλυση προβλημάτων σε οικονομικούς οργανισμούς.

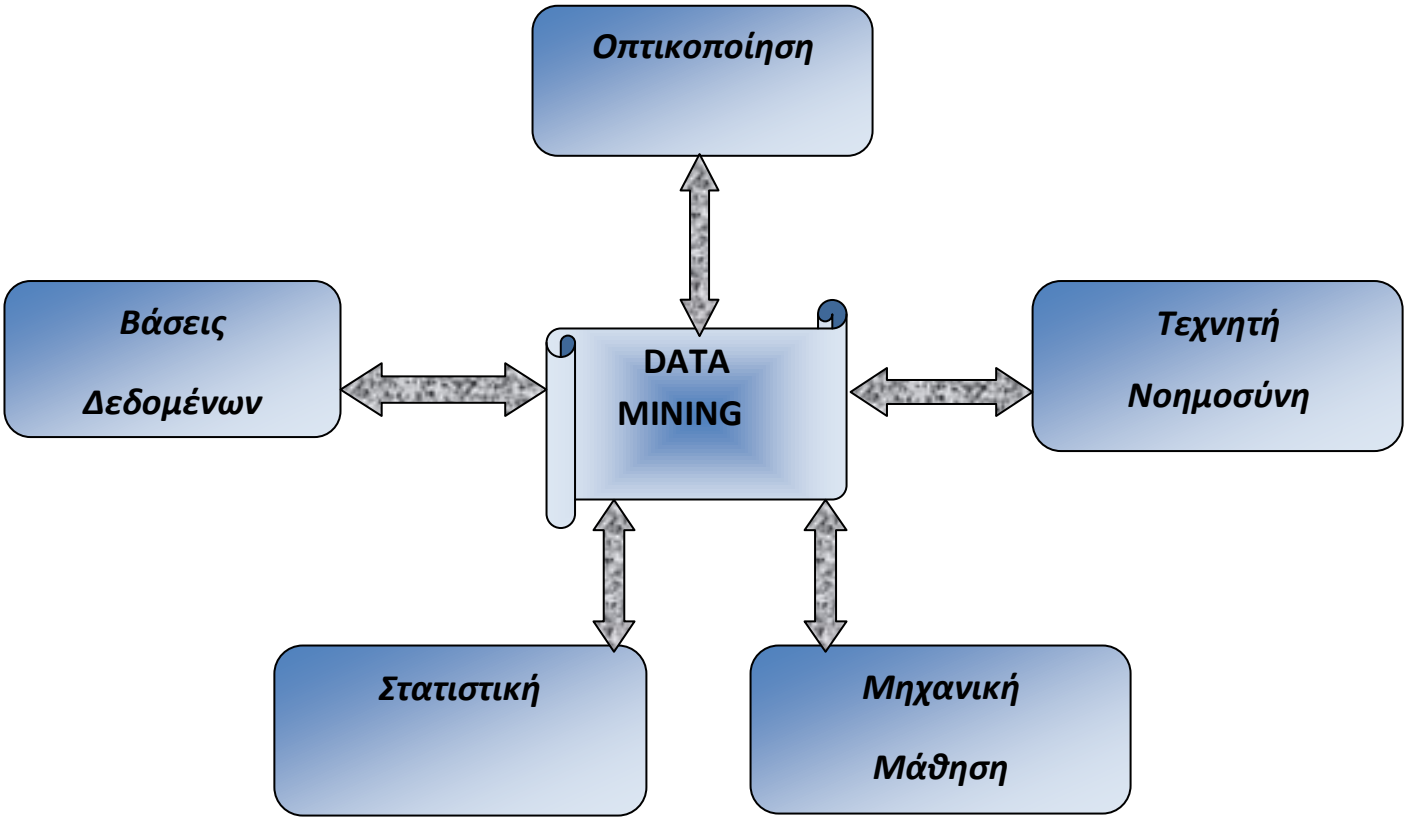
Τέλος , στο τέταρτο κεφάλαιο αναλύονται οι λειτουργίες και τα περιβάλλοντα του λογισμικού WEKA και εφαρμόζονται κάποιες βάσεις δεδομένων σε αυτό για την εξαγωγή πρόβλεψης.

1ο Κεφάλαιο

Γενικά για την Εξόρυξη Δεδομένων

1.1 Εισαγωγή

Ζούμε στην κοινωνία της πληροφορίας καθώς τεράστιος όγκος δεδομένων έχει αποθηκευτεί σε βάσεις δεδομένων. Το γεγονός αυτό οφείλεται στη διαρκή εξέλιξη της τεχνολογίας, και έτσι προβάλει επιτακτική η ανάγκη της μετατροπής των δεδομένων σε πληροφορία, πράγμα το οποίο αποτελεί προαπαιτούμενο βήμα για την μετατροπή της πληροφορίας σε γνώση. Η ανάγκη αυτή οδήγησε στην διαδικασία της Εξόρυξης Δεδομένων (Data Mining). Η εξόρυξη δεδομένων αποτελεί ένα μεμονωμένο και ταχέως αναπτυσσόμενο πεδίο. Αντλεί το όνομα της από την ομοιότητα που έχει η αναζήτηση πολύτιμων πληροφοριών σε μια μεγάλη βάση δεδομένων με την εξόρυξη πολύτιμων ορυκτών από μια ορεινή μάζα. Η Εξόρυξη Δεδομένων είναι ένα επιστημονικό πεδίο που συνδυάζει στοιχεία από τη Στατιστική, τη Μηχανική Μάθηση, τις Βάσεις Δεδομένων, την Οπτικοποίηση και την Τεχνητή Νοημοσύνη.



Εικόνα 1: Οι κυριότεροι τομείς αλληλεπίδρασης του Data Mining

1.2 Γενικές αρχές εξόρυξης δεδομένων

1.2.1 Τι είναι εξόρυξη δεδομένων

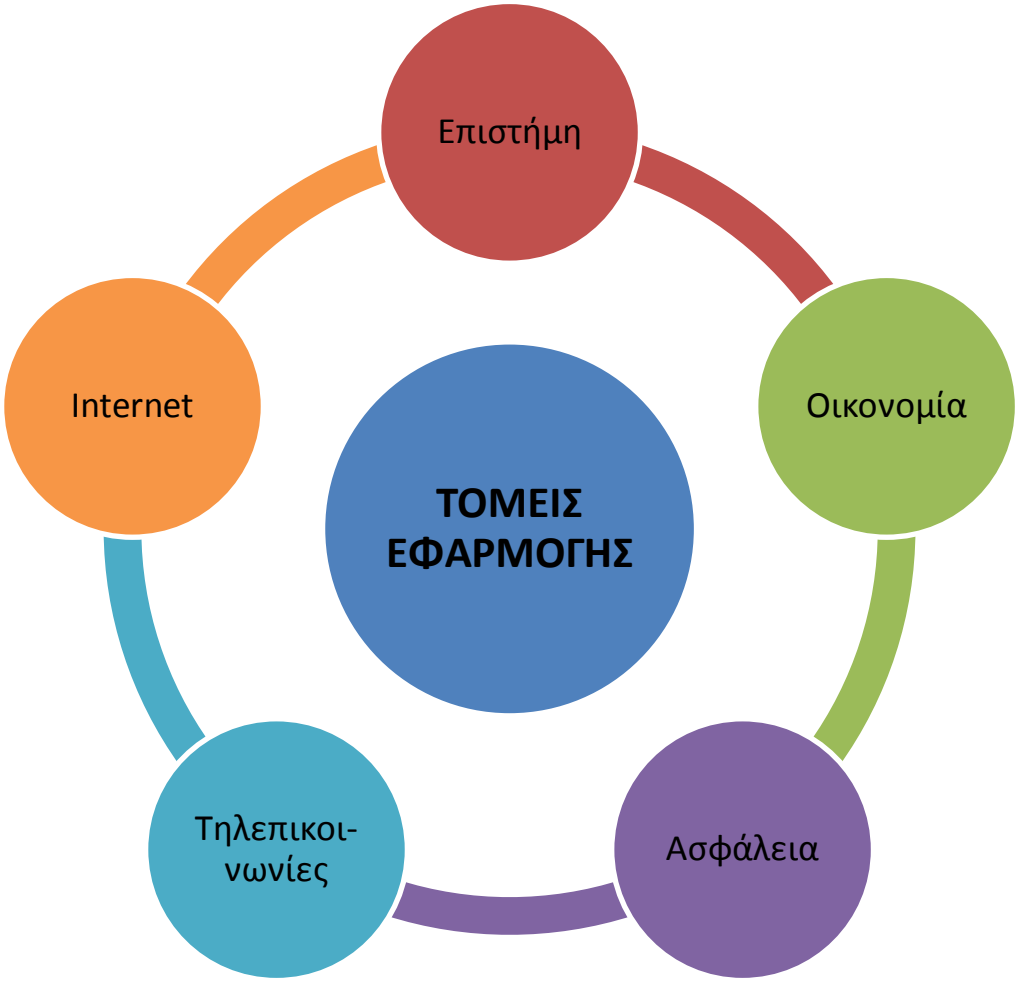
Γενικά, η Εξόρυξη Δεδομένων είναι η διαδικασία της ανάλυσης δεδομένων από διαφορετικές οπτικές γωνίες παρέχοντας τη δυνατότητα εξαγωγής χρήσιμης πληροφορίας. Πιο συγκεκριμένα, με τη χρήση κάποιων ορισμών γίνετε σαφέστερη η κατανόηση του όρου.

«**Εξόρυξη δεδομένων** (ή ανακάλυψη γνώσης από βάσεις δεδομένων) είναι η εξεύρεση μιας ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης, καθώς και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της εκμάθησης μηχανής και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι, η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν, να έχουν δομή κατανοητή από τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.»(Βικιπαίδεια, 2012)

«**Εξόρυξη δεδομένων** είναι η διαδικασία εξαγωγής υπονοούμενης και εν πολλοίς άγνωστης, αλλά ενδεχομένως χρήσιμης γνώσης, υπό την μορφή συσχετίσεων, προτύπων και τάσεων, μέσω της εξέτασης, ανάλυσης και επεξεργασίας βάσεων δεδομένων, συνδυάζοντας και χρησιμοποιώντας τεχνικές από την μηχανική μάθηση, την αναγνώριση προτύπων, τη στατιστική, τις βάσεις δεδομένων και την οπτικοποίηση».(Μαστρογιάννης Ν, 2009)

1.2.2 Τομείς εφαρμογής της Εξόρυξης Δεδομένων (Data Mining)

Η διαδικασία της εξόρυξης δεδομένων έχει πολλές εφαρμογές που έχουν αναπτυχθεί για μεγάλης κλίμακας προβλήματα του πραγματικού κόσμου σε διάφορους τομείς, όπως απεικονίζονται στο διάγραμμα που ακολουθεί:



Εικόνα 2: Τομείς εφαρμογής εξόρυξης δεδομένων

Επιστήμη

Τα τελευταία χρόνια, η εξόρυξη δεδομένων χρησιμοποιείται ευρέως σε διάφορους τομείς της επιστήμης, όπως: Ιατρική, Μηχανική, Αστρονομία, Βιολογία. Οι ερευνητές των τομέων αυτών συσσωρεύουν με ραγδαίο ρυθμό δεδομένα, τα οποία αποτελούν κλειδί για νέες ανακαλύψεις. Η εξόρυξη δεδομένων αποτελεί σημαντικό εργαλείο, αφού βοηθάει στην βελτίωση της διάγνωσης, της πρόληψης και της θεραπείας των ασθενειών, καθώς επίσης και στη σύγκριση της συμπεριφοράς χιλιάδων γονιδίων κάτω από διάφορες καταστάσεις. Σημαντικό επίτευγμα αποτελεί η δημιουργία ενός συστήματος το οποίο χρησιμοποιείται από τους αστρονόμους για ανάλυση εικόνας.

Χαρακτηριστικό παράδειγμα είναι αυτό της NASA που έχει στρατολογήσει δορυφόρους στην τροχιά της γης οι οποίοι παράγουν παρατηρήσεις για το σύνολο της επιφάνειας της γης, τους

ωκεανούς ,καθώς επίσης, και της ατμόσφαιρας. Με τη βοήθεια της εξόρυξης δεδομένων ανακαλύφθηκε ποιά είναι η σχέση μεταξύ της συχνότητας και της σφοδρότητας των διαταραχών του οικοσυστήματος (ξηρασίες, καταιγίδες και αύξηση θερμοκρασίας). Επίσης, με ποιόν τρόπο η βροχόπτωση και η θερμοκρασία της επιφάνειας του εδάφους επηρεάζονται από τη θερμοκρασία της επιφάνειας των ωκεανών και ,τέλος, πόσο καλά μπορούμε να προβλέψουμε την αρχή και το τέλος της περιόδου καλλιέργειας μιας περιοχής.

Ένα άλλο παράδειγμα εντοπίζεται στον τομέα της βιολογίας όπου οι επιστήμονες, βάση των νέων μεθόδων εξόρυξης δεδομένων κατάφεραν χρησιμοποιώντας μεγάλες ποσότητες γενωμικών δεδομένων να συγκρίνουν τη συμπεριφορά των γονιδίων, καθώς επίσης επετεύχθη και η πρόβλεψη πρωτεϊνικής δομής, η μοντελοποίηση βιοχημικών μονοπατιών και η φυλογενετική.

Οικονομία

Είναι σημαντικό να αναφέρουμε, πως η πληθώρα δεδομένων που υπάρχει στις επιχειρήσεις λόγω του μεγάλου αριθμού πελατών και των οικονομικών στοιχείων, δημιούργησε την ανάγκη για χρήση συστημάτων διαχείρισης δεδομένων, έτσι ώστε να βελτιστοποιηθεί η ανάλυση και η χρήση των δεδομένων αυτών. Στην περίπτωση των χρηματιστηριακών εφαρμογών η εξόρυξη δεδομένων γίνεται από κείμενα και τεχνικές αναφορές επιχειρήσεων, για την επίτευξη μιας πρόβλεψης της τάσης των μετοχών.(Βαρσάμη Ε, 2010)

Αξίζει να σημειωθεί πως η χρήση των τεχνικών εξόρυξης δεδομένων επέτρεψε στους εμπόρους λιανικής ,συλλέγοντας πληροφορίες και δεδομένα , να κατανοήσουν τις ανάγκες των πελατών. Όσο αφορά το marketing και τις επιχειρήσεις οι τεχνικές αυτές βοηθούν στην απάντηση βασικών ερωτημάτων όπως:

- Ποιοι πελάτες είναι οι πιο επικερδείς ;
- Ποιών προϊόντων οι πωλήσεις πρέπει να αναβαθμιστούν;
- Ποια είναι η πρόβλεψη για τα έσοδα της εταιρείας το επόμενο έτος ;

Ασφάλεια

Η εξόρυξη δεδομένων έχει παίξει σημαντικό ρόλο στην πρόληψη και την αποφυγή διαφόρων τύπων απάτης ,όπως, διαδικτυακές και οικονομικές απάτες. Η πρόληψη του δυσάρεστου

αυτού φαινομένου επιτυγχάνεται με την χρήση διαφόρων συστημάτων: ένα από τα συστήματα αυτά ονομάζεται FAIS.(Βαρσάμη Ε,2010)

Τηλεπικοινωνίες

Η ανάπτυξη της τηλεπικοινωνιακής βιομηχανίας και της τεχνολογίας (φαξ, κινητό τηλέφωνο, ηλεκτρονικό ταχυδρομείο) είναι ραγδαία. Η εξόρυξη δεδομένων βοηθάει στην καταπολέμηση παράνομων δραστηριοτήτων, στην αποδοτικότερη χρήση των πόρων και στην βελτίωση της ποιότητας των υπηρεσιών.

Τα τηλεπικοινωνιακά δεδομένα που συλλέγονται, περιλαμβάνουν τον τύπο κλήσης, την τοποθεσία του καλούντος και του κληθέντος, τον χρόνο κλήσης, την διάρκεια κλήσης κλπ. Η ανάλυση των δεδομένων αυτών μπορεί να δείξει διαγράμματα και γράφους των πόρων του συστήματος κάνοντας χρήση των εργαλείων οπτικοποίησης της εξόρυξης δεδομένων. Τέτοια εργαλεία είναι η συσχετισμένη οπτικοποίηση και η συσταδοποίηση. Επίσης , με τα εργαλεία της εξόρυξης δεδομένων είναι δυνατή η δημιουργία προφίλ των πελατών και ο εντοπισμός βλαβών στο δίκτυο.

Το κυρίως πρόβλημα που αντιμετωπίστηκε από την βιομηχανία τηλεπικοινωνιών είναι οι σκόπιμες κλήσεις κατά την ώρα αιχμής, περιοδικές κλήσεις κ.α. Μέθοδοι, όπως η συσταδοποίηση, συνεισφέρουν στην ανίχνευση παράνομων προτύπων βελτιώνοντας την αποτελεσματικότητα των υπηρεσιών τηλεπικοινωνίας.

Internet

Στο διαδίκτυο είναι διαθέσιμος μεγάλος όγκος δεδομένων και είναι αδύνατη η ακριβής μέτρησή τους. Παρόλα αυτά, οι μηχανές αναζήτησης αποτελούν παράδειγμα εξόρυξης δεδομένων, αφού παράγουν αποτελέσματα σε πολύ μικρό χρονικό διάστημα. Ένα χαρακτηριστικό παράδειγμα είναι και η Google η οποία παράγει αποτελέσματα η παρουσίαση των οποίων δεν ξεπερνά τα δύο δευτερόλεπτα.

1.2.3 Τι δεν είναι εξόρυξη δεδομένων

Η Εξόρυξη Δεδομένων (Data Mining) αποτελεί αναπόσπαστο κομμάτι της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (KDD Διαδικασία), η οποία θα αναλυθεί στην επόμενη παράγραφο. Στο σημείο αυτό, αξίζει να σημειωθεί η ύπαρξη της λέξης «γνώση», δημιουργώντας το ερώτημα γιατί να είναι «γνώση» και όχι «πληροφορίες» ή «δεδομένα»; Αυτό συμβαίνει επειδή υπάρχουν διαφορές μεταξύ των όρων «δεδομένα», «πληροφορίες» και «γνώση». Η ανάκτηση πληροφοριών ή δεδομένων από μεγάλες βάσεις δεδομένων, όσο περίπλοκή κι αν είναι η διαδικασία ανάκτησης πληροφοριών, δεν αποτελεί εξόρυξη δεδομένων. Αντίθετα, τα δεδομένα και οι πληροφορίες αυτές χρησιμοποιούνται για την εξαγωγή χρήσιμης γνώσης.

1.2.4 Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Databases- KDD)

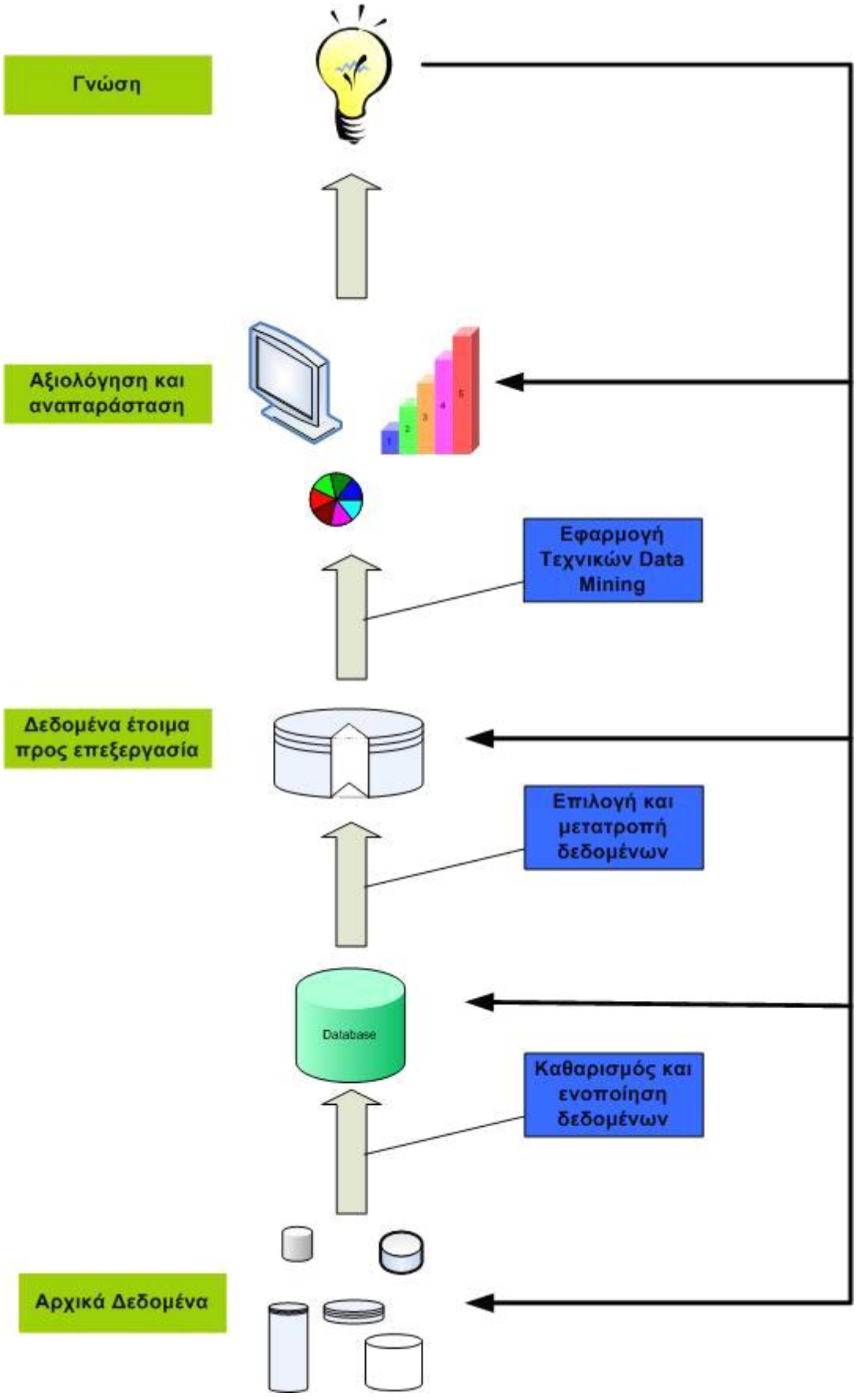
Όπως προαναφέρθηκε, η Εξόρυξη Δεδομένων αποτελεί βασικό βήμα της KDD Διαδικασίας . Μετά την επεξεργασία των δεδομένων, είναι πολύ πιθανό να ανακαλυφθεί «κρυμμένη γνώση», που σημαίνει πιθανές συσχετίσεις , αλληλεξαρτήσεις ή ομαδοποιήσεις μεταξύ αυτών. Όλα αυτά υποστηρίζονται με την εφαρμογή αλγορίθμων.

Πιο συγκεκριμένα, η KDD διαδικασία είναι: «η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων ,ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα»(Μ. Χαλκίδη- Μ. Βαζιριάννης, 2005)

Η ανακάλυψη γνώσης σαν διαδικασία αποτελείται από μια επαναληπτική ακολουθία βημάτων:

1. ***Καθαρισμός δεδομένων (Data Cleaning):*** όπου απομακρύνεται ο θόρυβος και τα ακατάλληλα δεδομένα.
2. ***Ενοποίηση δεδομένων (Data Integration):*** συνδυασμός πολλαπλών πηγών δεδομένων.
3. ***Επιλογή Δεδομένων(Data Selection):*** επιλογή και ανάκτηση των δεδομένων σχετικά με τη διαδικασία της ανάλυσης.
4. ***Μετατροπή Δεδομένων (Data Transformation):*** μετατροπή των δεδομένων σε κατάλληλη μορφή προς επεξεργασία.

5. ***Εξόρυξη Δεδομένων (Data Mining):***εφαρμογή ευφυών μεθόδων ψάχνοντας για ενδιαφέροντα πρότυπα γνώσης.
6. ***Αξιολόγηση Προτύπων(Evaluation of Standards):*** αξιολόγηση εξαγόμενων προτύπων για να προσδιοριστούν τα αληθινά ενδιαφέροντα πρότυπα.
7. ***Αναπαράσταση Γνώσης (knowledge presentation):*** εφαρμογή τεχνικών οπτικοποίηση και αναπαράστασης γνώσης με σκοπό την παρουσίαση της εξορυγμένης γνώσης στο χρήστη.

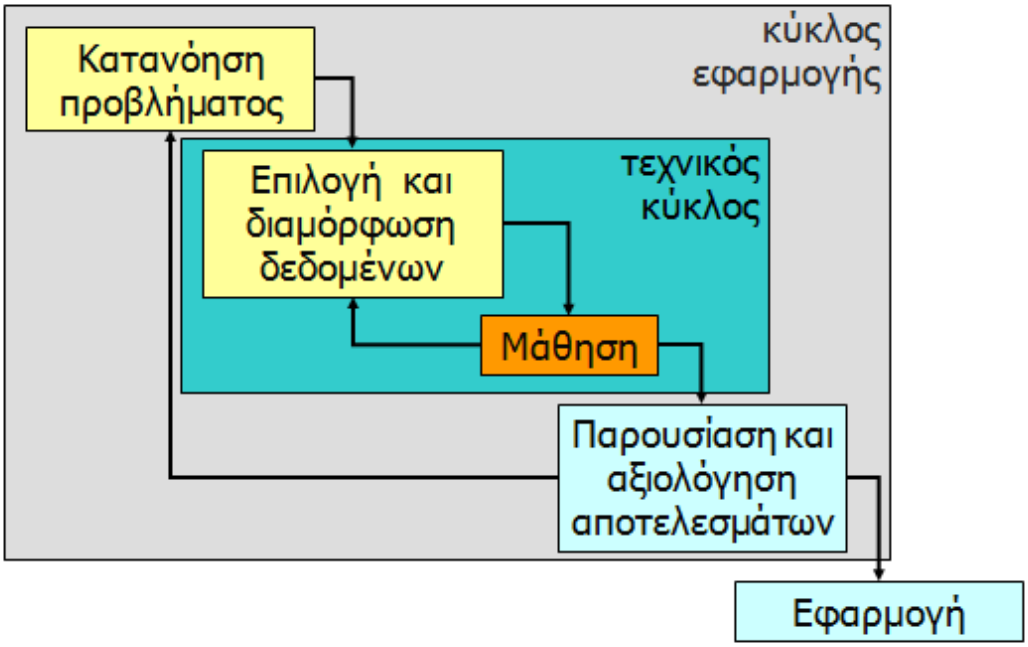


Εικόνα 3:Η Διαδικασία Ανακάλυψης Γνώσης (KDD)

Τα βήματα που απεικονίζονται στην εικόνα 3 είναι διαδοχικά, αλλά ο εκάστοτε χρήστης μπορεί να επανέλθει σε οποιοδήποτε βήμα, εάν δεν είναι ευχαριστημένος από το αποτέλεσμα ή μπορεί να ξεκινήσει τη διαδικασία από οποιοδήποτε ενδιάμεσο βήμα. Όπως αναφέρθηκε και παραπάνω, η εξόρυξη δεδομένων αποτελεί βασικό βήμα της διαδικασίας εξόρυξης γνώσης, αλλά αυτό δεν σημαίνει ότι τα υπόλοιπα βήματα δεν είναι εξίσου σημαντικά για την επιτυχή εφαρμογή της KDD διαδικασίας.

1.2.4 Βασικά βήματα διαδικασίας εξόρυξης γνώσης από βάσεις δεδομένων

Παρά το γεγονός ότι η εξόρυξη γνώσης (data mining) αποτελεί βασικό βήμα της Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (KDD Διαδικασία), η ίδια η εξόρυξη γνώσης αποτελείται κι αυτή από κάποια βασική διαδικασία. Αρχικά, γίνεται η κατανόηση του προβλήματος, το οποίο αποτελεί και το βασικότερο στοιχείο για την καλή πορεία της διαδικασίας. Στη συνέχεια, επιλέγονται και διαμορφώνονται τα δεδομένα, όπου απομακρύνονται οι θόρυβοι και τα ασυνεπή στοιχεία, έτσι ώστε τα δεδομένα να είναι κατάλληλα για την εξαγωγή χρήσιμης πληροφορίας. Επιπλέον, παρουσιάζονται και αξιολογούνται τα αποτελέσματα, έτσι ώστε να καταλήξουμε στην επιλογή και εφαρμογή της κατάλληλης τεχνικής.



Εικόνα 4: Διαδικασία εξόρυξης γνώσης

1.2.5 Προκλήσεις: Κίνητρα για ανάπτυξη της εξόρυξης γνώσης

Με την πάροδο του χρόνου , οι παραδοσιακές μέθοδοι ανάλυσης δεδομένων αντιμετώπισαν δυσκολίες οι οποίες αποτέλεσαν κίνητρο για την ανάπτυξη της εξόρυξης των δεδομένων. Κάποιες από τις προκλήσεις αυτές αναλύονται παρακάτω:

- **Κλιμάκωση (Scalability)**. Πλέον, τα σύνολα δεδομένων που χρησιμοποιούνται, είναι πολλές φορές μεγέθους gigabyte, terabyte και peta–byte. Για το λόγο αυτό, οι αλγόριθμοι πρέπει να είναι κλιμακωτοί, έτσι ώστε να μπορούν να προσαρμοστούν και να διαχειριστούν δεδομένα τέτοιου όγκου. Χαρακτηριστικό παράδειγμα αποτελούν οι αλγόριθμοι εκτός πυρήνα (out - of - core), οι οποίοι επεξεργάζονται σύνολα δεδομένων τα οποία δεν χωρούν στην κύρια μνήμη. Τέλος , η βελτίωση της κλιμάκωσης μπορεί να γίνει χρησιμοποιώντας δειγματοληψία ή αναπτύσσοντας παράλληλους και κατανεμημένους αλγόριθμους.
- **Πολλές Διαστάσεις (High Dimensionality).** Μερικές δεκαετίες πριν συναντούσαμε σύνολα δεδομένων με λίγα χαρακτηριστικά , σε αντίθεση με τη σημερινή εποχή, όπου τα χαρακτηριστικά αυτά τείνουν να γίνουν εκατοντάδες ή χιλιάδες . Το πρόβλημα αυτό αντιμετώπισαν πολλοί κλάδοι , όπως για παράδειγμα, η βιοπληροφορική, όπου

τα γονίδια περιέχουν χιλιάδες χαρακτηριστικά. Έτσι, συμπεραίνουμε, πως οι παραδοσιακές τεχνικές ανάλυσης δεν είναι πλέον κατάλληλες.

- **Ετερογενή και Πολύπλοκα Δεδομένα (Heterogeneous and Complex data)**. Τα δεδομένα που διαχειρίζονται οι παραδοσιακές μέθοδοι ανάλυσης περιέχουν χαρακτηριστικά ίδιου τύπου, είτε συνεχή, είτε κατηγορικά. Με το πέρας των ετών, η εξόρυξη δεδομένων έχει γίνει πολύ σημαντική για διάφορα πεδία, όπως , η επιστήμη και οι επιχειρήσεις, που περιέχουν σύνολο δεδομένων με ετερογενή χαρακτηριστικά, δημιουργώντας έτσι την ανάγκη ανάπτυξης νέων τεχνικών κατάλληλων για τα δεδομένα αυτά.
- **Κυριότητα και Διανομή δεδομένων (Data Ownership and Distribution).** Πολλά δεδομένα δεν είναι αποθηκευμένα σε μια μόνο θέση ή δεν αποτελούν ιδιοκτησία κάποιου οργανισμού. Για το λόγο αυτό, απαιτείται η ανάπτυξη νέων τεχνικών εξόρυξης δεδομένων που θα αντιμετωπίσουν το συγκεκριμένο πρόβλημα.
- **Μη Παραδοσιακή Ανάλυση (Non – Traditional Analysis).** Στην παραδοσιακή στατιστική προσέγγιση προτείνεται μια υπόθεση, σχεδιάζεται ένα πείραμα για τη συλλογή δεδομένων και στη συνέχεια, τα δεδομένα αναλύονται σε σχέση με την υπόθεση. Η διαδικασία που μόλις περιγράψαμε είναι ιδιαίτερα χρονοβόρα, αφού στις σύγχρονες αναλύσεις δεδομένων θα πρέπει να γίνεται δημιουργία και αξιολόγηση χιλιάδων υποθέσεων . (Steinbach Vipin Kumar,2007)

1.3 Μέθοδοι και Τεχνικές Ανακάλυψης Γνώσης

Η ***πρόβλεψη*** (prediction) περιλαμβάνει τη χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (predictive data mining tasks), προσπαθούν να κάνουν εκτιμήσεις, βγάζοντας συμπεράσματα από διαθέσιμα δεδομένα.

Η ***περιγραφή*** (description) επικεντρώνεται στην ανακάλυψη προτύπων κι αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων, με όσο το δυνατόν πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων

(descriptive data mining tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων. (Μαστρογιάννης Ν,2009)

Οι βασικότερες τεχνικές που χρησιμοποιεί η εξόρυξη δεδομένων περιγράφονται πιο αναλυτικά παρακάτω.

1.3.1 Μέθοδοι Πρόβλεψης- Περιγραφικοί Μέθοδοι

- ***Κατηγοριοποίηση (classification).*** Αποτελεί μια από τις βασικότερες εργασίες (tasks) εξόρυξης δεδομένων. Γίνεται εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένου), το οποίο, με βάση τα χαρακτηριστικά αυτά, αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων (classes). Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαριστάνονται γενικά από τις εγγραφές της βάσης δεδομένων κι έτσι, η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κατηγορίες. Οι βασικές μέθοδοι που χρησιμοποιούνται είναι: Μέθοδος Bayers, Δέντρα Αποφάσεων, Αλγόριθμοι διανυσμάτων Υποστήριξης (Support Vector Machines), Νευρωνικά Δίκτυα.

- ***Συσταδοποίηση (clustering).*** Είναι η εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων (clusters). Στην κατηγοριοποίηση, ο πληθυσμός διαιρείται σε κατηγορίες, αναθέτοντας κάθε στοιχείο ή εγγραφή σε μια προκαθορισμένη κατηγορία, με βάση ένα μοντέλο που αναπτύσσεται μέσω της εκπαίδευσής του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων. Αντιθέτως, στη συσταδοποίηση, οι εγγραφές ομαδοποιούνται σε σύνολα, με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Αξίζει να σημειωθεί πως μπορεί να χρησιμοποιηθεί και σαν εισαγωγή σε κάποια άλλη διαδικασία εξόρυξης γνώσης ή μοντελοποίησης. Αντιπροσωπευτικοί αλγόριθμοι: K-Means και παραλλαγές, PAM, DBSCAN, COBWEB.

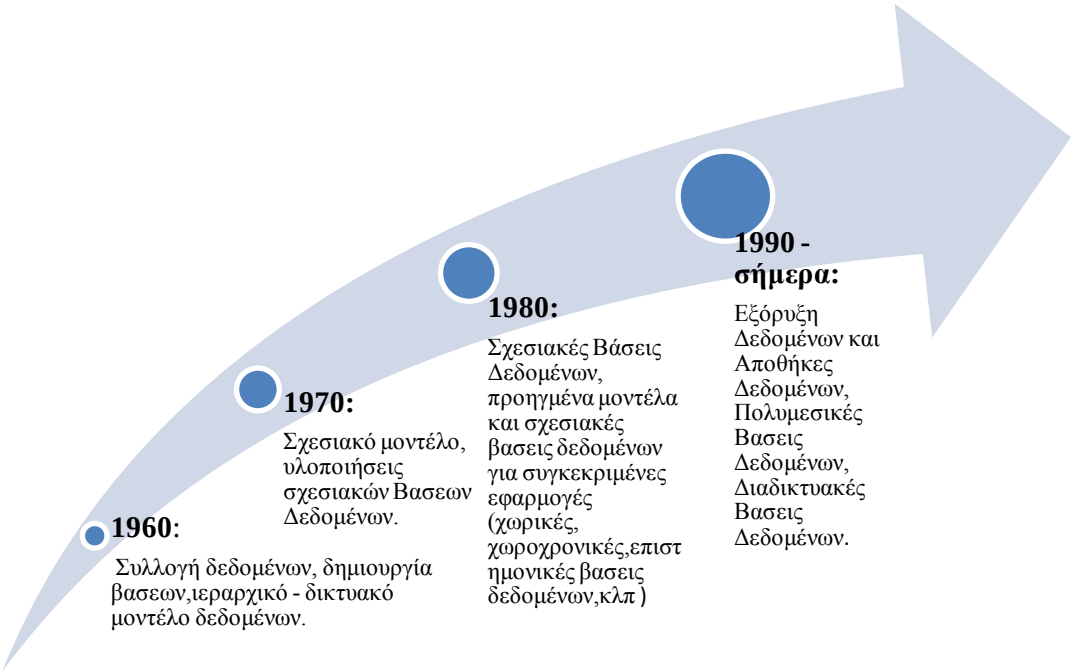
- ***Κανόνες Συσχέτισης (association rules).*** Θεωρείται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων, καθώς παρέχουν ένα συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές

από τους τελικούς χρήστες. Πιο συγκεκριμένα, οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Οι συσχετισμοί αυτοί έχουν τη μορφή « If *A* then *B*», όπου το *A* και το *B* αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα. Τέλος, η ανάλυση συσχέτισης είναι γνωστή στον επιχειρηματικό κόσμο και ως *ανάλυση συνάφειας* (affinity analysis) με πολλές εφαρμογές. (Berry and Linoff, 2004)

- ***Παλινδρόμηση (regression).*** Είναι η παλαιότερη και η πιο γνωστή στατιστική τεχνική που υλοποιείται εντός των πλαισίων της εξόρυξης δεδομένων. Χρησιμοποιώντας μια βάση δεδομένων, αναπτύσσεται μια μαθηματική σχέση που ταιριάζει στα δεδομένα αυτά, η οποία βοηθά στην πρόβλεψη μελλοντικής συμπεριφοράς, εφαρμόζοντας σε αυτή νέα αριθμητικά δεδομένα. Η συγκεκριμένη τεχνική εφαρμόζεται καλά μόνο σε συνεχή ποσοτικά δεδομένα (ζήτηση, δαπάνες διαφήμισης), ενώ δεν λειτουργεί καλά με κατηγορικά δεδομένα (Μαστρογιάννης Ν,2009).

1.4 Ιστορική Αναδρομή: Βάσεις δεδομένων & Εξόρυξη γνώσης

Οι Βάσεις Δεδομένων έκαναν την εμφάνισή τους το 1960, ενώ η Εξόρυξη Δεδομένων εμφανίστηκε το 1990.



Εικόνα 5: Ιστορική Εξέλιξη

1.4.1 Οι 10 καλύτεροι Αλγόριθμοι Εξόρυξης Δεδομένων (Data Mining)

1. C4.5 (61 votes)- Ταξινόμηση (δέντρο απόφασης)
2. K-Means (60 votes)- Συσταδοποίηση
3. SVM (58 votes)- Ταξινόμηση
4. Apriori (52 votes)- Κανόνες συσχέτισης
5. EM (48 votes)- Στατιστική, Συσταδοποίηση
6. PageRank (46 votes)- Ιστοσελίδες
7. AdaBoost (45 votes)- Μετα-Ταξινομητής
8. KNN (45 votes)- Συσταδοποίηση
9. Naïve Bayers (45 votes)- Στατιστική , Ταξινόμηση
10. CART (34 votes)- Ταξινόμηση (δέντρο απόφασης)

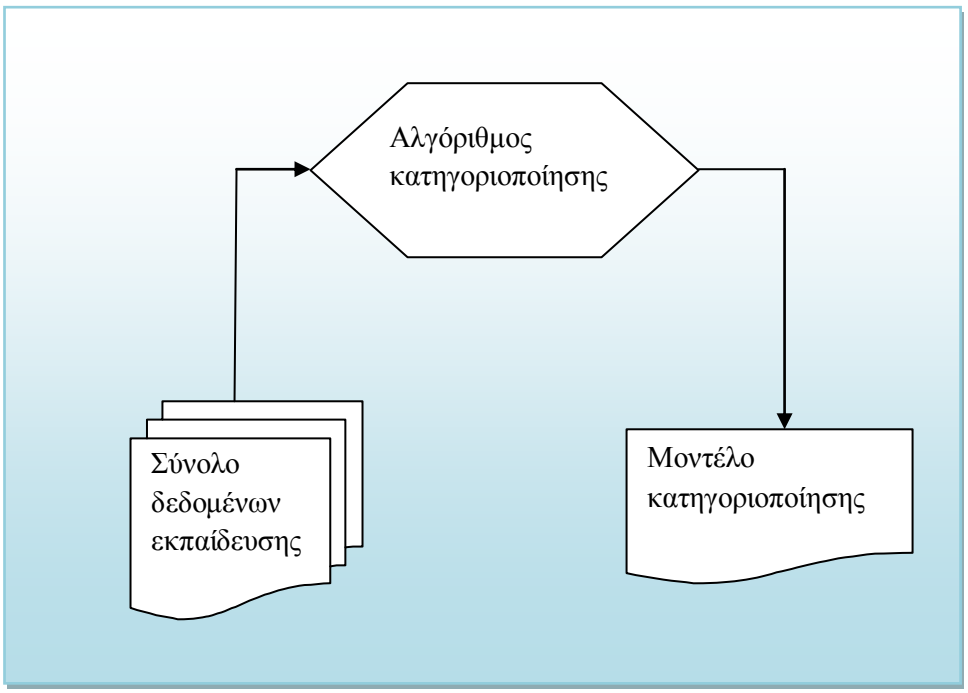
2ο Κεφάλαιο

Κατηγοριοποίηση

2.1 Εισαγωγή

Είναι γνωστό, πως η κατηγοριοποίηση είναι ένα από τα βασικά βήματα στη διαδικασία εξόρυξης γνώσης, όπου στο βήμα αυτό γίνεται η αντιστοιχία ενός στοιχείου με ένα καθορισμένο σύνολο κατηγοριών. Βασικός στόχος της είναι η δημιουργία ενός μοντέλου, το οποίο θα είναι δυνατόν να χρησιμοποιηθεί προκειμένου να κατηγοριοποιηθούν μελλοντικά δεδομένα τα οποία δεν έχουν κατηγοριοποιηθεί ακόμη. Η κατηγοριοποίηση επιτυγχάνεται έπειτα από το πέρας δυο βημάτων:

- Βήμα 1ο: Εκμάθηση (Learning). Το βήμα αυτό αποτελεί την τμηματική δημιουργία ενός μοντέλου, που περιγράφει ένα προκαθορισμένο σύνολο από κατηγορίες δεδομένων. Αρχικά, με βάση έναν αλγόριθμο κατηγοριοποίησης τα δεδομένα εκπαίδευσης (training data) αναλύονται έτσι ώστε να κατασκευαστεί εν συνεχεία το μοντέλο. Κατόπιν τα στοιχεία επιλέγονται τυχαία και κατατάσσονται σε μια από τις προκαθορισμένες κατηγορίες. Το βήμα που μόλις περιγράψαμε είναι γνωστό και ως «εποπτευμένη μάθηση» (supervised learning), ενώ το μοντέλο, ως κατηγοριοποιητής (classifier).
- Βήμα 2ο : Κατηγοριοποίηση (Classification). Το βήμα αυτό αποτελεί μια δοκιμή , ώστε να διαπιστωθεί κατά πόσο η ακρίβεια του μοντέλου είναι αποδεκτή. Εφόσον αποδειχθεί ότι είναι αποδεκτή, το μοντέλο μπορεί πλέον να κατηγοριοποιήσει μελλοντικά δείγματα δεδομένων. Για την εκτίμηση της ακρίβειας του κατηγοριοποιητή επιλέγονται τυχαία τα δεδομένα εκπαίδευσης . Στη συνέχεια, το μοντέλο κατηγοριοποιεί ένα από τα δοκιμαστικά παραδείγματα (training samples). Τέλος , συγκρίνεται η κατηγορία στην οποία ανήκουν τα δεδομένα με την πρόβλεψη την οποία έκανε το μοντέλο για την κατηγορία αυτή.



Εικόνα 6: Διαδικασία κατηγοριοποίησης (Εκμάθηση)

2.2 Bayesian κατηγοριοποίηση

Η Bayesian κατηγοριοποίηση έχει ως σκοπό να κατηγοριοποιεί ένα δείγμα X σε μια από τις προκαθορισμένες κατηγορίες C_1, C_2, .., C_n . Η απόδοση αυτού του είδους κατηγοριοποίησης είναι αρκετά υψηλή και χαρακτηρίζεται από την μεγάλη ταχύτητα της διαδικασίας κατηγοριοποίησης σε μεγάλες Βάσεις Δεδομένων. Κάθε κατηγορία χαρακτηρίζεται από μια εκ των προτέρων πιθανότητα (a prior probability) παρατήρησης της κλάσης C_i. Επίσης, υποθέτουμε ότι το δεδομένο δείγμα X ανήκει σε μια κλάση C_i με την υπό συνθήκη συνάρτηση πυκνότητας: p (X/Ci) ∈ [0,1]. Επιπρόσθετα, χρησιμοποιώντας τα παραπάνω και στηριζόμενοι στη θεωρεία Bayes, καθορίζουμε την εκ των υστέρων πιθανότητα p (ci/x). Ο τύπος δίνεται παρακάτω:

$$p(c_i|X)= p(c_i|X)p(c_i)/p(X)$$

2.2.1 Bayesian Κατηγοριοποιητές

Υπάρχουν δύο ειδών κατηγοριοποιητές κατά Bayes, ο Naïve Bayesian κατηγοριοποιητής και τα Bayesian Belief Networks.

- Naïve Bayesian: είναι γνωστός ως ο απλούστερος κατηγοριοποιητής που υποθέτει πως η επίδραση ενός γνωρίσματος (attribute) σε μια δεδομένη κατηγορία. είναι ανεξάρτητη από τις τιμές των άλλων γνωρισμάτων. Η παραπάνω υπόθεση γίνεται για να απλοποιήσει τους υπολογισμούς που εμπλέκονται και είναι γνωστή ως «υπό συνθήκη ανεξαρτησία» (conditional independence) κατηγορίας.
- Bayesian Belief Networks: πρόκειται για γραφικά μοντέλα που προσδιορίζουν τις συνδεδεμένες υπό συνθήκη κατανομές πιθανότητας, έχοντας ως στόχο, να λάβουν υπόψη τις εξαρτήσεις που μπορούν να υπάρξουν μεταξύ των μεταβλητών.

2.2.2 Πλεονεκτήματα

- Εύκολη χρήση.
- Απαιτείται μόνο ένα πέρασμα των δεδομένων εκπαίδευσης.
- Η προσέγγιση αυτή μπορεί εύκολα να χειριστεί ελλιπή δεδομένα, απλά παραλείποντας τις αντίστοιχες πιθανότητες.
- Σε περιπτώσεις όπου υπάρχουν απλές συσχετίσεις στα δεδομένα, η τεχνική συνήθως δίνει καλά αποτελέσματα κατηγοριοποίησης, σε σύντομο χρονικό διάστημα.

2.2.3 Μειονεκτήματα

- Σπάνιες είναι οι περιπτώσεις όπου τα χαρακτηριστικά δεν είναι ανεξάρτητα. Μια προσέγγιση είναι να αγνοήσουμε τα χαρακτηριστικά τα οποία εξαρτώνται από άλλα.
- Επιπρόσθετα, η τεχνική αυτή δεν μπορεί να χειριστεί συνεχή δεδομένα. Το μειονέκτημα αυτό λύνεται, με το να χωρίσουμε τα συνεχή χαρακτηριστικά σε διαστήματα, ωστόσο αυτό δεν είναι κάτι απλό και ο τρόπος με το οποίον θα γίνει, είναι πολύ πιθανό να επηρεάσει τα αποτελέσματα.

2.3 Δέντρα απόφασης

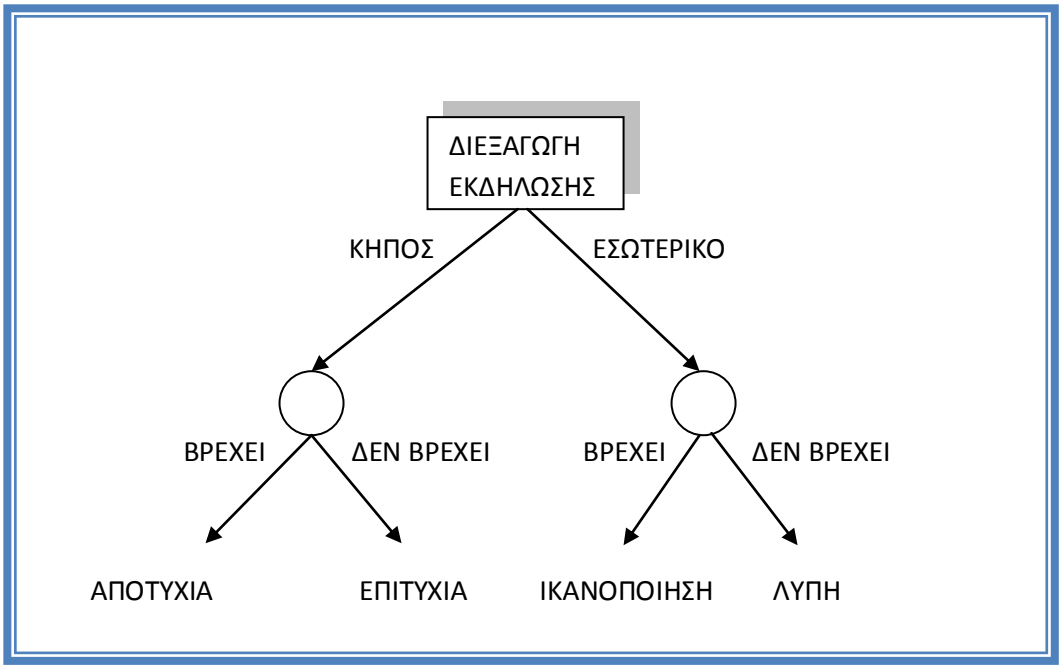
Μια ευρέως χρησιμοποιούμενη τεχνική κατηγοριοποίησης αποτελούν τα δέντρα απόφασης. Η δημιουργία του δέντρου απόφασης γίνεται με βάση ένα σύνολο εκπαίδευσης προ-κατηγοριοποιημένων δεδομένων. Σε κάθε εσωτερικό κόμβο του δέντρου απόφασης υπάρχει ο έλεγχος ενός γνωρίσματος. Κάθε κλαδί που επεκτείνεται από τον κόμβο αυτό αντιστοιχεί σε μια από τις πιθανές τιμές για το συγκεκριμένο γνώρισμα. Τέλος , κάθε φύλλο αντιστοιχεί σε μια από τις κατηγορίες που έχουν οριστεί.

2.3.1 Διαδικασία με βάση ένα δέντρο απόφασης

Η διαδικασία για την κατηγοριοποίηση ενός νέου δείγματος με βάση ένα δέντρο απόφασης είναι η εξής :

1. Έχοντας ως αφετηρία τη ρίζα του δέντρου όπου γίνεται και ο έλεγχος των γνωρισμάτων .
2. Έπειτα από τον έλεγχο γίνεται ο καθορισμός των εσωτερικών κόμβων που θα ακολουθήσουν στη συνέχεια.
3. Η διαδικασία αυτή επαναλαμβάνεται όσες φορές χρειαστεί, έως ότου καταλήξουμε στον τελικό κόμβο (φύλλο του δέντρου)

Αξιοσημείωτο είναι ότι, η κατηγορία του τελικού κόμβου είναι η κατηγορία του υπό μελέτη δείγματος .Ένα απλό παράδειγμα για την κατανόηση της παραπάνω διαδικασίας είναι το ακόλουθο:



Εικόνα 7: Παράδειγμα δέντρου απόφασης

Οι περισσότεροι από τους αλγόριθμους έχουν δύο διακριτές φάσεις, τη φάση οικοδόμησης (building phase) και τη φάση περικοπής (pruning phase). Στη φάση οικοδόμησης, τα δεδομένα χωρίζονται κατ' επανάληψη, μέχρις ότου όλα τα δείγματα σε ένα τμήμα (partition) να ανήκουν στην ίδια κατηγορία, με αποτέλεσμα τη δημιουργία ενός δέντρου, το οποίο κατηγοριοποιεί κάθε στοιχείο του συνόλου εκπαίδευσης. Όμως, οι περισσότεροι αλγόριθμοι εκτελούν μια φάση περικοπής μετά από τη φάση κατασκευής του δέντρου, η οποία δημιουργεί ένα δέντρο με υψηλότερη ακρίβεια.

2.3.2 Πλεονεκτήματα

- Μη παραμετρική προσέγγιση: Δε στηρίζεται σε υπόθεση εκ των προτέρων γνώσης σχετικά με τον τύπο της κατανομής πιθανότητας, που ικανοποιεί η κλάση ή τα άλλα γνωρίσματα.
- Η κατασκευή του βέλτιστου δέντρου απόφασης είναι ένα NP-complete πρόβλημα.
- Αφού το δέντρο κατασκευαστεί, η ταξινόμηση νέων εγγραφών είναι πολύ γρήγορη.
- Εύκολα στην κατανόηση (ιδιαίτερα τα μικρά δέντρα).

2.3.3 Μειονεκτήματα

- Δεν μπορεί να χειριστεί περίπλοκες σχέσεις μεταξύ των γνωρισμάτων.
- Απλά όρια απόφασης (decision boundaries).
- Προβλήματα όταν λείπουν πολλά δεδομένα.

2.3.4 Βασικοί αλγόριθμοι κατασκευής δέντρων απόφασης

Ορισμένοι από τους πιο γνωστούς αλγόριθμους κατασκευής δέντρων απόφασης είναι: ID3, SLIQ, C4.5, SPRINT, οι οποίοι θα αναλυθούν παρακάτω.

Αλγόριθμος ID3

Μια από τις πιο διαδεδομένες και ταυτόχρονα απλές τεχνικές που χρησιμοποιείται για την κατασκευή δέντρων απόφασης, είναι ο αλγόριθμος ID3. Αυτό που προσπαθεί να επιτύχει αυτός ο αλγόριθμος είναι να ελαχιστοποιήσει τον αριθμό των συγκρίσεων. Η βασική ιδέα ενός αλγορίθμου επαγωγής είναι να κάνει ερωτήσεις, των οποίων οι απαντήσεις να περιέχουν τις περισσότερες πληροφορίες. Λέγοντας περισσότερες πληροφορίες, εννοούμε ερωτήσεις που απορρίπτουν μεγάλο μέρος του χώρου αναζήτησης. Η βασική ιδέα του αλγορίθμου ID3 είναι η επιλογή χαρακτηριστικών διάσπασης, που περιέχουν μεγαλύτερο κέρδος πληροφορίας. Το ποσό της πληροφορίας, το οποίο σχετίζεται με την τιμή ενός χαρακτηριστικού, εξαρτάται από την πιθανότητα εμφάνισης του.

Χρησιμοποιούμε το μέτρο της εντροπίας, ώστε να μετρήσουμε το πόσο ανομοιογενές είναι ένα σύνολο δεδομένων. Το μέτρο αυτό παίρνει τιμές στο διάστημα [0,1]. Ο ορισμός της εντροπίας είναι:

Με δεδομένες τις πιθανότητες $p_1, p_2, \ldots, p_s$ με

$$\sum_{i=1}^{s} p_i = 1$$

Η εντροπία ορίζεται ως:

$$H(p_1, p_2, \ldots, p_s) = \sum_{i=1}^{s} (p_i \log(\frac{1}{p_i}))$$

Δεδομένης μια κατάστασης *D*, το *H(D)* , βρίσκει την ποσότητα της τάξης σε αυτή την κατάσταση. Όταν η κατάσταση *D*, χωρίζεται σε *s* νέες καταστάσεις ($S = \{D_1, D_2, \ldots, D_s\}$), το μέτρο της εντροπίας μπορεί να εφαρμοστεί σε κάθε μια από αυτές τις νέες καταστάσεις. Κάθε βήμα του ID3 επιλέγει την κατάσταση, η οποία διατάσσει περισσότερο τη διάσπαση. Μια κατάσταση της Βάσης Δεδομένων είναι απολύτως διατεταγμένη, αν όλες οι πλειάδες σε αυτήν ανήκουν στην ίδια κατηγορία. Ο ID3 επιλέγει το χαρακτηριστικό διάσπασης με το μεγαλύτερο κέρδος πληροφορίας. Το ***Κέρδος (gain)*** πληροφορίας μετρά την μείωση της εντροπίας που θα προκληθεί, αν χωριστεί το σύνολο δεδομένων με βάση κάποιο χαρακτηριστικό. Καταλήγοντας, ο ID3 αλγόριθμος υπολογίζει το κέρδος μιας διάσπασης χρησιμοποιώντας τον εξής τύπο:

$$Gain\ (D, S) = H(D) - \sum_{i=1}^{s} P(D_i) H(D_i)$$

Διαδικασία:

Βήμα 1ο: Αρχικά, πρέπει να επιλεγεί το πιο κατάλληλο χαρακτηριστικό για έλεγχο στη ρίζα.

Βήμα 2ο: Στη συνέχεια, για κάθε δυνατή τιμή του χαρακτηριστικού δημιουργούνται οι αντίστοιχοι απόγονοι της ρίζας. Τα δεδομένα μοιράζονται στους νέους κόμβους, ανάλογα με την τιμή που έχουν για το χαρακτηριστικό που ελέγχεται στη ρίζα.

Βήμα 3ο: Η όλη διαδικασία επαναλαμβάνεται για κάθε νέο κόμβο. Η επιλογή του χαρακτηριστικού θα γίνει βάσει των δεδομένων που ανήκουν στον κάθε κόμβο.

Βήμα 4ο: Ένας κόμβος γίνεται φύλλο, όταν όλα τα δεδομένα που ανήκουν σε αυτόν ανήκουν στην ίδια κατηγορία. Η κατηγορία αυτή γίνεται και η τιμή του φύλλου.

Βήμα 5ο: Αν σε κάποιο σημείο τελειώσουν τα χαρακτηριστικά προς έλεγχο, τότε ο κόμβος γίνεται τερματικός και σαν τιμή παίρνει εκείνη που έχει την πλειοψηφία με βάση τα δεδομένα του κόμβου αυτού.

Ο αλγόριθμος ID3 ο οποίος θεωρείται ένας από τους βασικούς αλγόριθμους κατηγοριοποίησης, προσφέρει κάποια πλεονεκτήματα και κάποια μειονεκτήματα, ορισμένα από τα οποία είναι:

+ Ένα δένδρο απόφασης μπορεί εύκολα να αναπαρασταθεί και σαν ένα σύνολο κανόνων.

+ Είναι εξαιρετικά αποδοτικός.

– Απαιτεί από την αρχή το σύνολο των δεδομένων εκπαίδευσης, καθώς η λειτουργία του βασίζεται σε συγκεντρωτικά μεγέθη αυτού του συνόλου.
– Είναι ισχυρά εξαρτώμενος από τον μηχανισμό διαχωρισμού που θα επιλεγεί.

Αλγόριθμος SLIQ

Αρχικά, στον αλγόριθμο αυτό εφαρμόζεται προ-κατηγοριοποίηση των γνωρισμάτων. Ορίζεται ο κόμβος-ρίζα του δέντρου, ενώ, κατά τη διάρκεια του διαχωρισμού χρησιμοποιείται η λίστα κατηγοριών, ώστε να γίνει ο βέλτιστος διαχωρισμός .Η ενημέρωση των ετικετών των φύλλων γίνεται κάθε φορά από τον αντίστοιχο κατάλογο του τρέχοντος γνωρίσματος. Μετά τον διαχωρισμό ενός κόμβου γίνεται τροποποίηση της λίστας κατηγοριών, ώστε να υποδείξουν τον κόμβο στον οποίο ανήκει η εγγραφή. Με βάση το δείκτη GINI γίνεται επιλογή του χαρακτηριστικού που θα εξεταστεί.

Έστω **c** κλάσεις και **n** αντικείμενα.

$\mathbf{P_i}$: σχετική συχνότητα της κλάσης **i** στο σύνολο **S**

$$gini(S) = 1 - \sum_{i=1}^{c} p_i^{\,2}$$

Ένα από τα βασικά μειονεκτήματα του αλγόριθμου SLIQ είναι πως πρέπει να βρίσκεται συνεχώς στη μνήμη, έτσι ώστε να επιτύχουμε μια καλή απόδοση, γεγονός που περιορίζει το μέγιστο επιτρεπτό μέγεθος του συνόλου εκπαίδευσης. Επίσης, ενημερώνει μόνο τη λίστα κλάσεων και δεν παραλληλίζεται εύκολα.

Αλγόριθμος C4.5

Ο Αλγόριθμος C4.5 εφαρμόζει μια απλή κατά-βάθος μέθοδο για την κατασκευή του δέντρου απόφασης και βελτιώνει τον αλγόριθμο ID3. Πιο συγκεκριμένα :

- **Ελλιπή δεδομένα:** Όταν το δέντρο απόφασης χτίζεται, τα ελλιπή δεδομένα αγνοούνται. Αυτό σημαίνει ότι το κέρδος υπολογίζεται λαμβάνοντας υπόψη μόνο τις εγγραφές που έχουν τιμή. Για να κατηγοριοποιήσουμε ένα σύνολο με ελλιπή

τιμή σε ένα χαρακτηριστικό, η τιμή αυτή μπορεί να προβλεφτεί με βάση των υπόλοιπων τιμών αυτού του χαρακτηριστικού.

- **Συνεχή δεδομένα:** Τα χαρακτηριστικά που παίρνουν συνεχής τιμές, χωρίζονται σε διαστήματα.
- **Περικοπή:** Υπάρχουν δύο σημαντικές στρατηγικές περικοπής που χρησιμοποιεί ο αλγόριθμος C4.5, οι οποίες είναι:
  - **Αντικατάσταση του υποδέντρου**: ένα υποδέντρο αντικαθιστάται από ένα φύλλο, αν αυτή η αντικατάσταση έχει ως αποτέλεσμα σφάλμα κοντά σε αυτό του αρχικού υποδέντρου. Η τεχνική αυτή εφαρμόζεται ξεκινώντας από τα φύλλα και ανεβαίνοντας προς τη ρίζα.
  - **Ανύψωση υποδέντρου:** αντικαθιστά ένα υποδέντρο με το περισσότερο χρησιμοποιούμενο υποδέντρο του. Έτσι ένα υποδέντρο ανυψώνεται, αφού αντικαθιστά ένα άλλο που βρίσκεται σε ψηλότερο επίπεδο. Και σε αυτή την περίπτωση πρέπει να λάβουμε υπόψη την αύξηση στη συχνότητα λαθών.
- **Κανόνες:** Ο C4.5 επιτρέπει την κατηγοριοποίηση είτε μέσω δέντρων αποφάσεων είτε μέσω κανόνων που δημιουργούνται από αυτό. Επίσης, προτείνονται κάποιες τεχνικές που απλουστεύουν τους πολύπλοκους κανόνες.
- **Διάσπαση:** Ο ID3 προτιμά τα χαρακτηριστικά με πολλές διαιρέσεις. Ωστόσο αυτό μπορεί να οδηγήσει σε υπερπροσαρμογή. Μια οριακή περίπτωση είναι να έχουμε ένα χαρακτηριστικό που έχει μια μοναδική τιμή για κάθε σύνολο. Το χαρακτηριστικό αυτό θα είναι το καλύτερο, αφού θα υπήρχε μόνο ένα σύνολο (και έτσι μόνο μια κατηγορία) για κάθε διαίρεση. Μια βελτίωση θα μπορούσε να γίνει αν λάβουμε υπόψη την πληθικότητα της κάθε διαίρεσης. Αυτή η προσέγγιση χρησιμοποιεί το GainRatio και όχι το Gain που χρησιμοποιείται στον ID3.

$$GainRatio\ (D,S) = \frac{Gain\,(D,S)}{H(\frac{|D_1|}{|D|},\ldots,\frac{|D_s|}{|D|})}$$

Για την διάσπαση, ο C4.5 χρησιμοποιεί το μεγαλύτερο GainRatio το οποίο εξασφαλίζει ένα μεγαλύτερο από το μέσο όρο κέρδος στην πληροφορία. Αυτό αντισταθμίζει το γεγονός ότι η τιμή του GainRatio κλίνει προς διασπάσεις, όπου το μέγεθος του ενός υποσυνόλου είναι κοντά προς αυτό του αρχικού. Τέλος, ο αλγόριθμος αυτός κρίνεται

ακατάλληλος για μεγάλα σύνολα δεδομένων, αφού η ακρίβεια που παρουσιάζει είναι πολύ μικρή.

Αλγόριθμος SPRINT

Ο αλγόριθμος SPRINT , δημιουργήθηκε για να απαντήσει στα προβλήματα μνήμης που είχαν παλαιότεροι αλγόριθμοι κατηγοριοποίησης, αφού δεν απαιτεί να υπάρχουν όλα τα δεδομένα στη μνήμη του συστήματος, για την υλοποίησή του. Είναι γρήγορος, ενώ μπορεί να χρησιμοποιηθεί σε οποιοδήποτε όγκο δεδομένων και για αριθμητικά και για κατηγορικά δεδομένα. Επιπλέον, μπορεί να χρησιμοποιηθεί με τέτοιο τρόπο, ώστε να επιτρέπει την παράλληλη χρήση πολλών επεξεργαστών. Για τη δημιουργία ενός κόμβου και τον διαχωρισμό των δεδομένων, χρησιμοποιείται και εδώ η παράμετρος GINI. Με αυτό τον τρόπο, έχουμε το πλεονέκτημα, ότι απαιτείται ο υπολογισμός μόνο της κατανομής της κλάσης σε κάθε διαχωρισμό. Βασικό μειονέκτημα του SPRINT είναι, ότι απαιτεί να υπάρχει συνεχής πρόσβαση σε όλα τα δεδομένα. Αυτό σημαίνει ότι θα πρέπει να διατηρούνται όλες οι λίστες στη μνήμη και στο δίσκο και δεν παρέχει τρόπους να μειωθεί το μέγεθός τους κατά την εκτέλεση του. Επίσης με τη χρήση και μνήμης και σκληρού, εκτελούνται πολλές πράξεις εισόδου/ εξόδου δεδομένων και με αυτό τον τρόπο μειώνεται η ταχύτητα και η απόδοσή του.

Λοιποί αλγόριθμοι βασιζόμενοι σε δέντρα απόφασης

Μια συνέχεια των παραπάνω βασικών αλγορίθμων, αποτελούν και οι αλγόριθμοι που θα αναφέρουμε στη συνέχεια οι οποίοι αναπτύχθηκαν τα τελευταία χρόνια και χρησιμοποιούνται σε λογισμικά συστήματα που διατίθενται στο εμπόριο ή από εταιρείες οι οποίες ασχολούνται με τον τομέα της εξόρυξης δεδομένων και τα εκμεταλλεύονται για λογαριασμό διαφόρων πελατών τους.

Πιο συγκεκριμένα:

Αλγόριθμος PUBLIC

Στον αλγόριθμο αυτό η διαδικασία κλαδέματος εκτελείται περιοδικά, καθώς αναπτύσσεται το δέντρο. Ένας κόμβος δεν αναπτύσσεται αν υπολογιστεί ότι θα πρέπει να κλαδευτεί στην επόμενη φάση. Για να υπολογιστεί αν ένας κόμβος πρέπει ή όχι να αναπτυχθεί, υπολογίζεται

ένα κάτω όριο, του ελάχιστου ορίου, του υπό-δέντρου που θα αναπτυσσόταν κάτω από το συγκεκριμένο κόμβο. Με αυτό τον τρόπο προβλέπεται ποιοι κόμβοι θα κλαδευτούν σε επόμενη φάση, οπότε δεν είναι αναγκαίο να αναπτυχθούν.

Αλγόριθμος BOAT

Ένας άλλος αλγόριθμος είναι ο αλγόριθμος BOAT, ο οποίος κάνει χρήση δειγμάτων από τα δεδομένα. Αρχικά, δημιουργεί πρόχειρα δέντρα αποφάσεων, χρησιμοποιώντας δείγματα. Αν υπάρχουν αριθμητικά δεδομένα, υπολογίζει τα διαστήματα στα οποία θα γίνει διαχωρισμός. Μετά με ένα πέρασμα στα δεδομένα, καθορίζει την ακριβή τιμή στην οποία θα γίνει ο διαχωρισμός. Ύστερα, επαληθεύεται αν ο διαχωρισμός που επιλέχτηκε είναι όντως ο βέλτιστος και αν δεν είναι ξαναδημιουργείται το υπό-δέντρο που έχει σα ρίζα το συγκεκριμένο κόμβο.

2.4 Παράλληλη Κατηγοριοποίηση

Η κατηγοριοποίηση και η εξόρυξη δεδομένων αφορούν μεγάλα σύνολα δεδομένων. Αυτό πρακτικά σημαίνει ότι θα πρέπει οι διάφοροι αλγόριθμοι και μεθοδολογίες να επικεντρωθούν σε θέματα ταχύτητας και οικονομίας κατά την υλοποίηση, με τέτοιο τρόπο ώστε να μπορούν να εφαρμοστούν σε οποιαδήποτε δεδομένα, οποιουδήποτε όγκου. Οι φυσικοί περιορισμοί στη μνήμη στρέφουν τις μεθοδολογίες στη χρήση πολλών πράξεων εισόδου και εξόδου, γεγονός που μειώνει την ταχύτητά τους. Επομένως, προβάλλει επιτακτική η ανάγκη για αλγόριθμους που θα μπορούν να είναι αποδοτικοί, γρήγοροι και να μπορούν να χρησιμοποιηθούν σε οποιοδήποτε όγκο δεδομένων. Γενικά, μεθοδολογίες με τη δυνατότητα να χειρίζονται περισσότερα δεδομένα με περισσότερη ακρίβεια και σε μικρότερα χρονικά διαστήματα. Γι' αυτό και εισάγεται το θέμα της παράλληλης επεξεργασίας των δεδομένων, από παραπάνω από έναν επεξεργαστές.

Υπάρχουν δύο τρόποι:

- ✓ Ο ένας τρόπος είναι να διαχωρίσουμε τα δεδομένα και να τα κατηγοριοποιήσουμε παράλληλα. Δημιουργούνται ,όμως, πολλά δέντρα αποφάσεων και αυτό αυξάνει το κόστος.

- ✓ Ένας δεύτερος τρόπος είναι καθώς δημιουργείται το δέντρο να γίνεται παράλληλη επεξεργασία πολλών κόμβων ενός επιπέδου, αφού έτσι και αλλιώς πρόκειται για δεδομένα διαφορετικά μεταξύ τους.

Ο δεύτερος τρόπος είναι πιο φυσικός για την κατηγοριοποίηση, αλλά δημιουργούνται διάφορα προβλήματα ταχύτητας, λόγω της μεταφοράς των δεδομένων από τους γονείς στα παιδιά, αλλά και της ανισορροπίας που μπορεί να υπάρξει λόγω της ανισότητας του όγκου των πράξεων. Για την αντιμετώπιση αυτών των προβλημάτων δημιουργήθηκαν μερικές νέες πρακτικές. Μία από αυτές είναι η προσέγγιση της *συγχρονισμένης δημιουργίας δέντρων* (Synchronous Tree Construction). Σύμφωνα με αυτή, τα δεδομένα μοιράζονται ίσα στους επεξεργαστές και υπάρχει παράλληλη επεξεργασία τους. Τα μειονεκτήματα αυτής της προσέγγισης είναι ότι μπορεί να δημιουργηθεί και εδώ ανισορροπία στους επεξεργαστές και επίσης να υπάρξουν δυσλειτουργία και μείωση της ταχύτητας στο επίπεδο των φύλλων. Μία άλλη προσέγγιση είναι *ο διαχωρισμός του δέντρου* (Partitioned Tree Construction), δηλαδή ο διαχωρισμός και των δεδομένων και των κόμβων ίσα στους επεξεργαστές. Αυτή η προσέγγιση έχει αποδώσει αρκετά καλά, αλλά υπάρχει και εδώ μείωση της ταχύτητας λόγω του μεγάλου όγκου ανταλλαγής πληροφοριών. (Γολέμη Ε,2010)

2.5 Νευρωνικά Δίκτυα

Μια άλλη προσέγγιση της κατηγοριοποίησης που χρησιμοποιείται σε πολλές εφαρμογές εξόρυξης γνώσης για πρόβλεψη και κατηγοριοποίηση βασίζεται στα νευρωνικά δίκτυα, τα οποία βοηθούν στην κατασκευή ενός μοντέλου κατηγοριοποίησης ή πρόβλεψης. Κατάλληλα βήματα που συντελούν στη διαδικασία αυτή είναι:

Βήμα 1^{ο}: Αναγνώριση των χαρακτηριστικών εισόδου και εξόδου.

Βήμα 2^{ο}: Κατασκευή ενός δικτύου με την κατάλληλη τοπολογία.

Βήμα 3^{ο}: Επιλογή του σωστού συνόλου εκπαίδευσης.

Βήμα 4^{ο}: Εκπαίδευση του δικτύου, με βάση ένα αντιπροσωπευτικό σύνολο δεδομένων.

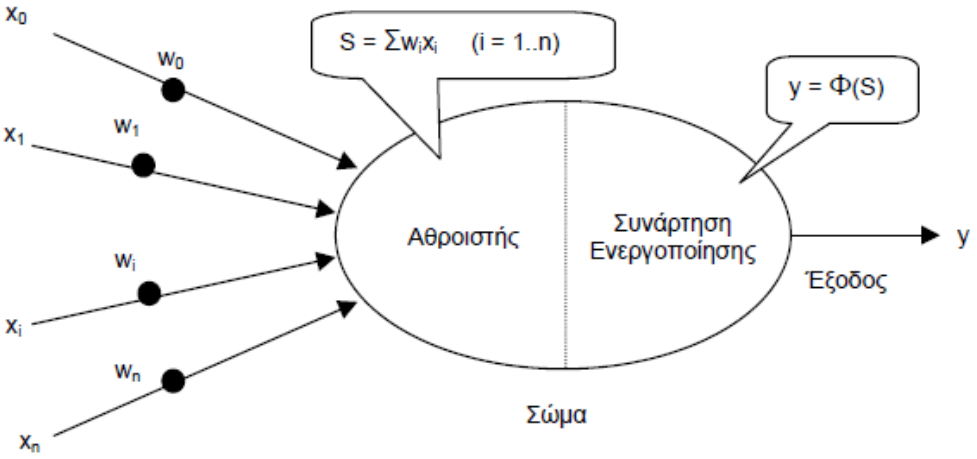
Βήμα 5^{ο}: Έλεγχος δικτύου, χρησιμοποιώντας ένα σύνολο ελέγχου (test data set) το οποίο είναι ανεξάρτητο από το σύνολο εκπαίδευσης (training data set).

2.5.1 Κατηγοριοποίηση με βάση Νευρωνικά Δίκτυα

Είναι μια αρχιτεκτονική δομή, αποτελούμενη από ένα πλήθος διασυνδεδεμένων μονάδων (τεχνητοί νευρώνες). Κάθε μονάδα χαρακτηρίζεται από εισόδους και εξόδους και υλοποιεί τοπικά έναν απλό υπολογισμό. Κάθε σύνδεση μεταξύ δύο μονάδων χαρακτηρίζεται από μια τιμή βάρους. Οι τιμές των βαρών των συνδέσεων αποτελούν τη γνώση που είναι αποθηκευμένη στο δίκτυο και καθορίζουν τη λειτουργικότητά του. Η έξοδος κάθε μονάδας καθορίζεται από τον τύπο της μονάδας, τη διασύνδεση με τις υπόλοιπες μονάδες και πιθανώς κάποιες εξωτερικές εισόδους.

Μοντέλο Τεχνητού Νευρώνα

- Σήματα εισόδου $x_0, x_1, ..., x_n$: Συνεχείς μεταβλητές.
- Τιμή βάρους w_i (weight): Αντίστοιχο των συνάψεων.
- Σώμα του τεχνητού νευρώνα:
  - ✓ Αθροιστής (sum): προσθέτει τα επηρεασμένα από τα βάρη σήματα εισόδου και παράγει την ποσότητα *S*.
  - ✓ Συνάρτηση ενεργοποίησης ή κατωφλίου (activation ή threshold function): μη γραμμικό φίλτρο, που διαμορφώνει το σήμα εξόδου *y*, σε συνάρτηση με την ποσότητα *S*.



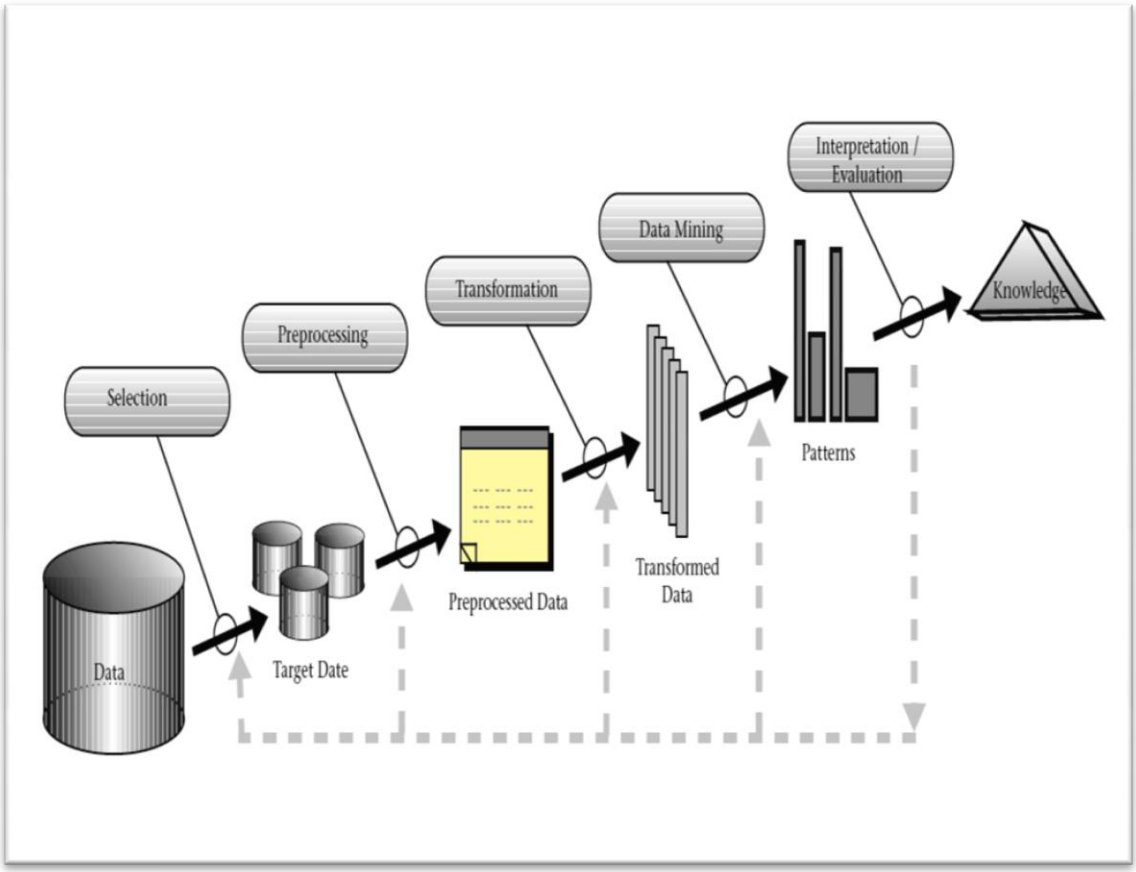
Εικόνα 8: Δομή νευρωνικού δικτύου

Συσταδοποίηση

2.6 Εισαγωγή

Η Συσταδοποίηση (clustering), είναι μια από τις βασικές εργασίες της διαδικασίας εξόρυξης γνώσης, η οποία εφαρμόζει καταμερισμό ενός ετερογενούς πληθυσμού σε ένα σύνολο συστάδων, έτσι ώστε τα στοιχεία του συνόλου των δεδομένων, που ανήκουν σε μια συστάδα, να είναι περισσότερο όμοια μεταξύ τους από ότι είναι με τα στοιχεία των άλλων συστάδων. Σκοπός της διαδικασίας αυτής είναι να οργανώσει τα δεδομένα σε «λογικές» συστάδες, έτσι ώστε να γίνονται αντιληπτές οι ομοιότητες και οι διαφορές, δίνοντας χρήσιμα συμπεράσματα. Στη βιβλιογραφία τη συσταδοποίηση τη συναντάμε και ως : *μη εποπτευμένη μάθηση (unsupervised learning)* στην αναγνώριση προτύπων, *αριθμητική ταξινόμηση (numerical taxonomy)* στην οικολογία και στη βιολογία, *τυπολογία (typology)* στις κοινωνικές επιστήμες και *τμηματοποίηση (partition)* στη θεωρία γράφων. Η συσταδοποίηση, ανάλογα με το κριτήριο που χρησιμοποιείται, χωρίζει τα δεδομένα σε διαφορετικά τμήματα, γι' αυτό το λόγο, χρειάζεται αρχικά η προ-επεξεργασία του συνόλου των δεδομένων.

Στο παρακάτω σχήμα απεικονίζονται τα βασικά βήματα ανάπτυξης της διαδικασίας της συσταδοποίησης:



Εικόνα 9: Βήματα διαδικασίας συσταδοποίησης

- ***Επιλογή χαρακτηριστικών γνωρισμάτων (selection):*** Γίνεται επιλογή των κατάλληλων γνωρισμάτων, τα οποία πρόκειται να συσταδοποιηθούν, ώστε να κωδικοποιηθεί όσο το δυνατόν περισσότερη πληροφορία, σχετικά με το πρόβλημα που μας ενδιαφέρει. Σε ορισμένες περιπτώσεις η προ-επεξεργασία κρίνεται απαραίτητη.
- ***Επιλογή αλγορίθμου συσταδοποίησης (clustering algorithm selection):*** Γίνεται επιλογή του αλγορίθμου για την ανακάλυψη της δομής των ομάδων που υπάρχουν στα δεδομένα. Το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης χαρακτηρίζουν τον αλγόριθμο, όπως:
    1. Το μέτρο γειτνίασης προσδιορίζει την ομοιότητα δύο διανυσμάτων γνωρισμάτων. Πρέπει να εξασφαλίσουμε ότι όλα τα επιλεγμένα γνωρίσματα συμβάλλουν εξίσου στον υπολογισμό μέτρου γειτνίασης και δεν υπάρχει κανένα γνώρισμα που να κυριαρχεί σε άλλα.

2. Το κριτήριο συσταδοποίησης εκφράζεται μέσω μιας συνάρτησης κόστους ή κάποιου άλλου κανόνα, λαμβάνοντας υπόψη τον τύπο των συστάδων, που πρόκειται να εμφανιστούν στο σύνολο των δεδομένων.

- ***Αξιοπιστία (validation) αποτελεσμάτων :*** Εφόσον καθοριστούν οι συστάδες, θα πρέπει να αξιολογηθούν, εάν τα αποτελέσματά τους είναι ακριβή.
- ***Ερμηνεία αποτελεσμάτων (interpretation):*** Γίνεται ερμηνεία και ανάλυση των αποτελεσμάτων, προκειμένου να προκύψει το σωστό συμπέρασμα.

2.6.1 Εφαρμογές συσταδοποίησης

Η συσταδοποίηση είναι ένα σημαντικό εργαλείο με πολλές εφαρμογές σε πολλά πεδία, όπως : επιχειρήσεις , βιολογία, χωρική ανάλυση στοιχείων, εξόρυξη στον παγκόσμιο ιστό, στατιστική ανάλυση δεδομένων, ανάκτηση πληροφοριών, χημεία, ιατρικές επιστήμες , κοινωνικές επιστήμες ,επιστήμες γης και επιστήμες μηχανικών.

Παρακάτω περιγράφονται οι εφαρμογές της συσταδοποίησης , όπως:

- Μείωση δεδομένων (data reduction): Συμπίεση της πληροφορίας των δεδομένων, χωρίζοντας το σύνολο των δεδομένων σε ένα αριθμό συστάδων.
- Παραγωγή υπόθεσης (hypothesis generation): Η συσταδοποίηση χρησιμοποιείται σε αυτή την περίπτωση έτσι, ώστε προκύπτοντας κάποιες υποθέσεις, να αντλήσουμε χρήσιμα συμπεράσματα.
- Έλεγχος υποθέσεων (hypothesis testing): Η συσταδοποίηση χρησιμοποιείται για την επαλήθευση της αξιοπιστίας μιας συγκεκριμένης υπόθεσης .
- Πρόβλεψη βασισμένη σε συστάδες (prediction based on clusters): Η συσταδοποίηση εφαρμόζεται σε σύνολα δεδομένων και οι συστάδες που προκύπτουν χαρακτηρίζονται από τα χαρακτηριστικά των προτύπων που ανήκουν στις συστάδες αυτές. Εν συνεχεία, όταν τα άγνωστα πρότυπα εμφανιστούν, ταξινομούνται στις προσδιοριζόμενες συστάδες σύμφωνα με την ομοιότητά τους στα χαρακτηριστικά των συστάδων.

2.6.2 Μέθοδοι συσταδοποίησης

Οι αλγόριθμοι συσταδοποίησης ταξινομούνται βάσει:

- Τον ***τύπο δεδομένων*** που εισάγονται στον αλγόριθμο:
  - ***Συσταδοποίηση αριθμητικών δεδομένων***, όπου εφαρμόζεται σε βάσεις δεδομένων με τύπο γνωρισμάτων αριθμητικές τιμές . Η ομοιότητα ή η απόσταση μεταξύ των αντικειμένων μετράται με τη χρήση της Ευκλείδειας Απόστασης ή της City- block απόστασης .
  - ***Εννοιολογική συσταδοποίηση***, όπου εφαρμόζεται σε βάσεις δεδομένων με τύπο γνωρισμάτων κείμενο: η απόσταση δεν είναι κατάλληλη, οπότε, εναλλακτικά, χρησιμοποιείται ο αριθμός των γνωρισμάτων, που δεν είναι κοινά σε δύο αντικείμενα.
- Τη ***μέθοδο που καθορίζει τη συσταδοποίηση*** του συνόλου των δεδομένων:
  - ***Διαιρετική συσταδοποίηση (partitional clustering),*** λειτουργεί αποσυνθέτοντας το σύνολο των δεδομένων σε ένα σύνολο που περιέχει μη συσχετιζόμενες συστάδες. Επομένως, το γενικό κριτήριο είναι η ελαχιστοποίηση κάποιων μέτρων ανομοιότητας, καθώς και η μεγιστοποίηση της ανομοιότητας μεταξύ των δειγμάτων μέσα σε κάθε μία από τις συστάδες .
  - ***Ασαφής συσταδοποίηση (fuzzy clustering)*** , η οποία χρησιμοποιεί τεχνικές ασαφούς λογικής, προκειμένου να ομαδοποιήσει δεδομένα, θεωρώντας ότι ένα αντικείμενο μπορεί να ταξινομηθεί σε πολλές συστάδες.
  - ***Μη ασαφής συσταδοποίηση (crisp clustering),***θεωρεί ότι ένα στοιχείο του συνόλου δεδομένων είτε ανήκει σε μία κατηγορία είτε σε καμία. Οι περισσότεροι αλγόριθμοι ανήκουν στην κατηγορία αυτή.
  - ***Συσταδοποίηση βασισμένη στα δίκτυα Kohonen (Kohonen net clustering),*** βασίζεται στην έννοια των νευρωνικών δικτύων. Το δίκτυο Kohonen έχει κόμβους εισόδου και εξόδου. Το επίπεδο εισόδου έχει ένα κόμβο για κάθε γνώρισμά μιας εγγραφής τα οποία συνδέονται με κάθε κόμβο εξόδου. Κάθε σύνδεση σχετίζεται με ένα βάρος, το οποίο καθορίζει τη θέση του αντίστοιχου κόμβου εξόδου. Επομένως, βάση ενός αλγορίθμου που αλλάζει κατάλληλα τα βάρη, οι κόμβοι εξόδου σχηματίζουν συστάδες.

- ***Ιεραρχική συσταδοποίηση (hierarchical clustering),*** όπου οι αλγόριθμοι βασίζονται στη διαδοχική σύνδεση μικρότερων συστάδων ή τη διάσπαση μεγαλύτερων σε μικρότερες. Οι μέθοδοι συσταδοποίησης χρησιμοποιούν διαφορετικό κανόνα, με βάση τον οποίο γίνεται η διάσπαση ή η συγχώνευση. Το αποτέλεσμα του αλγορίθμου είναι ένα δενδρογράφημα το οποίο αν το κόψουμε σε κάποιο επίπεδο, έχουμε ως αποτέλεσμα μη συσχετιζόμενες συστάδες.
- ***Συσταδοποίηση βασισμένη στην πυκνότητα (Density-based clustering),*** οργανώνει γειτονικά αντικείμενα ενός συνόλου δεδομένων σε συστάδες με βάση ορισμένα κριτήρια πυκνότητας .
- ***Συσταδοποίηση βασισμένη σε πλέγμα (Grid- based clustering),*** αυτός ο τύπος αλγορίθμων χωρίζει το χώρο σε έναν πεπερασμένο αριθμό κελιών και στη συνέχεια γίνονται όλες τις διαδικασίες στο χώρο αυτό (ανάλυση χωρικών δεδομένων).
- ***Συσταδοποίηση υποχώρων (Subspace clustering),*** οι αλγόριθμοι αυτοί προσπαθούν να βρουν τα υποσύνολα του αρχικού χώρου, όπου τα αποτελέσματα συσταδοποίησης είναι καλύτερα.

➢ ***Τη θεωρία και τις θεμελιώδης έννοιες*** στις οποίες είναι βασισμένες οι τεχνικές ανάλυσης συστάδας .

2.6.3 Αλγόριθμοι Συσταδοποίησης

Οι αλγόριθμοι συσταδοποίησης μπορεί να είναι ιεραρχικοί (hierarchical) ή μη ιεραρχικοί/διαιρετικοί (non-hierarchical/partitional). Οι ιεραρχικοί αλγόριθμοι βρίσκουν διαδοχικές ομάδες, χρησιμοποιώντας κάθε φορά τις ήδη καθιερωμένες ομάδες, ενώ οι μη ιεραρχικοί/διαιρετικοί καθορίζουν τις ομάδες αμέσως.

2.6.3.1 Διαιρετικοί αλγόριθμοι

Οι διαιρετικοί αλγόριθμοι ξεκινούν με όλα τα στοιχεία να εμπεριέχονται μέσα σε μία συστάδα και σταδιακά τη διαιρούν σε μικρότερες συστάδες, έως ότου ικανοποιηθεί η

συνθήκη τερματισμού. Η βασική ιδέα είναι, ότι μια συστάδα διασπάται, όταν κάποια από τα στοιχεία της δεν βρίσκονται αρκετά κοντά στα υπόλοιπα στοιχεία της. Κάποιες φορές, μπορεί να περιλαμβάνουν τεχνικές κλαδέματος και συγχώνευσης, ώστε να επιτευχθεί ένα πιο βελτιωμένο τελικό αποτέλεσμα. Οι βασικοί διαιρετικοί αλγόριθμοι είναι οι εξής: K-Means, PAM, CLARA, CLARANS.

K-MEANS

Ο αλγόριθμος k-means (k-μέσων) είναι ένας αλγόριθμος, που ομαδοποιεί αντικείμενα, βάσει των χαρακτηριστικών των k συστάδων. Ο αλγόριθμος υποθέτει ότι τα χαρακτηριστικά του αντικειμένου δημιουργούν ένα χώρο διανυσμάτων και ο σκοπός του είναι να ελαχιστοποιήσει τη συνολική διακύμανση της ομάδας ή τη συνάρτηση τετραγωνικού σφάλματος:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left| x_j - \mu_i \right|^2$$

Όπου υπάρχουν k ομάδες S_i, i = 1,2,...,k και μ_i είναι το κεντροειδές ή το μεσαίο σημείο από όλα τα σημεία.

Τα βήματα του αλγορίθμου είναι τα εξής:

i. Ανάθεση των αρχικών κέντρων, $\mathbf{v_i}$ **i = 1, 2, ..., c**, για τις **c** συστάδες.
   Για κάθε επανάληψη **r = 1, ..., $\mathbf{r_{max}}$**:
ii. Υπολογισμός της απόστασης κάθε στοιχείου του συνόλου δεδομένων από το κέντρο κάθε συστάδας $\mathbf{d_{ki} = (x_k - v_i)^2}$**, k = 1, 2, .., n i = 1, 2, ..., c**
iii. Κάθε στοιχείο $\mathbf{x_k}$ αντιστοιχίζεται στη συστάδα με την ελάχιστη απόσταση.
iv. Υπολογισμός των νέων κέντρων των συστάδων.

$$\mathbf{m_i^{(r+1)} = \frac{\sum_{k=1}^{n_i} x_k}{n_i}}$$

Όπου $\mathbf{n_i}$, ο αριθμός των στοιχείων που ανήκουν στην **i** συστάδα, μέχρι στιγμής.

v. If $|| m_i^{(r)} - m_i^{(r+1)}|| < \varepsilon$ then
Stop
Else
R= r+1, goto2.

Ο αλγόριθμος ξεκινά διαχωρίζοντας τα αρχικά σημεία σε k αρχικά σύνολα. Στη συνέχεια, υπολογίζει το μεσαίο ή το κεντροειδές του κάθε συνόλου και υλοποιεί νέο διαχωρισμό, ώστε το κάθε σημείο να σχετίζεται με το κοντινότερο κεντροειδές. Έπειτα, τα κεντροειδή επαναϋπολογίζονται για τις νέες ομάδες και ο αλγόριθμος επαναλαμβάνει τα δύο βήματα, ωσότου τα σημεία δεν μπορούν να αλλάξουν ομάδες. Ένα από τα βασικά πλεονεκτήματα είναι, πως ο αλγόριθμος αυτός τείνει σε κάποιο όριο πολύ γρήγορα. Παρόλα αυτά, όσον αφορά στην απόδοση, ο αλγόριθμος δεν εγγυάται ότι θα αγγίξει το βέλτιστο. Η ποιότητα της τελικής λύσης εξαρτάται πολύ από το αρχικό σύνολο ομάδων και μπορεί να είναι πολύ χαμηλότερη από το συνολικό βέλτιστο. Επίσης, ένα άλλο μειονέκτημα του αλγόριθμου είναι, ότι ο αριθμός των ομάδων πρέπει να οριστεί εξαρχής.

Παραλλαγές K-Means

Ορισμένες χαρακτηριστικές παραλλαγές του K-Means είναι:

- Ο *αλγόριθμος* ***ISODATA,*** ο οποίος περιλαμβάνει μία διαδικασία για αναζήτηση του καλύτερου αριθμού συστάδων, με βάση κάποιο κόστος εκτέλεσης.
- Ο ***Fuzzy C-Means,*** ο οποίος επεκτείνει τον κλασικό αλγόριθμο K-Means, χρησιμοποιώντας την θεωρία της ασαφής λογικής, ο οποίος θα αναλυθεί παρακάτω και τέλος,
- Ο ***SAS PROC FASTCLUS,*** ο οποίος ελέγχει την διαδικασία συσταδοποίησης, υιοθετώντας δύο ακόμα παραμέτρους, την **max_rad** και **min_size**. Η πρώτη παράμετρος ελέγχει τον ελάχιστο αριθμό στοιχείων που μπορεί να έχει κάθε συστάδα, ενώ η δεύτερη καθορίζει, ότι η απόσταση κάθε στοιχείου μίας συστάδας από το κέντρο της συστάδας δεν πρέπει να είναι μεγαλύτερη του **max_rad**.
- Ο ***k-windows,*** που χρησιμοποιεί τεχνικές υπολογιστικής γεωμετρίας.

Fuzzy k-means

Ο αλγόριθμος αυτός είναι μια μέθοδος συσταδοποίησης που επιτρέπει τα δεδομένα να ανήκουν σε περισσότερες από μια συστάδες. Αυτή η μέθοδος χρησιμοποιείται συχνά στην αναγνώριση προτύπων και βασίζεται στην ελαχιστοποίηση της παρακάτω συνάρτησης:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \|x_i - c_j\|^2 \qquad 1 \le m < \infty$$

Όπου *m* είναι πραγματικός αριθμός πάνω από 1, *uij* είναι ο βαθμός της ιδιότητας μέλους του στην ομάδα *j*, *xi* είναι το *i*th της δ-διάστασης των καταγεγραμμένων δεδομένων, *cj* είναι το κέντρο της ομάδας δ-διάστασης, *cj* και ||*|| είναι κάθε κανόνας που εκφράζει την ομοιότητα μεταξύ των δεδομένων και του κέντρου.

Βήματα αλγορίθμου:

i. Επιλέγουμε μία αρχικά fuzzy διαίρεση των *N* αντικειμένων σε *k* συστάδες επιλέγοντας ένα *N*x*K* πίνακα γειτνίασης *U*. Η τιμή u_{ij} καθορίζει το βαθμό συμμετοχής του αντικειμένου x_i στο cluster c_j

ii. Με βάση το *U* υπολογίζουμε την τιμή ενός fuzzy κριτηρίου:

$$E^2(\mathcal{X}, \mathbf{U}) = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ij} \|\mathbf{x}_i - \mathbf{c}_k\|^2,$$

όπου

$$\mathbf{c}_k = \sum_{i=1}^{N} u_{ik} \mathbf{x}_i$$

iii. Επαναϋπολογίζουμε τα cluster centers για να μειώσουμε το κριτήριο αυτό.

iv. Επαναλαμβάνουμε το βήμα ii.

Αλγόριθμος PAM

Ο αλγόριθμος PAM (Partitioning Around Medoids) αναπτύχθηκε από τους Kaufman και Rousseeuw. Για την εύρεση των k συστάδων, η μεθοδολογία του PAM είναι ο καθορισμός ενός αντιπροσωπευτικού σημείου για κάθε συστάδα. Αυτό το αντιπροσωπευτικό σημείο, που ονομάζεται medoid, θεωρείται ότι είναι το πιο κεντρικό σημείο της συστάδας . Αφού βρεθούν τα medoids, κάθε ένα από τα υπόλοιπα αντικείμενα ομαδοποιείται μαζί με το κοντινότερο του medoid.

Το κρίσιμο βήμα του αλγορίθμου είναι η επιλογή των k medoids. Για την εύρεση τους, ο αλγόριθμος PAM αρχικά επιλέγει τυχαία k αντικείμενα. Κατόπιν σε κάθε βήμα, γίνεται εναλλαγή μεταξύ ενός επιλεγμένου αντικειμένου Oi ($1 \leq i \leq N$) και ενός μη επιλεγμένου Oh ($1 \leq h \leq N$ και $h \neq i$), εφόσον αυτή η εναλλαγή οδηγεί σε βελτίωση της ποιότητας της συσταδοποίησης. Ο απλοποιημένος αλγόριθμος PAM για την ομαδοποίηση N αντικειμένων σε k συστάδες, θα αναλυθεί παρακάτω. Το κόστος αντικατάστασης Ci,h για το ζευγάρι αντικειμένων (Oi,Oh) είναι η (θετική ή αρνητική) αύξηση του ολικού αθροίσματος των αποστάσεων σε κάθε συστάδα, αν υποτεθεί, ότι το Oh γίνεται αυτό νέο medoid στη θέση του (παλιού medoid) Oi.

Βασικά βήματα αλγορίθμου PAM:

i. Τυχαία επιλογή K αντιπροσώπων για τις συστάδες.
ii. Υπολογισμός του συνολικού κόστους $\mathbf{TC_{ih}}$ για όλα τα ζεύγη των αντικειμένων O_i, O_h όπου το O_i είναι το τρέχον επιλεγμένο αντικείμενο και το O_h είναι ένα μη επιλεγμένο αντικείμενο.
iii. Επιλέγουμε το ζεύγος O_i, O_h, το οποίο αντιστοιχεί στο $\min_{Oi,\ Oh} TC_{ih}$. Εάν το συνολικό κόστος είναι αρνητικό αντικαθιστούμε το O_i με το O_h και επιστρέφουμε στο βήμα ii.
iv. Διαφορετικά, για κάθε μη επιλεγμένο αντικείμενο, βρίσκουμε το αντικείμενο αντιπρόσωπο, που προσεγγίζει περισσότερο. Τότε ο αλγόριθμος σταματά.

```
Input:
   D = {t_1, t_2, ..., t_n}   // Set of elements
   A      // Adjacency matrix showing distance between elements.
   k      // Number of desired clusters.
Output:
   K      // Set of clusters.
PAM Algorithm:
   arbitrarily select k medoids from D;
   repeat
      for each t_h not a medoid do
         for each medoid t_i do
            calculate TC_ih;
      find i, h where TC_ih is the smallest;
      if TC_ih < 0 then
         replace medoid t_i with t_h;
   until TC_ih ≥ 0;
   for each t_i ∈ D do
      assign t_i to K_j where dis(t_i, t_j) is the smallest over all medoids;
```

Εικόνα 10: Κώδικας Αλγορίθμου PAM

Αλγόριθμος CLARA

Ο αλγόριθμος CLARA (Clustering LARge Applications), είναι μία παραλλαγή του αλγορίθμου k-medoids για πολλά δεδομένα. Λειτουργεί με το να ομαδοποιεί ένα δείγμα του συνόλου και έπειτα εκχωρεί όλα τα δεδομένα του συνόλου σε αυτές τις ομάδες. Ο αλγόριθμος λειτουργεί ως ακολούθως:

1. Για i = 1... 5, επαναλαμβάνουμε τα ακόλουθα βήματα:
2. Επιλέγουμε ένα δείγμα 40 + 2k αντικειμένων με τυχαίο τρόπο από το σύνολο των δεδομένων και καλούμε τον αλγόριθμο PAM για να βρούμε τους k αντιπροσώπους για τις συστάδες.
3. Για κάθε αντικείμενο Oj στο σύνολο δεδομένων, καθορίζουμε ποιο από τα k medoids προσεγγίζει περισσότερο το Oj.
4. Υπολογίζουμε τη συνολική ανομοιότητα για την συσταδοποίηση που λαμβάνεται από το προηγούμενο βήμα. Εάν αυτή η τιμή είναι μικρότερη από το τρέχον ελάχιστο, χρησιμοποιούμε αυτή την τιμή του ελαχίστου σαν τρέχον ελάχιστο και διατηρούμε τα

k medoids που βρήκαμε στο βήμα 2 σαν το καλύτερο σύνολο των medoids που έχουμε μέχρι στιγμής.

5. Επιστρέφουμε στο βήμα 1 και ξεκινάμε με την επόμενη επανάληψη.

Αλγόριθμος CLARANS

Ο αλγόριθμος αυτός λειτουργεί σαν μια βελτιωμένη μέθοδος της μεθόδου CLARA. Ο CLARANS (Clustering Large Applications based on RANdomized Search) θεωρείται ότι ομαδοποιεί καλύτερα, με λιγότερες πράξεις. Στην ουσία, αναζητεί ένα τυχαίο υποσύνολο των γειτόνων μιας λύσης S. Έχει δύο παραμέτρους Maxneighbor, που δηλώνουν τον μέγιστο αριθμό των γειτόνων του, και numlocal, που δηλώνουν τον αριθμό των τοπικών λύσεων που αναμένει. Η λειτουργία του έχει ως εξής:

1. Αρχικοποίηση των παραμέτρων numlocal (αριθμός τοπικών *βέλτιστων, που θα αναζητηθούν*) και *maxneighbor (μέγιστος αριθμός γειτόνων, που μπορούν να εξεταστούν*). Αρχικοποιούμε το i σε 1 και θέτουμε ως ελάχιστο κόστος mincost ένα μεγάλο αριθμό.
2. Καθορισμός της μεταβλητής current (τρέχον κόμβος προς *εξέταση*), ώστε να αναφέρεται σε έναν αρχικό κόμβο $G_{n,k}$.
3. Θέτουμε το j ίσο με 1.
4. Θεωρούμε έναν τυχαίο γείτονα S του τρέχοντος και υπολογίζουμε το κόστος αντικατάστασης του τρέχοντος κόμβου από το γειτονικό κόμβο.
5. Εάν ο S έχει μικρότερο κόστος, θέτουμε ως τρέχον κόμβο (current) τον S και επιστρέφουμε στο βήμα 3.
6. Διαφορετικά, αυξάνουμε το j κατά 1. Εάν j ≤ maxneighbor, επιστρέφουμε στο *βήμα 4*.
7. Διαφορετικά, όταν το *j > maxneighbor*, συγκρίνουμε το κόστος του τρέχοντος κόμβου *current* με το ελάχιστο κόστος *mincost*. Εάν το πρώτο είναι μικρότερο από το mincost, θέτουμε ως mincost το κόστος του current και ορίζουμε ως καλύτερο κόμβο (*bestnode*) τον current.
8. Αυξάνουμε το ***i*** κατά 1. Εάν *i > numlocal*, εξάγουμε τον καλύτερο κόμβο και η διαδικασία σταματά. Διαφορετικά, επιστρέφουμε στο βήμα 2.

2.6.3.2 Ιεραρχικοί αλγόριθμοι

Οι ιεραρχικοί αλγόριθμοι χωρίζονται στους συσσωρευτικούς (agglomerative) και στους διαχωριστικούς (divisive). Οι συσσωρευτικοί αντιμετωπίζουν κάθε στοιχείο σαν μια ομάδα από μόνο του και στη συνέχεια συγχωνεύεται σε μεγαλύτερες ομάδες. Οι διαχωριστικοί ξεκινούν με ολόκληρο το σύνολο και το διασπούν σε μικρότερες ομάδες. Η κλασσική μορφή αυτής της ιεραρχίας είναι το δενδρόγραμμα, όπου τα μεμονωμένα στοιχεία είναι από τη μια μεριά και η ομάδα με το κάθε στοιχείο από την άλλη. Οι συσσωρευτικοί αλγόριθμοι ξεκινάνε από την κορυφή του δέντρου, ενώ οι διαχωριστικοί από τις ρίζες.

Στο σημείο αυτό αξίζει να αναφέρουμε τους βασικούς ιεραρχικούς αλγόριθμους της συσταδοποίησης, οι οποίοι είναι: CURE, BIRCH, CHAMELEON,C^2P, DBSCAN, STING, CLIQUE και ROCK.

Αλγόριθμος CURE

Ο αλγόριθμος αυτός δεν χρησιμοποιεί το μέσο της συστάδας ή ένα αντικείμενο για να αναπαρασταθεί μια συστάδα, αλλά επιλέγεται ένα καθορισμένο σύνολο από αντιπροσωπευτικά αντικείμενα. Η απόσταση μεταξύ δύο συστάδων είναι η απόσταση μεταξύ των δύο πιο κοντινών αντιπροσώπων των δύο συστάδων και όχι μεταξύ των κέντρων τους .

1. Αρχικά κάθε σημείο αποτελεί και μια συστάδα .
2. Στη συνέχεια, υπολογίζονται όλες οι αποστάσεις μεταξύ των συστάδων και οι δυο πιο κοντινές συστάδες ενώνονται σε μια.
3. Η διαδικασία επαναλαμβάνεται μέχρι να φτάσουμε στον επιθυμητό αριθμό από συστάδες .
4. Κάθε φορά που δύο συστάδες ενώνονται, αν το σύνολο των σημείων ξεπερνά τον αριθμό των αντιπροσώπων, μια συνάρτηση μειώνει τον αριθμό των αντιπροσώπων στον επιθυμητό.

```
Input:
   D = {t_1, t_2, ..., t_n}    //Set of elements.
   k      // Desired number of clusters.
Output:
   Q      //Heap containing one entry for each cluster.
CURE Algorithm:
   T = build(D);              // Put each point in Tree
   Q = heapify(D);            // Initially build heap with one entry per item;
   repeat
      u = min(Q);
      delete(Q, u.close);
      w = merge(u, v);
      delete(T, u);
      delete(T, v);
      insert(T, w);
      for each x ∈ Q do
         x.close = find closest cluster to x;
         if x is closest to w then
            w.close = x;
      insert(Q, w);
   until number of nodes in Q is k;
```

Εικόνα 11: Κώδικας Αλγορίθμου CURE

Επεκτάσεις για μεγάλα σύνολα δεδομένων:

- Επιλογή τυχαίου δείγματος από τα δεδομένα.
- Διαχωρισμός του δείγματος σε p διαιρέσεις ίδιου μεγέθους .
- Τμηματοποίηση των σημείων σε κάθε διαίρεση, σε n/*pq* ομάδες, χρησιμοποιώντας την ιεραρχική εκδοχή του CURE, λαμβάνοντας έτσι ένα σύνολο από n/*q* ομάδες.
- Εφαρμογή στους εκπροσώπους των ομάδων του αλγορίθμου CURE, για να γίνει η τμηματοποίηση σε n/*q* ομάδες, μέχρις ότου μείνουν μόνο *K* ομάδες.
- Απομάκρυνση ακραίων σημείων.
- Ανάθεση των υπολοίπων σημείων στην κοντινότερη ομάδα, για να παραχθεί μία πλήρης ομαδοποίηση.

Αλγόριθμος BIRCH

Ο αλγόριθμος αυτός εφαρμόζεται σε «μετρικά» χαρακτηριστικά και είναι σχεδιασμένος για μεγάλο όγκο δεδομένων. Επιπλέον , θεωρεί πως τα δεδομένα δεν είναι ομοιόμορφα στο

χώρο(πυκνή περιοχή-συστάδα, αραιά σημεία δεν λαμβάνονται υπόψη) και χρησιμοποιεί όλη τη διαθέσιμη μνήμη. Το πρόβλημα που καλείται να λύσει ο αλγόριθμος αυτός είναι: δεδομένου του αριθμού K των συστάδων, ενός συνόλου δεδομένων N και μιας συνάρτησης μέτρησης βασισμένης στην απόσταση, το ζητούμενο είναι να γίνει τμηματοποίηση των δεδομένων, έτσι ώστε η συνάρτηση μέτρησης να έχει ελάχιστη τιμή.

Αλγόριθμος CHAMELEON

Ο αλγόριθμος CHAMELEON βρίσκει τις συστάδες του συνόλου δεδομένων χρησιμοποιώντας έναν αλγόριθμο δύο φάσεων. Κατά τη διάρκεια της πρώτης φάσης, ο CHAMELEON χρησιμοποιεί έναν αλγόριθμο συσταδοποίησης βασισμένο σε γράφους, για να τμηματοποιήσει τα δεδομένα σε έναν μεγάλο αριθμό σχετικά μικρών υποσυστάδων. Ο αλγόριθμος της πρώτης φάσης προσπαθεί να ελαχιστοποιήσει το βάρος κάθε ομάδας. Κατά τη διάρκεια της δεύτερης φάσης, χρησιμοποιεί ένα συσσωρευτικό ιεραρχικό αλγόριθμο, για να βρει τις συστάδες από επαναληπτικούς συνδυασμούς των υποσυστάδων, που προέκυψαν από την πρώτη φάση. Η ομοιότητα μεταξύ των συστάδων καθορίζεται με τον έλεγχο της *σχετικής ενδοσυνδετικότητας* (*inter-connectivity*) και της *σχετικής εγγύτητας* (*closeness*) αυτών. Η αναπαράσταση των δεδομένων βασίζεται στη προσέγγιση του k-πλησιέστερου γράφου γειτνίασης (*k-nearest neighbor graph*).

Αλγόριθμος C^2P

Ο C^2P εκμεταλλεύεται τις δομές ευρετηρίων και την επεξεργασία των ερωτήσεων του πιο κοντινού ζευγαριού CPQ στις χωρικές βάσεις δεδομένων. Επίσης, οργανώνει το αποτέλεσμα του CPQ πάνω σε μια χωρική μέθοδο προσπέλασης (R-Tree) σε μια δομή γράφου. Η δομή αυτή αναπαριστά τα Closest Pairs. Κατόπιν, η συσταδοποίηση εκτελείται με τον προσδιορισμό των συστάδων ως συστατικά του γράφου. Οι δύο βασικές φάσεις του αλγόριθμου C^2P είναι οι ακόλουθες :

- ✓ *Φάση Ι*: Παράγει διάφορες υποσυστάδες, οι οποίες είναι μια αποτελεσματική αντιπροσώπευση των τελικών συστάδων. Είναι μια επαναληπτική διαδικασία, κατά τη οποία διάφορες συστάδες συγχωνεύονται. Ο αλγόριθμος χρησιμοποιεί τον αλγόριθμο

DEPTH FIRST SEARCH (DFS) στο γράφο, για να βρει τα συνδεμένα στοιχεία του γράφου, ο οποίος περιλαμβάνει επίσης τις υποσυστάδες του συνόλου δεδομένων.

- ✓ *Φάση ΙΙ* : Είναι μια εξειδικευμένη περίπτωση της πρώτης φάσης, που χρησιμοποιεί μια διαφορετική αναπαράσταση συστάδας, ώστε να παραχθεί το λεπτομερές τελικό σχήμα συσταδοποίησης. Επιπλέον, συγχωνεύει δύο συστάδες σε κάθε βήμα, με σκοπό να ελεγχθεί η διαδικασία συσταδοποίησης.

Αλγόριθμος DBSCAN

Ο αλγόριθμος αυτός είναι βασισμένος στην πυκνότητα. Η λογική του αλγορίθμου είναι πως η περιοχή που εκτείνεται σε συγκεκριμένη ακτίνα (Eps) γύρω από κάθε αντικείμενο μιας συστάδας θα πρέπει να περιέχει έναν ελάχιστο αριθμό από αντικείμενα(Minpts).

Τα αντικείμενα διαχωρίζονται σε :

- Βασικά (core): ένα αντικείμενο για το οποίο υπάρχουν περισσότερα από ένα προκαθορισμένο αριθμό (Minpts) αντικείμενα σε ακτίνα Eps.
- Οριακά (border): ένα αντικείμενο για το οποίο υπάρχουν λιγότερα από ένα προκαθορισμένο αριθμό (Minpts) αντικείμενα σε ακτίνα Eps, αλλά είναι στη γειτονία ενός βασικού αντικειμένου.
- Θορύβου (noise): ένα αντικείμενο που δεν είναι ούτε βασικό, ούτε οριακό.

Βασικός αλγόριθμος

- Χαρακτηρισμός κάθε αντικειμένου ως βασικό, οριακό ή θόρυβο.
- Διαγραφή των αντικειμένων θορύβου.
- Τοποθέτηση μιας ακμής μεταξύ όλων των βασικών αντικειμένων, που είναι σε απόσταση έως Eps μεταξύ τους .
- Μετατροπή κάθε ομάδας συνδεδεμένων βασικών αντικειμένων σε μια διαφορετική συστάδα.
- Ανάθεση κάθε οριακού αντικειμένου σε μια από τις συστάδες των συσχετιζόμενων των βασικών αντικειμένων.

Αλγόριθμος STING

Ο αλγόριθμος STING (STatistical Information Grid-based), είναι αλγόριθμος βασισμένος σε πλέγμα. Χρησιμοποιεί μια ιεραρχική τεχνική κατά την οποία κβαντοποιεί το διάστημα σε ένα πεπερασμένο αριθμό κελιών και εν συνεχεία κάνει όλες τις διαδικασίες στο κβαντοποιημένο διάστημα. Κάθε κόμβος στη δομή πλέγματος συνοψίζει την πληροφορία για τα στοιχεία εντός της. Μπορεί να θεωρηθεί ως τεχνική ιεραρχικής συσταδοποίησης.

```
Input:
   T        // Tree.
   q        // Query.
Output:
   R        // Regions of relevant cells.
STING Algorithm
   i = 1;
   repeat
      for each node in level i do
         determine if this cell is relevant to q and mark as such;
         i = i + 1;
   until all layers in the tree have been visited;
   identify neighboring cells of relevant cells to create regions of cells;
```

Εικόνα 12: Κώδικας Αλγορίθμου STING

Αλγόριθμος CLIQUE

Ο αλγόριθμος CLIQUE προχωρά από χαμηλότερης έως υψηλότερης διάστασης υποχώρους και ανακαλύπτει τις πυκνές περιοχές σε κάθε υποχώρο. Για να προσεγγίσει την πυκνότητα των σημείων, το διάστημα εισόδου χωρίζεται στα κελιά, με τη διαίρεση κάθε διάστασης στον ίδιο αριθμό, x_i, ίσου μήκους διαστημάτων. Για ένα δεδομένο σύνολο διαστάσεων, ο συνδυασμός των αντίστοιχων διαστημάτων (ένα για κάθε διάσταση του συνόλου) καλείται *μονάδα* (unit) στον αντίστοιχο υποχώρο. Μια μονάδα είναι πυκνή, εάν ο αριθμός σημείων είναι επάνω από ένα δεδομένο όριο t. Τα x_i και t είναι παράμετροι, που καθορίζονται από το χρήστη. Ο αλγόριθμος βρίσκει όλες τις πυκνές μονάδες σε κάθε k-διάστατο υποχώρο με τη

δημιουργία των πυκνών (k-1)-διάστατων υποχώρων, και στη συνέχεια, τις συνδέει, για να περιγράψουν συστάδες ως ένωση των μέγιστων ορθογωνίων.

Διαδικασία Αλγορίθμου:

1. Find all the dense areas in the one-dimensional spaces corresponding to each attribute. This is the set of dense one-dimensional cells.
2. k=2
3. **repeat**
4. Generate all candidate dense k-dimensional cells from dense (k-1)-dimensional cells.
5. Eliminate cells that have fewer than ξ points
6. k=k+1
7. **Until** there are no candidate dense k-dimensional cells
8. Find clusters by taking the union of adjacent high-density cells
9. Summarize each cluster using a small set of inequalities that describe the attribute ranges of the cells in the cluster.

Αλγόριθμοι K-modes, ROCK

Υπάρχουν διάφορες προσεγγίσεις για την αντιμετώπιση του ζητήματος της συσταδοποίησης κατηγορικών δεδομένων. Οι χαρακτηριστικότεροι αλγόριθμοι που έχουν κατασκευαστεί έτσι, ώστε να διαχειρίζονται αποτελεσματικά τα κατηγορικά δεδομένα, είναι οι **k-modes** και **ROCK** (*RObust Clustering using linKs*).

Ο αλγόριθμος **k-modes** είναι ο πρώτος αλγόριθμος που δημοσιεύτηκε για τη διαχείριση ενός πολύ μεγάλου συνόλου δεδομένων με κατηγορικές μεταβλητές. Πρόκειται για μια επέκταση του γνωστού αλγορίθμου k-means που αφορά στη συσταδοποίηση συνεχών δεδομένων. Επομένως, είναι ένας διαιρετικός αλγόριθμος, όπως και ο k-means. Στόχος του είναι η ανακάλυψη συστάδων, ενώ υιοθετεί νέες έννοιες, όπως, την αντικατάσταση των κέντρων των

συστάδων με τα «modes». Επίσης, εισάγει ένα νέο μέτρο ανομοιότητας για την εξέταση των κατηγορικών δεδομένων.

Ο **ROCK** είναι ένας αντιπροσωπευτικός ιεραρχικός αλγόριθμος συσταδοποίησης κατηγορικών δεδομένων. Η νέα έννοια που εισάγει είναι τα «links» (δεσμοί ή σύνδεσμοι), ώστε να μετρήσει την ομοιότητα / εγγύτητα ανάμεσα σε ζεύγη σημείων. Έτσι, η μέθοδος συσταδοποίησης του ROCK βασίζεται σε μη μετρικά κριτήρια ομοιότητας, τα οποία είναι εφαρμόσιμα σε κατηγορικά σύνολα δεδομένων.

Ένα κοινό των δύο αυτών αλγορίθμων είναι, ότι απαιτούν ως παράμετρο εισόδου τον αριθμό των συστάδων. Όμως, ο ROCK δουλεύει εντελώς διαφορετικά απ' ότι ο k-modes, όχι μόνο επειδή είναι ιεραρχικός, αλλά επειδή δουλεύει και σε δείγματα των δεδομένων. Βέβαια, ο k-modes έχει καλύτερες δυνατότητες κλιμάκωσης από τον ROCK, ενώ το μόνο αρνητικό του ROCK είναι ότι τα αποτελέσματα βασίζονται –κατά πολύ- στη δειγματοληψία που διενεργεί. Τα αδύνατα σημεία του k-modes, είναι ότι δε μπορεί να χειριστεί το θόρυβο και τις έκτροπες παρατηρήσεις, καθώς και ότι δε μπορεί να χειριστεί συστάδες αυθαίρετου σχήματος.

Αξιολόγηση Συσταδοποίησης

Σύμφωνα με τα όσα προαναφέρθηκαν, δεν υπάρχει ένας μοναδικός αλγόριθμος ο οποίος να δίνει κάθε φορά την καλύτερη δυνατή συσταδοποίηση των διαθέσιμων αντικειμένων. Ανάλογα με τη φύση και τις απαιτήσεις του προβλήματος και των δεδομένων υπάρχει ένα πλήθος τεχνικών, οι οποίες μπορούν να οδηγήσουν σε λιγότερο ή περισσότερο ικανοποιητικά αποτελέσματα. Συνήθως, χρησιμοποιούνται δυο είδη τεχνικών για την αξιολόγηση μιας ομαδοποίησης, όπως :

- Η πρώτη τεχνική δεν απαιτεί αναφορά σε εξωτερική γνώση, δηλαδή, ο χωρισμός των αντικειμένων σε ομάδες γίνεται με τέτοιο τρόπο, ώστε τα αντικείμενα που ανήκουν στην ίδια ομάδα να μοιάζουν όσο το δυνατό περισσότερο μεταξύ τους, ενώ τα αντικείμενα που ανήκουν σε διαφορετικές ομάδες να διαφέρουν όσο το δυνατό περισσότερο το ένα από το άλλο.
- Η δεύτερη τεχνική απαιτεί εξωτερική γνώση, δηλαδή, γίνεται αξιολόγηση μιας συσταδοποίησης, αν γνωρίζουμε ήδη με κάποιο τρόπο σε ποια κατηγορία ανήκει ή σε ποια κατηγορία θα θέλαμε να ανήκει κάθε αντικείμενο.

Κανόνες Συσχέτισης

2.7 Εισαγωγή

Οι αλγόριθμοι εξαγωγής κανόνων συσχέτισης (association rules) είναι μία από τις σημαντικότερες τεχνικές εξόρυξης από δεδομένα. Είναι μια σύγχρονη μέθοδος, καθότι εμφανίστηκε μόλις το 1993 κι αναφερόταν στην εξαγωγή συσχετίσεων στα πεδία βάσεων δεδομένων. Οι πληροφορίες, που μπορούν να περιγραφούν και να συγκεντρωθούν από τους κανόνες συσχέτισης, είναι ιδιαίτερα σημαντικές και αφορούν πολλαπλές εφαρμογές. Το πιο χαρακτηριστικό παράδειγμα εφαρμογής των κανόνων συσχέτισης είναι η ανάλυση του «καλαθιού της νοικοκυράς» (market-basket analysis), όπου μια συναλλαγή, δηλαδή η αγορά των προϊόντων (για παράδειγμα το περιεχόμενο ενός καλαθιού υπεραγοράς), αντιμετωπίζεται σα μία μεμονωμένη συναλλαγή. Αναλύεται ένας αριθμός τέτοιων συναλλαγών, ώστε να εξαχθούν πρότυπα, τα οποία θα αναδείξουν τις αγοραστικές τάσεις των πελατών. Το κατάστημα μπορεί να χρησιμοποιεί τέτοιες πηγές πληροφοριών για διάφορους σκοπούς, όπως την προώθηση των προϊόντων, την τοποθέτηση των προϊόντων στα ράφια ενός καταστήματος και τη διαχείριση των αποθεμάτων.
Οι κανόνες συσχέτισης παρουσιάζονται με τη μορφή $X \rightarrow Y$, όπου τα X και Y είναι οι τιμές των πεδίων που παρουσιάζονται μέσα στους κανόνες. Ο κανόνας $X \rightarrow Y$ δείχνει ότι οι τιμές αυτές παρουσιάζονται μαζί μέσα στις εγγραφές. Με βάση τους κανόνες αυτούς, μετά από διαλογή, γίνεται η συσχέτιση των πεδίων κι ο ορισμός των προτύπων.

Οι κανόνες συσχέτισης μπορούν να θεωρηθούν τριών ειδών:

- ✓ Οι χρήσιμοι κανόνες συσχέτισης: περιέχουν πληροφορία υψηλής ποιότητας που μπορεί να χρησιμοποιηθούν και με την εύρεσή τους τις περισσότερες φορές μπορούν και να εξηγηθούν σχετικά εύκολα. Για παράδειγμα, ο κανόνας τις Πέμπτες, όσοι αγοράζουν πάνες, αγοράζουν και μπύρες, εξηγείται , ότι τα βράδια της Πέμπτης, τα νεαρά ζευγάρια ετοιμάζονται για το σαββατοκύριακο, αγοράζοντας πάνες για τα μωρά και μπύρα για τους πατέρες. Τοποθετώντας το ράφι με τις πάνες κοντά στα ράφια με τις μπύρες, μπορεί να αυξηθούν οι πωλήσεις και οι πελάτες δεν πρόκειται

να ξεχάσουν να αγοράσουν ούτε το ένα ούτε το άλλο προϊόν. Επίσης θα μπορούσαν να τοποθετηθούν και άλλα παρόμοια προϊόντα δίπλα στις πάνες, όπως πατατάκια κτλ.

- ✓ Οι ασήμαντοι κανόνες: είναι αυτοί που είναι ήδη γνωστοί στον καθένα που ασχολείται με μια επιχείρηση. Για παράδειγμα, ασήμαντος κανόνας μπορεί να είναι ο πελάτες που υπογράφουν συμφωνίες συντήρησης, συνήθως αγοράζουν μεγάλες οικιακές συσκευές. Γνωρίζουμε ότι κάποιος που αγοράζει μία μεγάλη οικιακή συσκευή θα υπογράψει και συμφωνία για τη συντήρησή της, αφού δεν υπάρχει και άλλος λόγος να υπογράψει κανείς συμφωνία συντήρησης. Εξ' άλλου οι συμφωνίες συντήρησης, σπάνια υπογράφονται ξεχωριστά από μία αγορά μεγάλης συσκευής. Οπότε, αυτός ο κανόνας αν και είναι έγκυρος είναι πρακτικά άχρηστος. Μια υποκατηγορία αυτών των κανόνων συσχέτισης είναι και οι κανόνες που προκύπτουν λόγω Marketing. Για παράδειγμα, κάποιος που κάνει νέα σύνδεση κινητής τηλεφωνίας, αγοράζει και συσκευή. Αυτή η τάση προκύπτει λόγω του τρόπου λειτουργίας, της αγοράς της κινητής τηλεφωνίας, με λίγα λόγια είναι κατευθυνόμενη. Τα αποτελέσματα των κανόνων συσχέτισης, μπορεί να είναι και μέτρα για την επιτυχία προηγουμένων τρόπων προωθήσεων των προϊόντων.

- ✓ Οι ανεξήγητοι κανόνες: είναι αυτοί που δεν έχουν εξήγηση αλλά ούτε προτείνουν και κάτι νέο. Για παράδειγμα, ο κανόνας σε ένα καινούργιο κατάστημα, το πιο κοινό προϊόν που πωλείται είναι το γάλα, δεν μας προσφέρει κάποια σημαντική πληροφορία για τη συμπεριφορά των πελατών, ούτε προτείνει και κάτι νέο στο κατάστημα. Με λίγα λόγια, το κατάστημα δεν μπορεί να κερδίσει με αυτό τον τρόπο. Ίσως περισσότερη έρευνα στο θέμα να μας δώσει νέες πληροφορίες, αλλά ο ίδιος ο κανόνας από μόνος του δεν μας προσφέρει κανενός είδους πληροφορία.

Κάθε κανόνας έχει δύο μετρικές, την υποστήριξη (support) και την εμπιστοσύνη (confidence). Αυτές οι μετρικές καθορίζουν το ποσοστό εφαρμογής του κανόνα στο σύνολο των εγγραφών. Η υποστήριξη μετράει ουσιαστικά την ισχύ του κανόνα, δηλαδή είναι το ποσοστό των συναλλαγών που περιέχουν το Y επί του αριθμού των συναλλαγών που περιέχουν το X. Η εμπιστοσύνη είναι το ποσοστό εμφάνισης του X και του Y μαζί στο σύνολο της βάσης δεδομένων, δηλαδή πόσο
συχνά συμβαίνει το πρότυπο αυτό στη βάση δεδομένων. Όσο πιο μεγάλοι είναι αυτοί οι αριθμοί, τόσο πιο «δυνατός» είναι ο κανόνας. Ο χρήστης πρέπει να καθορίσει την ελάχιστη

εμπιστοσύνη και υποστήριξη που επιθυμεί να έχει ο κανόνας. Εδώ πρέπει να σημειωθεί, ότι δεν υπάρχει κάποιος προκαθορισμένος αριθμός, που πρέπει να χρησιμοποιηθεί από τον χρήστη. Ο ίδιος, ανάλογα με το πρόβλημα που μελετά, τα δεδομένα που διαθέτει και το τι επιθυμεί να αναδείξει, θέτει την ελάχιστη εμπιστοσύνη και υποστήριξη, που κρίνει σωστή. Σίγουρα, θεωρείται πολύ σημαντική η εμπειρία που έχει ο χρήστης, ώστε να γίνει η σωστή επιλογή του κατώτατου ορίου. Η πολυπλοκότητα των αλγορίθμων και η δυσκολία επιλογής των χρήσιμων κανόνων, από το σύνολο των κανόνων που προκύπτουν, είναι βασικά προβλήματα, που αφορούν την εύρεση κανόνων συσχέτισης. Το πρώτο πρόβλημα αφορά τον αριθμό των κανόνων, ο οποίος αυξάνεται εκθετικά με τον αριθμό των πεδίων. Οι πιο πρόσφατοι αλγόριθμοι που εξάγουν κανόνες συσχέτισης, μπορούν να μειώσουν αποτελεσματικά τον αριθμό αυτό, με τον καθορισμό ενός κατώτατου ορίου στην εμπιστοσύνη και την υποστήριξη, που αφορά τη μέτρηση της ποιότητας των κανόνων. Το δεύτερο πρόβλημα αφορά τους χρήσιμους κανόνες, που συνήθως προκύπτουν και αποτελούν μόνο ένα μικρό ποσοστό του συνόλου των κανόνων. Το πρόβλημα αυτό ερευνάται σε σχέση με την υποστήριξη προς το χρήστη, όταν αναζητά κανόνες μέσα στους κανόνες που έχουν εξαχθεί, καθώς και με την ανάπτυξη επιπλέον μέτρων ποιότητας στους κανόνες.

2.7.1 Αλγόριθμοι Κανόνων συσχέτισης

Αλγόριθμος Apriori

Ο αλγόριθμος Apriori έρχεται να λύσει το πρόβλημα των πολλών λιστών που δημιουργούνται. Η εύρεση των μεγάλων λιστών βασίζεται στο ότι μία λίστα από προϊόντα είναι μεγάλη, αν κάθε υποσύνολό της είναι μεγάλη λίστα από δεδομένα. Η εύρεση αυτή γίνεται μετά από πολλές επαναλήψεις στη βάση δεδομένων. Κατά την πρώτη επανάληψη, υπολογίζεται η εμπιστοσύνη κάθε προϊόντος κι επίσης, ποια από αυτά είναι μεγάλες λίστες. Σε κάθε επόμενη επανάληψη, λαμβάνονται υπόψη μόνο οι μεγάλες λίστες από προϊόντα που είχαν βρεθεί στην προηγούμενη επανάληψη, χωρίς να λαμβάνονται υπόψη οι εγγραφές. Από τις νέες λίστες δημιουργούνται νέες υποψήφιες μεγάλες λίστες. Έπειτα μετράται η εμπιστοσύνη των λιστών

αυτών και καθορίζεται ποιες από αυτές είναι τελικά μεγάλες λίστες. Ο αλγόριθμος ξεκινάει πάλι, λαμβάνοντας υπόψη τις μεγάλες λίστες που καθορίστηκαν στην προηγούμενη επανάληψη. Αναλυτικά, τα βήματα του αλγορίθμου Apriori είναι:

1. Εύρεση των προϊόντων που έχουν εμπιστοσύνη μεγαλύτερη από την ελάχιστη εμπιστοσύνη, δηλαδή το σύνολο L1= μεγάλες λίστες από ένα προϊόν.
2. Από k=2 κι όσο το L_{k-1} δεν είναι κενό:
3. α) εύρεση του συνόλου C_k των υποψηφίων μεγάλων λιστών από k προϊόντα με βάση το L_{k-1}

   β) εύρεση της εμπιστοσύνης των υποψήφιων μεγάλων λιστών και δημιουργία συνόλου L_k = μεγάλες λίστες από k προϊόντα.
4. Για κάθε στοιχείο των $L_1 \ldots L_n$ εύρεση εκείνων που έχουν υποστήριξη μεγαλύτερη από την ελάχιστη υποστήριξη.

Στο πρώτο βήμα, ο αλγόριθμος μετράει την εμπιστοσύνη του κάθε προϊόντος ξεχωριστά, ώστε να σχηματιστούν οι μεγάλες λίστες μεγέθους ενός προϊόντος. Στο δεύτερο βήμα (που αποτελείται από δύο υπό-βήματα) μεγάλες λίστες από k-1 προϊόντα, που βρέθηκαν στην προηγούμενη επανάληψη, χρησιμοποιούνται για να δημιουργηθούν οι υποψήφιες μεγάλες λίστες από k προϊόντα (C_k). Έπειτα, υπολογίζεται η εμπιστοσύνη των υποψηφίων μεγάλων λιστών από k προϊόντα. Το βήμα αυτό τερματίζεται, όταν δεν υπάρχουν υποψήφιες μεγάλες λίστες. Τέλος, στο τρίτο βήμα, υπολογίζεται η υποστήριξη κάθε μεγάλης λίστας προϊόντων κι εξάγονται κανόνες, από τους οποίους γίνονται αποδεκτοί εκείνοι που έχουν υποστήριξη μεγαλύτερη από την ελάχιστη υποστήριξη.

Αλγόριθμος AprioriTID

Ο αλγόριθμος AprioriTID είναι μία παραλλαγή του βασικού αλγορίθμου Apriori. Σε αυτόν τον αλγόριθμο η βάση δεδομένων χρησιμοποιείται στην αρχή. Μετά την πρώτη επανάληψη δεν χρησιμοποιείται για υπολογισμό της εμπιστοσύνης των υποψήφιων μεγάλων λιστών, αλλά γίνεται χρήση μίας κωδικοποίησης των υποψηφίων μεγάλων λιστών, που είχε χρησιμοποιηθεί στην προηγούμενη επανάληψη. Σε επόμενες επαναλήψεις, το μέγεθος της κωδικοποίησης αυτής μπορεί να γίνει πολύ μικρότερο από τον αριθμό των συναλλαγών στη βάση δεδομένων.

```
1. L1 = {large 1-itemsets}
2. Ĉ1 = database D;
3. for (k=2; Lk-1≠Ø; k++) do begin
4. Ck = apriori-gen(Lk-1); // δημιουργία υποψηφίων
5. Ĉk = Ø;
6. forall entries t ∈ Ĉk-1 do begin // εύρεση των υποψηφίων στο Ck
   που περιέχονται στη συναλλαγή t με αναγνωριστικό t.TID
7.   Ct = {c ∈ Ck | (c-c[k]) ∈ t.set_of_itemsets ^ (c-c[k-1]) ∈ t.set_of_itemsets};
8.   forall candidates c ∈ Ct do
9.     c.count++;
10.    if (Ct≠Ø) then Ck+= <t.TID, Ct>;
11.  end
12.  Lk = {c ∈ Ck | c.count ≥ minsup}
13. end
14. return UkLk;
```

Εικόνα 13: Κώδικας Αλγορίθμου AprioriTID

Μία επέκταση των Apriori και AprioriTID είναι ο αλγόριθμος AprioriHybrid, που συνδυάζει τα καλύτερα χαρακτηριστικά και των δύο αλγορίθμων. Συγκεκριμένα, ενώ ο Apriori στις αρχικές επαναλήψεις δίνει γρηγορότερα αποτελέσματα, ο AprioriTID δίνει γρηγορότερα αποτελέσματα σε μεταγενέστερες επαναλήψεις. Έτσι, ο AprioriHybrid χρησιμοποιεί τον Apriori σε αρχικά στάδια και στη συνέχεια χρησιμοποιεί τον AprioriTID. Η αλλαγή αυτή από τον έναν αλγόριθμο στον άλλον, φυσικά, συμπεριλαμβάνει ένα κόστος.

Αλγόριθμος PARTITION

Ο αλγόριθμος αυτός χωρίζει τη βάση δεδομένων σε μικρά κομμάτια, έτσι ώστε να μπορούν να αναλύονται ξεχωριστά και αποτελεσματικά, προκειμένου να βρεθούν οι μεγάλες λίστες. Οι μεγάλες αυτές λίστες, στη συνέχεια, συνδυάζονται, ώστε να δημιουργηθούν οι υποψήφιες μεγάλες λίστες. Όμως, χρειάζεται ακόμη ένα επιπλέον σάρωμα της βάσης, για να επιβεβαιωθεί ότι οι τοπικές μεγάλες λίστες από προϊόντα είναι επίσης και ολικές.

Αλγόριθμος FP-growth

Είναι βασισμένος σε ένα πρόθεμα αναπαράστασης δέντρου της βάσης δεδομένων (που ονομάζεται FP-tree), το οποίο μπορεί να σώσει μεγάλα ποσά μνήμης για την αποθήκευση των συναλλαγών. Η βασική ιδέα του αλγορίθμου FP-growth μπορεί να περιγραφεί, ως ένα επαναληπτικό σύστημα εξουδετέρωσης: σε ένα στάδιο προεπεξεργασίας διαγράφει όλα τα στοιχεία από τις συναλλαγές, που δεν είναι συχνά σε ατομική βάση, δηλαδή, δεν εμφανίζονται με βάση ενός ελαχίστου αριθμού συναλλαγών, που καθορίζει ο χρήστης. Στη συνέχεια, επιλέγει όλες εκείνες τις συναλλαγές, που περιέχουν το λιγότερο συχνά αντικείμενο και διαγράφει αυτό το στοιχείο από αυτές. Επιστρέφει, για να επεξεργαστεί την νέα μειωμένη βάση δεδομένων . Στην επιστροφή, αφαιρεί το επεξεργασμένο στοιχείο επίσης από τη βάση δεδομένων όλων των συναλλαγών κι αρχίζει πέρα από την αρχή, δηλαδή, επεξεργάζεται το δεύτερο συχνό στοιχείο.

Ένας επιπλέον αλγόριθμος, που είναι βασισμένος σε ένα πρόθεμα αναπαράστασης της βάσης δεδομένων είναι ο **Eclat** ο οποίος χρησιμοποιεί την κατά βάθος αναζήτηση στο δέντρο κλάσεων. Επιπλέον, ο χωρισμός του πίνακα σε τμήματα δεν είναι αναγκαίος αφού το κόστος στη μνήμη είναι χαμηλότερο.

2.7.2 Ποσοτικοί Κανόνες Συσχέτισης

Μια λύση είναι η κατάτμηση του συνόλου τιμών των ποσοτικών γνωρισμάτων σε διαστήματα κι η δημιουργία λογικών γνωρισμάτων που να έχουν την μορφή <γνώρισμα_1, διάστημα_1>. Το πρόβλημα συνεπώς ανάγεται στην εύρεση της κατάλληλης κατάτμησης.

Υπάρχουν δυο θέματα για τα οποία πρέπει να δοθεί η βέλτιστη λύση.

- «Minsup». Αν ο αριθμός των διαστημάτων για ένα ποσοτικό γνώρισμα είναι μεγάλος, τότε η υποστήριξη ενός διαστήματος μπορεί να είναι μικρή. Κατά συνέπεια, αν δε χρησιμοποιηθούν μεγαλύτερα διαστήματα, μπορεί να μην παραχθούν κάποιοι κανόνες για αυτό το γνώρισμα, καθώς, δεν θα έχουν την ελάχιστη υποστήριξη.
- «Minconf». Υπάρχει κόστος που συνεπάγεται ο χωρισμός των τιμών σε διαστήματα. Η πληροφορία χάνεται όσο το μέγεθος των διαστημάτων μεγαλώνει, καθώς, μερικοί

κανόνες μπορεί να έχουν ελάχιστη εμπιστοσύνη (minimum confidence), μόνο όταν το πρώτο μέλος αποτελείται από μικρό διάστημα (μικρή εμπιστοσύνη).

Η λύση που προτάθηκε, είναι η κατάτμηση να μπορεί να δημιουργήσει μικρά διαστήματα, τα οποία, στη συνέχεια, ενώνονται, για να φτιάξουν καινούρια, μεγαλύτερα και με μεγαλύτερη υποστήριξη.

2.7.3 Αντιπροσωπευτικοί Κανόνες Συσχέτισης

Το σύνολο όλων των κανόνων συσχέτισης, που ικανοποιούν τις απαιτήσεις για ελάχιστη υποστήριξη ***s*** κι ελάχιστη εμπιστοσύνη ***c,*** θα το αποκαλούμε εν συντομία AR(s,c). Εάν τα s και c εννοούνται, τότε μπορούμε να γράφουμε απλά AR.

Η κάλυψη C ενός κανόνα $X \Rightarrow Y$, ορίζεται ως εξής:

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

<u>Ιδιότητες που σχετίζονται με τον τελεστή κάλυψης</u>

- Έστω r ένας κανόνας συσχέτισης με υποστήριξη s και εμπιστοσύνη c. Κάθε κανόνας r΄ που ανήκει στην κάλυψη C(r) είναι ένας κανόνας συσχέτισης που έχει υποστήριξη όχι μικρότερη από s κι εμπιστοσύνη όχι μικρότερη από c. Η άμεση συνέπεια της ιδιότητας αυτής είναι, ότι αν ένας κανόνας r ανήκει στο AR(s,c), τότε κάθε κανόνας r από το C(r) θα ανήκει επίσης στο AR(s,c).

- Έστω δύο κανόνες συσχέτισης r : $X \Rightarrow Y$ και r΄= $(X' \Rightarrow Y')$. Τότε ο r θα ανήκει στην κάλυψη του r΄ C(r΄) αν και μόνο αν $X \cup Y \subseteq X' \cup Y'$ και $X \supseteq X'$. Δηλαδή $r \in C(r') \Leftrightarrow X \cup Y \subseteq X' \cup Y' \wedge X \supseteq X'$.

  - (i) Αν ένας κανόνας συσχέτισης r είναι μεγαλύτερος (περιέχει περισσότερα αντικείμενα) από έναν κανόνα συσχέτισης r΄, τότε $r \notin C(r')$.

  (ii) Αν ένας κανόνας συσχέτισης r : $(X \Rightarrow Y)$ είναι μικρότερος από έναν κανόνα συσχέτισης r΄: $(X' \Rightarrow Y')$, τότε $r \in C(r')$ αν και μόνο αν $X \cup Y \subseteq X' \cup Y'$ και $X \supseteq X'$.

(iii) Αν r : (X⇒Y) και r΄: (X΄⇒Y΄) είναι διαφορετικοί κανόνες συσχέτισης με το ίδιο μήκος (ίδιος αριθμός από αντικείμενα), τότε r ∈ C(r΄), αν και μόνο αν X∪Y = X΄∪Y΄ και X ⊃ X΄.

Παραγωγή Αντιπροσωπευτικών Κ.Σ.

Ιδιότητα 1

Έστω ∅ ≠ **X**⊂ **Z** ⊆ **I** και r είναι ένας κανόνας της μορφής r : (X⇒Y) ∈ AR(s, c) . Τότε ο κανόνας αυτός θα ανήκει στο RR(s,c), αν ισχύουν οι δυο επόμενες προϋποθέσεις.

(i) maxSup ≤ s ή maxSup|sup(**X**) < c, όπου

sup maxSup = max**({sup(Z΄) | Z** ⊂ **Z΄ (I}** ∪**{0})**

(ii) ¬∃ **X΄**, ∅ ≠ **X΄**⊂ **X** τέτοιο, ώστε **(X΄⇒Z\X΄)** ∈ **AR(s,c)**

Η πρώτη συνθήκη εξασφαλίζει ότι ο κανόνας r δεν βρίσκεται στην κάλυψη κάποιου κανόνα μεγαλύτερου από τον r. Η δεύτερη συνθήκη εξασφαλίζει ότι ο κανόνας r δε βρίσκεται στην κάλυψη κάποιου κανόνα με μήκος ίσο με τον r.

Ιδιότητα 2

Έστω ∅ ≠ **Z** ⊂ **Z΄** ⊆ **I**. Αν sup(**Z**) = sup(**Z΄**), τότε κανένας κανόνας της μορφής **(X⇒Z\X)** ∈ **AR(s,c)** με ∅ ≠ **X**⊂ **Z** δεν ανήκει στο RR (s,c).

```
1.  procedure FastGenAllRepresentatives(all frequent
    itemsets L);
2.  forall Z ∈ F do begin
3.  k = |Z|; maxSup = max({sup(Z')|Z⊂Z'∈F_{k+1}}∪{0});
4.  if Z.sup ≠ maxSup then begin
5.  A_1 = {{Z[1]}, {Z[2]}, ... , {Z[k]}};
6.  for (i=1; (A_i.0) and (i<k); i++) do begin
7.  forall X∈A_i do begin
8.  find Y∈L_i such that Y=X;
9.  XCount = Y.count;
10. // is X⇒Z\X an association rule?
11. if (Z.count/XCount.c) then begin
12. // aren't there any rep. rules longer that X⇒Z\X
13. if (maxSup/XCount<c) then
14. print(X, «⇒», Z\X, «with support: «, Z.count, « and
    confidence: «, Z.count/XCount);
15. // antecedents of ass. rules are not extended
16. A_i = A_i\{X};
17. endif
18. endfor
19. A_{i+1} = AprioriGen(A_i);
20. endfor
21. endif
```

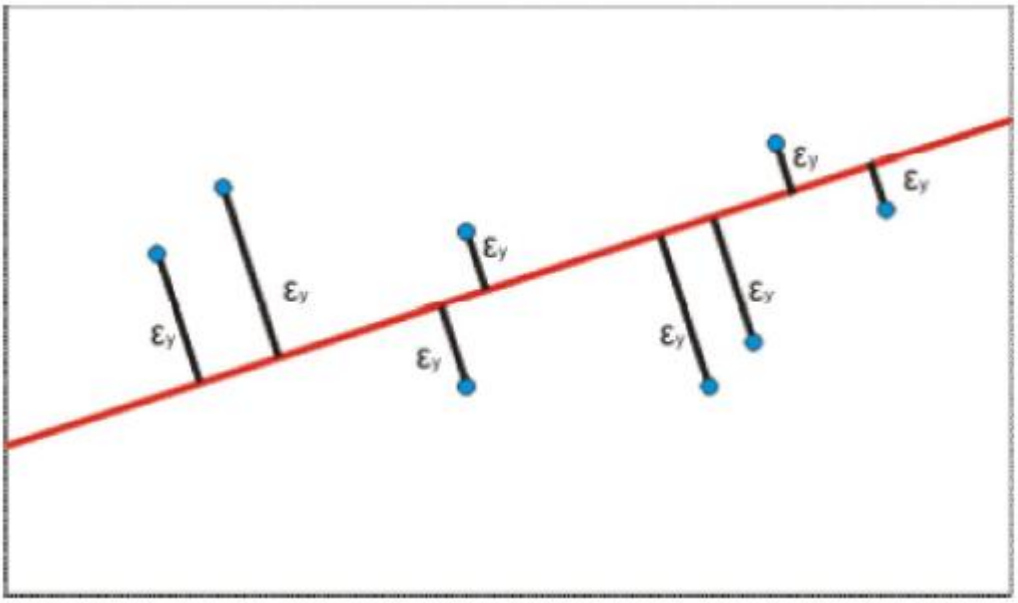
Εικόνα 14: Παραγωγή Αντιπροσωπευτικών Κανόνων Συσχέτισης

Παλινδρόμηση

2.8 Γραμμική Παλινδρόμηση

Η ανάλυση παλινδρόμησης είναι μια τεχνική, που χρησιμοποιείται για την μοντελοποίηση και την ανάλυση αριθμητικών δεδομένων μιας εξαρτημένης μεταβλητής και κάποιων ανεξάρτητων μεταβλητών. Το μοντέλο είναι μια συνάρτηση συσχέτισης της εξαρτημένης μεταβλητής από τις ανεξάρτητες. Η μοντελοποίηση μπορεί να γίνει χωρίς προηγούμενη γνώση για τον τρόπο με τον οποίο συνδέεται η εξαρτημένη μεταβλητή από τις ανεξάρτητες

και τότε ονομάζεται «εμπειρική μοντελοποίηση». Στην γραμμική παλινδρόμηση, η απαίτηση του μοντέλου που θα παραχθεί είναι: η εξαρτημένη μεταβλητή yi να είναι ένας γραμμικός συνδυασμός των ανεξαρτήτων μεταβλητών.



Εικόνα 15: Γραμμική Παλινδρόμηση

2.8.1 Απλή Γραμμική Παλινδρόμηση

Το μοντέλο της απλής γραμμικής παλινδρόμησης έχει τη μορφή

yi = β0 + β1xi + εi, i= 1,m

όπου:

- xi η ανεξάρτητη μεταβλητή,
- β0,β1 δύο παράμετροι,
- εi το σφάλμα της πρόβλεψης.

Ο στόχος λοιπόν, είναι να βρεθούν οι κατάλληλες παράμετροι, που θα ελαχιστοποιήσουν τη συνάρτηση του τετραγώνου του σφάλματος, ο οποίος δίνεται από τον τύπο:

$$SSE = \sum_{i=1}^{N} e_i^2$$

Τυπολόγιο παραμέτρων απλής γραμμικής παλινδρόμησης

$$\beta_1 = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sum (x_i - \bar{X})^2}$$

$$\bar{\sigma}_\varepsilon = \sqrt{\frac{SSE}{N-2}}$$

όπου X , ο μέσος όρος της ανεξάρτητης μεταβλητής και Y , ο μέσος όρος των τιμών y.

Υπό την προϋπόθεση ότι το σφάλμα έχει μία σταθερή απόκλιση, η εκτίμηση αυτής της απόκλισης δίνεται από τον τύπο:

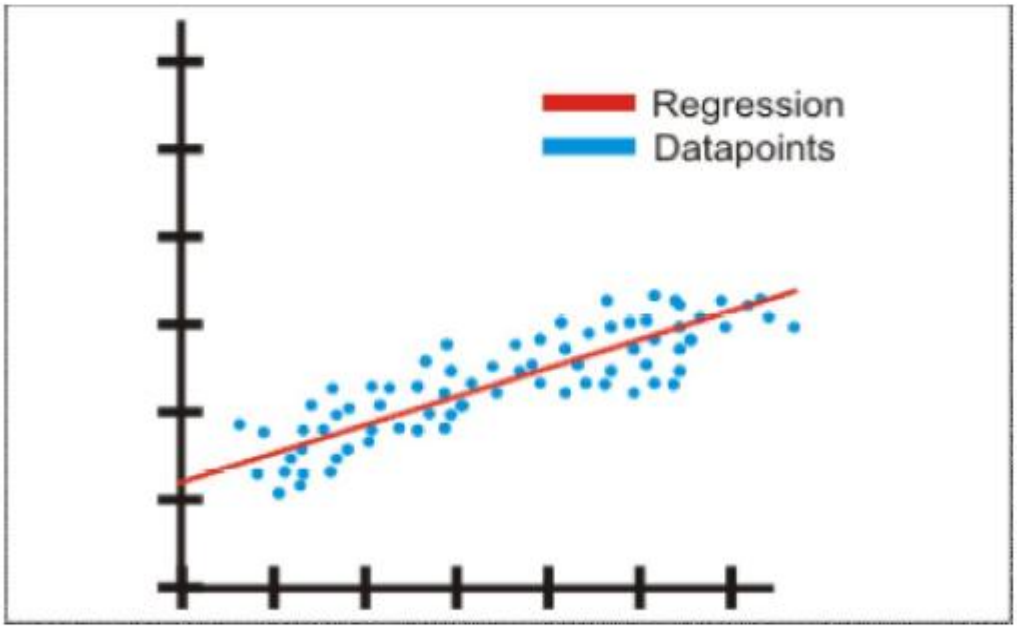
$$\bar{\sigma}_\varepsilon = \sqrt{\frac{SSE}{N-2}}$$

κι ονομάζεται τυπικό σφάλμα της πρόβλεψης.

Το τυπικό σφάλμα της εκτίμησης των παραμέτρων δίνεται από :

$$\bar{\sigma}_{\beta_0} = \bar{\sigma}_\varepsilon \sqrt{\frac{1}{N} + \frac{\bar{X}^2}{\sum (x_i - \bar{x})^2}}$$

$$\bar{\sigma}_{\beta_1} = \bar{\sigma}_\varepsilon \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}$$



Εικόνα 16: Απλή Γραμμική Παλινδρόμηση

Αξίζει να σημειωθεί πως τη λογική της Απλής Γραμμικής Παλινδρόμησης ακολουθεί κι ο αλγόριθμος **Least Median Square**, έχοντας όμως ως στόχο την ελαχιστοποίηση του σφάλματος της διαμέσου.

2.8.2 Πολλαπλή Γραμμική Παλινδρόμηση

Όταν το πρόβλημα αποτελείται από p ανεξάρτητες μεταβλητές τότε το μοντέλο θα έχει την εξής μορφή:

$$Y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_{pi} x_{pi} + \varepsilon_i,\ i = 1,m$$

Συνεπώς, ο τύπος του σφάλματος θα είναι της παρακάτω μορφής:

$$\varepsilon_i = Y_i \beta_0 - \beta_1 x_i - \ldots - \beta_{pi} x_{pi}$$

ενώ οι εξισώσεις των παραμέτρων διαμορφώνονται ως εξής:

$$\sum_{i=1}^{N} \sum_{k=i}^{p} X_{ij} X_{ik} \overline{\beta}_{\kappa} = \sum_{i=1}^{N} X_{ij} y_i, \quad j = 1, p$$

2.8.3 Pace Regression

Η συγκεκριμένη μοντελοποίηση ακολουθεί τη λογική της Γραμμικής Παλινδρόμησης, όμως, είναι περισσότερο κατάλληλη για την περίπτωση που οι ανεξάρτητες μεταβλητές είναι πάρα πολλές και τείνουν στο άπειρο. Τις αξιολογεί γρηγορότερα κι απορρίπτει τις ανεξάρτητες μεταβλητές, που δεν μπορούν να αξιοποιηθούν στην τελική πρόβλεψη. Ο αλγόριθμος αυτός δε μπορεί να διαχειριστεί

- Ελλιπή στιγμιότυπα ως προς τις ανεξάρτητες μεταβλητές
- Μη αριθμητικές τιμές

2.9 Support Vector Machines

Οι αλγόριθμοι **Support Vector** λειτουργούν ικανοποιητικά, κατασκευάζοντας αρκετά πολύπλοκα θεωρητικά μοντέλα, αλλά ταυτόχρονα με πολύ ικανοποιητικά απλή μαθηματική ανάλυση, καθώς, μπορούν να αντιστοιχίσουν ένα πολυδιάστατο μη γραμμικό χώρο σε μια απλή γραμμική συνάρτηση.

Τα **Support Vector** ουσιαστικά, είναι όλα εκείνα τα σημεία, που βρίσκονται κοντά στην επιφάνειας της συνάρτησης. Στην παλινδρόμηση στόχος παραμένει η εύρεση της συνάρτησης που προσεγγίζει τα σημεία εκπαίδευσης μέσω της ελαχιστοποίησης του σφάλματος πρόβλεψης.

Οι πιο αντιπροσωπευτικοί αλγόριθμοι είναι:

SMOreg

Ο αλγόριθμος **SMOreg** υλοποιεί τον διαδοχικού ελάχιστου αλγόριθμο βελτιστοποίησης για την εκπαίδευση ενός μοντέλου παλινδρόμησης με **support vector**. Αυτή η υλοποίηση

αντικαθιστά παντού όλες τις απούσες τιμές και μετατρέπει τις μη αριθμητικές ανεξάρτητες μεταβλητές σε δυαδικές, ενώ επίσης ομαλοποιεί εξ ορισμού όλα τα γνωρίσματα.

SVMreg

Ο **SVMreg** υλοποιεί τους **Support Vector Machine** για παλινδρόμηση, αλλά με εξελιγμένο αλγόριθμο βελτιστοποίησης και με περισσότερες επιλογές στην επιλογή της κατάλληλης συνάρτησης.

REPTree

Τα δέντρα απόφασης εκπαίδευσης αποτελούν μια κοινή μέθοδος της εξόρυξης γνώσης. Κάθε κόμβος αντιστοιχεί σε μία ανεξάρτητη μεταβλητή, ενώ κάθε κλάδος του σε μια πιθανή τιμή ή ένα πιθανό εύρος τιμών αυτής. Τέλος, το κάθε φύλλο αντιπροσωπεύει μία τιμή της εξαρτημένης μεταβλητής. Το δέντρο εκπαιδεύεται χωρίζοντας τις συστάδες σε υποσυστάδες, ανάλογα με την τιμή μιας ανεξάρτητης μεταβλητής κι αυτό επαναλαμβάνεται για κάθε υποσύνολο για όλες τις ανεξάρτητες μεταβλητές. Ο **REPTree** είναι ταχύς εκπαιδευτής δέντρου απόφασης. Δημιουργεί ένα δέντρο παλινδρόμησης αξιοποιώντας πληροφορία που εξάγεται από στοιχεία των ανεξάρτητων μεταβλητών.

M5P

Επίσης, κι ο **M5P** είναι αλγόριθμος δέντρου παλινδρόμησης που ακολουθεί την στρατηγική του «διαίρει και βασίλευε». Σε κάθε επανάληψη δημιουργεί ένα δέντρο και κρατάει το φύλλο, που δίνει τα καλύτερα αποτελέσματα.

Isotonic Regression

Ο αλγόριθμος **isotonic regression** αφορά την εύρεση σταθμισμένων leastsquares, κατάλληλου $x \in \Re^n$ σε ένα διάνυσμα $a \in \Re^n$ με το διάνυσμα βαρών $\omega \in \Re^n$, σε ένα σύνολο περιορισμών μονοτονίας, που οργανώνουν μερικώς τις μεταβλητές. Οι περιορισμοί

μονοτονίας καθορίζουν μια κατευθυνόμενη ακυκλική γραφική παράσταση (directed acyclic graph) G = (N,E) στους κόμβους ***N=1,2,…,n,*** που αντιστοιχούν στις μεταβλητές ***x=x1,x2,…,xn*** . Κατά συνέπεια, το πρόβλημα IR, όπου μια απλή διαταγή καθορίζεται, αντιστοιχεί στο ακόλουθο quadric programmers (QP):

$$\min \sum_{i=1}^{n} w_i (x_i - a_i)^2 \quad subject\ to\ x_i \geq x_j \forall (i,j) \in E$$

Στην περίπτωση όπου G = (N,E) είναι σε πλήρη διάταξη, ένας απλός επαναληπτικός αλγόριθμος καλείται για αυτό το QP πρόβλημα.

Multilayer perceptrons

Ο αλγόριθμος αυτός είναι ένα ανατροφοδοτούμενο νευρωνικό δίκτυο, το οποίο χαρτογραφεί τα σύνολα των δεδομένων εισόδου, (δηλαδή τις ανεξάρτητες μεταβλητές), από ένα κατάλληλο σύνολο της εξόδου (εξαρτημένη μεταβλητή). Χρησιμοποιεί τρία ή περισσότερα στρώματα των νευρώνων (κόμβοι), με μη γραμμικές συναρτήσεις ενεργοποίησης, κι είναι ισχυρότερο από αυτό, καθώς, μπορεί να διακρίνει τα στοιχεία που δεν είναι γραμμικώς διαχωριζόμενα ή διαχωριζόμενα από πολυδιάστατο επίπεδο.

K- nearest neighbor, K star, LWL

Στους παραπάνω αλγόριθμους η γενίκευση και η ερμηνεία των δεδομένων εκμάθησης πραγματοποιείται, αφού τεθεί το ερώτημα. Ουσιαστικά, δημιουργούν το μοντέλο πρόβλεψης ανάλογα με το κάθε πρόβλημα που τίθεται.

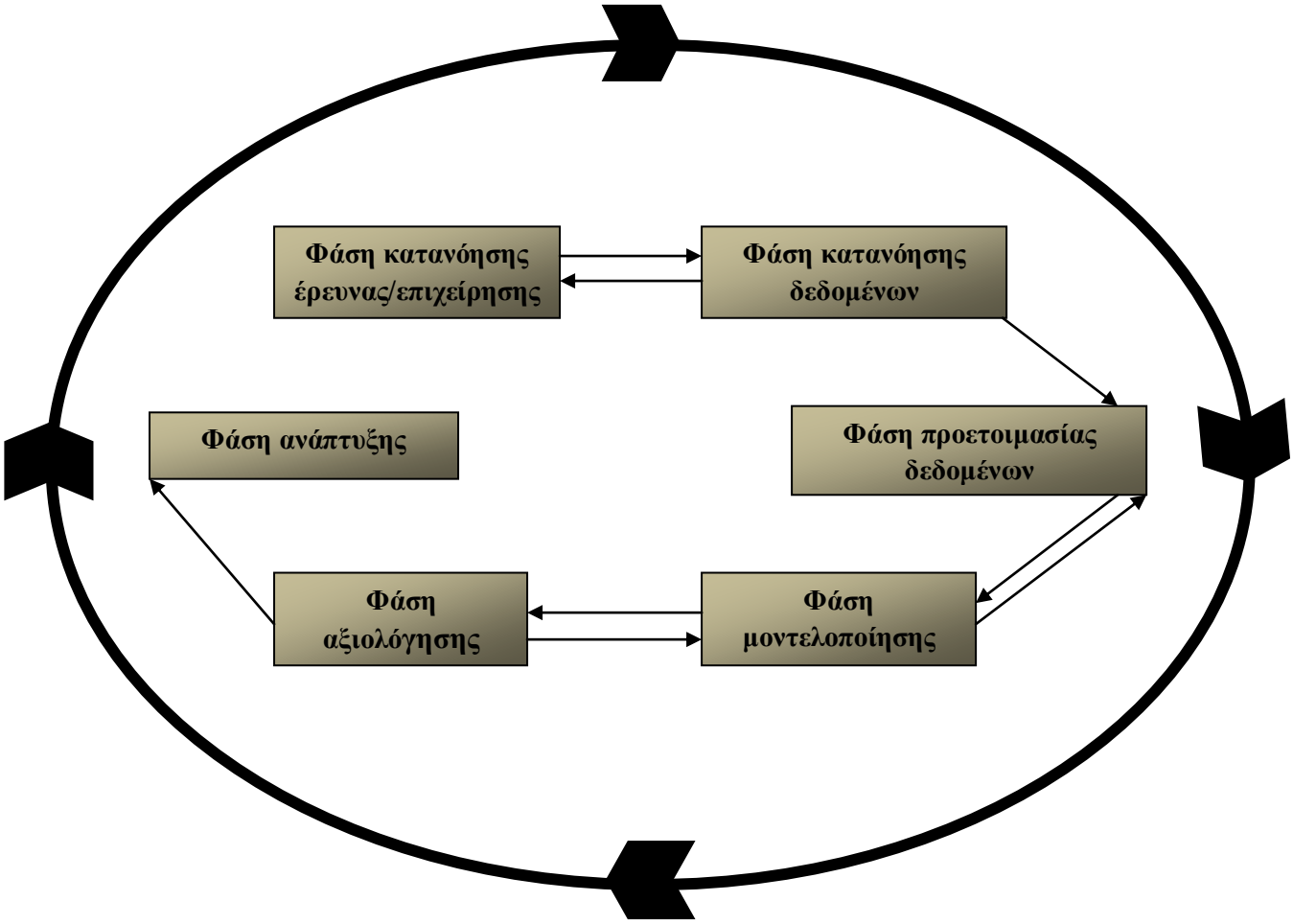
3ο Κεφάλαιο

Εφαρμογές εξόρυξης δεδομένων

Το CRISP-DM (Cross-Industry Standard Process for Data Mining) είναι ένα ουδέτερο εργαλείο που αναπτύχθηκε το 1996 από τους αναλυτές που εκπροσωπούν την DaimlerChrysler, την SPSS, και την NCR. Είναι ελεύθερα διαθέσιμο και χρησιμοποιείται για την εφαρμογή εξόρυξης δεδομένων στη γενική στρατηγική επίλυσης προβλημάτων μιας επιχειρηματικής μονάδας. Σύμφωνα με το CRISP-DM, ένα συγκεκριμένο σχέδιο εξόρυξης δεδομένων έχει έναν κύκλο ζωής που αποτελείται από έξι φάσεις, οι οποίες είναι:

1. Φάση κατανόησης έρευνας/επιχείρησης.
2. Φάση κατανόησης δεδομένων.
3. Φάση προετοιμασίας δεδομένων.
4. Φάση μοντελοποίησης.
5. Φάση αξιολόγησης.
6. Φάση ανάπτυξης.

Αξίζει να σημειωθεί ότι η ακολουθία των φάσεων είναι προσαρμοστική. Δηλαδή, η επόμενη φάση στην ακολουθία εξαρτάται συχνά από τα αποτελέσματα που συνδέονται με την προηγούμενη φάση. Οι πιο σημαντικές εξαρτήσεις μεταξύ των φάσεων υποδεικνύονται από τα βέλη, όπως απεικονίζεται στο παρακάτω σχήμα.



Εικόνα 17: Η επαναληπτική, προσαρμοστική διαδικασία CRISP-DM

3.1 Μελέτη 1η – Εφαρμογή Κανόνων συσχέτισης και Κατηγοριοποίησης

«Ανάλυση διεκδικήσεων ασφάλειας αυτοκίνητου»

Η ποιοτική ασφάλεια συνεχίζει να είναι μια προτεραιότητα των κατασκευαστών αυτοκινήτων, συμπεριλαμβανομένης της DaimlerChrysler. Ο Jochen Hipp του πανεπιστημίου Tubingen, της Γερμανίας, και ο Guido Lindner DaimlerChrysler στη Γερμανία, διερεύνησαν το μοτίβο των διεκδικήσεων ασφάλειας για τα αυτοκίνητα της DaimlerChrysler.

1. Φάση κατανόησης της επιχείρησης

Ο αντικειμενικός σκοπός της DaimlerChrysler είναι να μειωθούν οι δαπάνες που συνδέονται με τις διεκδικήσεις βελτιώνοντας ,έτσι, την ικανοποίηση των πελατών. Μέσω των συνομιλιών με τους μηχανικούς εργοστασίων, οι οποίοι είναι οι τεχνικοί με εμπειρία στην κατασκευή οχημάτων, οι ερευνητές είναι σε θέση να διατυπώσουν τα συγκεκριμένα επιχειρησιακά προβλήματα, όπως τα εξής:

- Υπάρχουν αλληλεξαρτήσεις μεταξύ των αξιώσεων εγγύησης;
- Οι προηγούμενες αξιώσεις εγγύησης συνδέονται με τις παρόμοιες αξιώσεις στο μέλλον;
- Υπάρχει συσχετισμός μεταξύ ενός ορισμένου τύπου διεκδίκησης και ενός συγκεκριμένου γκαράζ;

Το σχέδιο είναι να εφαρμοστούν κατάλληλες τεχνικές εξόρυξης δεδομένων για να προσπαθήσουν να αποκαλύψουν τέτοιες και άλλες πιθανές σχέσεις.

2.Φάση κατανόησης δεδομένων.

Οι ερευνητές χρησιμοποιούν το σύστημα πληροφοριών ποιότητας της DaimlerChrysler το όποιο περιέχει πληροφορίες για πάνω από 7000000 οχήματα και έχει μέγεθος περίπου 40 gigabytes .Το σύστημα περιέχει λεπτομέρειες παράγωγης για το πως κ που κατασκευάστηκε ένα όχημα συμπεριλαμβανόμενων περίπου 30 κ πάνω κώδικες πωλήσεων για κάθε οχήματα επίσης συμπεριλαμβάνει πληροφορίες διεκδικήσεων εγγύησης τις όποιες παρέχουν τα γκαράζ με τη μορφή μιας η κ περισσότερων από 5000 πιθανών αιτίων. Οι ερευνητές τόνισαν το γεγονός ότι η βάση δεδομένων ήταν τελείως ακατανόητη σε μη ειδικούς του χώρου δραστηριότητας έτσι οι ειδικοί από διαφορά τμήματα έπρεπε να τοποθετηθούν και να τους ζητηθεί η συμβουλή σε μια σύντομη αποστολή που αποδείχτηκε μάλλον δαπανηρή. Τονίζουν ότι οι αναλυτές δε θα πρέπει να υποτιμούν τη σημαντικότητα ,τη δυσκολία και το πιθανό κόστος αυτής της πρώτης φάσης της διαδικασίας άντλησης δεδομένων και ότι οι προχειρότητες εδώ ίσως οδηγήσουν σε ακριβές επαναλήψεις της ροής πληροφοριών.

3.Φάση προετοιμασίας δεδομένων

Οι ερευνητές διαπίστωσαν ότι αν και σχεσιακή, η βάση δεδομένων QUIS είχε περιορισμένη πρόσβαση στην SQL.Αυτοί χρειάστηκε να επιλέξουν τις υποθέσεις και τις μεταβλητές χειρωνακτικά και μετά χειρωνακτικά να αντλήσουν νέες μεταβλητές που θα μπορούσαν να χρησιμοποιηθούν για τη φάση μοντελοποίησης. Για παράδειγμα η μεταβλητή *αριθμός ήμερων από την ημερομηνία πώλησης μέχρι την πρώτη διεκδίκηση* έπρεπε να αντληθεί από τις κατάλληλες ιδιότητες της ημερομηνίας. Μετά στραφήκαν σε λογισμικό άντλησης ιδιοκτησιακών δεδομένων το οποίο είχε χρησιμοποιηθεί στην DaimlerChrysler σε προηγούμενα προγράμματα όπου και έτρεξαν τα δεδομένα. Το αποτέλεσμα ήταν η αναλυτική προεπεξεργασία των δεδομένων για να μετατρέψουν τις ιδιότητες σε μια μορφή χρησιμοποιήσιμη για μοντέλα αλγορίθμων. Οι ερευνητές αναφέρουν ότι η φάση προετοιμασίας δεδομένων πηρέ περισσότερο χρόνο από όσο είχαν σχεδιάσει.

4.Φάση μοντελοποίησης

Αφού το γενικό πρόβλημα της επιχείρησης από την πρώτη φάση ήταν να ερευνήσουν την εξάρτηση ανάμεσα στις διεκδικήσεις εγγύησης οι ερευνητές επέλεξαν να εφαρμόσουν τις παρακάτω τεχνικές:

- Τα δίκτυα Bayesian
- Κανόνες Συσχέτισης

Τα δίκτυα Bayesian διαμορφώνουν την αβεβαιότητα με το να παρουσιάζουν κατηγορηματικά τις εξαρτήσεις ανάμεσα σε διαφορά μέρη παρέχοντας έτσι γραφική οπτικοποίηση των σχέσεων εξάρτησης ανάμεσα στα στοιχεία. Έτσι τα δίκτυα Bayesian αντιπροσωπεύουν μια φυσική επιλογή για τη μοντελοποίηση της εξάρτησης ανάμεσα στις διεκδικήσεις ασφάλειας. Οι κανόνες συσχέτισης είναι, επίσης, ένας φυσικός τρόπος για να ερευνήσουμε την εξάρτηση ανάμεσα σε διεκδικήσεις εγγύησης αφού το μέτρο σιγουριάς αντιπροσωπεύει έναν τύπο πιθανότητας υπό συνθήκες παρόμοιο με τα δίκτυα Bayesian.Οι λεπτομέρειες των αποτελεσμάτων είναι εμπιστευτικές άλλα μπορούμε να πάρουμε μια ιδέα του τύπου των εξαρτήσεων που αποκαλύπτονται από τα μοντέλα. Η γνώση που αποκάλυψαν οι ερευνητές ήταν ότι ένας συγκεκριμένος συνδυασμός προδιαγραφών κατασκευής διπλασιάζει την πιθανότητα να αντιμετωπιστεί ένα ηλεκτρολογικό πρόβλημα στο αυτοκίνητο. Οι μηχανικοί της DaimlerChrysler έχουν αρχίσει να ερευνούν πως αυτός ο συνδυασμός παραγόντων μπορεί να οδηγήσει σε αύξηση των ηλεκτρολογικών προβλημάτων. Επίσης ερεύνησαν αν ορισμένα

γκαράζ είχαν περισσότερες διεκδικήσεις εγγυήσεων ενός ορισμένου τύπου από άλλα γκαράζ. Τα αποτελέσματα των κανόνων συσχέτισης έδειξαν, πράγματι, τα επίπεδα εμπιστοσύνης του κανόνα « If συνεργείο X, then ηλεκτρολογικό πρόβλημα» διαφοροποιήθηκε σημαντικά από γκαράζ σε γκαράζ.

5.Φάση αξιολόγησης

Αρχικά, οι ερευνητές δήλωσαν απογοητευμένοι διότι στην πραγματικότητα δεν βρέθηκε κάποιος κανόνας που οι εμπειρογνώμονες θα έκριναν ότι ήταν ενδιαφέρον, τουλάχιστον με την πρώτη ματιά . Συμφώνα με αυτό το κριτήριο, λοιπόν, βρέθηκε ότι τα μοντέλα δεν είχαν αποτελεσματικότητα και δεν ανταποκρίθηκαν στους σκοπούς που τεθήκαν για αυτά στη φάση κατανόησης της επιχείρησης .Για να το αιτιολογήσουν αυτό οι ερευνητές δείχνουν τη δομή «κληρονομιά» της βάσης δεδομένων για την όποια τα μέρη του αυτοκίνητου κατηγοριοποιήθηκαν από τα γκαράζ και τα εργοστάσια για ιστορικούς η τεχνικούς λόγους και δεν σχεδιαστήκαν για εξόρυξη δεδομένων .Προτείνουν την προσαρμογή και τον επανασχεδιασμό της βάσης δεδομένων για να την κάνουν πιο σχετική με ανακάλυψη γνώσης.

6.Φάση ανάπτυξης

Οι ερευνητές έχουν αναγνωρίσει το προηγούμενο σχέδιο ως πιλοτικό σχέδιο και έτσι δεν σκοπεύουν να αναπτύξουν καθόλου μεγάλης κλίμακας μοντέλα για αυτήν την πρώτη επανάληψη. Μετά το πιλοτικό σχέδιο ,ωστόσο, έχουν εφαρμόσει τα μαθήματα που έμαθαν από αυτό το σχέδιο με σκοπό να ενσωματώσουν τις μεθόδους τους με το υπάρχον περιβάλλον τεχνολογίας πληροφοριών στην DaimlerChrysler.Για να υποστηρίξουν περεταίρω τον αρχικό στόχο να χαμηλώσουν το κόστος διεκδικήσεων σκοπεύουν να αναπτύξουν μια ενδοδικτυακή ικανότητα προσφοράς άντλησης πληροφοριών του QUIS για όλους τους εταιρικούς υπάλληλους.

3.2 Μελέτη 2η – Εφαρμογή Κατηγοριοποίησης

«Πρόβλεψη ανώμαλων αποδόσεων στο χρηματιστήριο με τη χρήση νευρωνικών δικτύων»

1.Φάση κατανόησης έρευνας /επιχείρησης.

Ο Alan M. Safer του California State University, αναφέρει ότι οι συναλλαγές που γίνονται στο χρηματιστήριο από γνώστες συνήθως έχουν ανώμαλες αποδόσεις. Άνθρωποι έκτος χρηματιστήριου έχουν αυξημένα κέρδη χρησιμοποιώντας νόμιμες πληροφορίες συναλλαγής

από γνώστες ιδιαίτερα εστιάζοντας σε ιδιότητες όπως το μέγεθος της εταιρίας και το χρονικό πλαίσιο για πρόβλεψη. Ο Safer ενδιαφέρεται να χρησιμοποιήσει μεθοδολογία εξόρυξης δεδομένων για να αυξήσει την ικανότητα να προβλέψει ανώμαλες αποδόσεις στο χρηματιστήριο που προκύπτουν από νόμιμες συναλλαγές από γνώστες.

2.Φάση κατανόησης δεδομένων.

Ο Safer συνέλλεξε δεδομένα από 343 εταιρίες από τον Ιανουάριο του 1993 μέχρι τον Ιούνιο του 1997. Οι μετοχές που χρησιμοποιήθηκαν στη μελέτη ήταν όλες οι μετοχές που είχαν αρχεία από γνώστες για ολόκληρη την περίοδο και ήταν S&P 600 S&P 400 S&P 500 (μικρή, μεσαία και μεγάλη κεφαλαιοποίηση αντίστοιχα) από τον Ιούνιο του 1997. Από 946 μετοχές που πρόεκυψαν που ταίριαζαν σε αυτήν την περιγραφή ο Safer επέλεξε μόνο εκείνες τις μετοχές που υπέστησαν τουλάχιστον 2 εντολές αγοράς ανά χρόνο για να διασφαλίσει μια επαρκή ποσότητα δεδομένων συναλλαγής για τις αναλύσεις εξόρυξης δεδομένων. Αυτό οδήγησε στο να χρησιμοποιηθούν 343 μετοχές για τη μελέτη μεταβλητών στο αρχικό σετ δεδομένων περιλαμβάνοντας την εταιρία, όνομα και βαθμό του γνώστη ,ημερομηνία συναλλαγής ,τιμή μετοχής, αριθμός μετοχών που συναλλαχτήκαν, τύπος συναλλαγής (αγορά ή πώληση) και αριθμός των μετοχών που κρατηθήκαν μετά την συναλλαγή. Για να αξιολογηθεί το προηγούμενο μοτίβο συναλλαγής του γνώστη η μελέτη εξέτασε τις προηγούμενες 9 και 18 εβδομάδες ,για την πρόβλεψη ανώμαλων αποδόσεων καθιερώθηκαν ως 3,6,9 και 12 μήνες.

3.Φάση προετοιμασίας δεδομένων.

Ο Safer αποφάσισε ο βαθμός του γνώστη στην εταιρία δε θα χρησιμοποιούνταν ως ιδιότητα μελέτης αφού άλλες έρευνες είχαν δείξει ότι ήταν αμφιλεγόμενης άξιας πρόβλεψης για την πρόβλεψη ανώμαλων αποδόσεων των τιμών των μετοχών. Ομοίως παρέλειψε γνώστες οι οποίοι δεν συμμετείχαν σε αποφάσεις της εταιρείας.

4.Φάση μοντελοποίησης.

Τα δεδομένα χωριστήκαν σε ένα σετ εκπαίδευσης (80% των δεδομένων) και ένα σετ επικύρωσης (20%). Ένα μοντέλο νευρωνικού δικτύων εφαρμόστηκε το οποίο αποκάλυψε τα παρακάτω αποτελέσματα:

Α) ορισμένες βιομηχανίες είχαν τις πιο προβλέψιμες ανώμαλες αποδόσεις μετοχών συμπεριλαμβανόμενου:

- βιομηχανικό γκρουπ 36: ηλεκτρονικός εξοπλισμός εκτός από εξοπλισμό υπολογιστών
- βιομηχανικό γκρουπ 28: χημικά προϊόντα
- βιομηχανικό γκρουπ 37: εξοπλισμός μεταφοράς
- βιομηχανικό γκρουπ 73: υπηρεσίες επιχειρήσεων

Β) προβλέψεις που κοίταξαν πιο μακριά στο μέλλον (9 με 12 μήνες) είχαν αυξημένη ικανότητα να αναγνωρίσουν ασυνήθιστες μεταβολές στις συναλλαγές γνώστη από ότι έκαναν οι προβλέψεις που είχαν κοντύτερο χρονικό πλαίσιο (3 με 6 μήνες).

Γ) ήταν ευκολότερο να προβληθούν ανώμαλες αποδόσεις μετοχών από συναλλαγές γνώστη από μικρές εταιρίες πάρα από μεγάλες εταιρίες.

5. Φάση αξιολόγησης.

Ο Safer ταυτόχρονα εφάρμοσε ένα πολυπαραγοντικο μοντέλο οπισθοδρόμησης (MARS) στο ίδιο σετ δεδομένων. Το μοντέλο MARS αποκάλυψε πολλά ίδια ευρήματα με το μοντέλο νευρωνικού δικτυού συμπεριλαμβάνοντας τα αποτελέσματα Α και Β από τη φάση μοντελοποίησης. Μια τέτοια σύμπτωση αποτελεσμάτων είναι μια ισχυρή και κομψή μέθοδος για την αξιολόγηση της ποιότητας και της αποτελεσματικότητας του μοντέλου ανάλογη με το να παίρνεις δύο ανεξαρτήτους κριτές και να συμφωνούν σε μια απόφαση. Αυτοί που αντλούν δεδομένα θα πρέπει να προσπαθήσουν να παράγουν μια τέτοια σύμπτωση αποτελεσμάτων όποτε προκύπτει μια τέτοια ευκαιρία. Αυτό είναι δυνατόν επειδή συχνά περισσότερες από μια μέθοδοι άντλησης δεδομένων μπορεί να εφαρμοστούν κατάλληλα στο έκαστο πρόβλημα και τα δυο μοντέλα συμφωνούν στα αποτελέσματα αυτό ενδυναμώνει την σιγουριά στα ευρήματα. Αν τα μοντέλα διαφωνούν θα πρέπει ,πιθανόν, να διερευνήσουμε επιπλέον. Μερικές φορές ένας τύπος μοντέλου είναι απλά καλύτερος για να αποκαλύψει ένα ορισμένο είδος αποτελέσματος αλλά μερικές φορές η διαφωνία υποδηλώνει μεγαλύτερα προβλήματα κάτι το οποίο απαιτεί επιστροφή σε προηγούμενες φάσεις.

6.Φάση ανάπτυξης.

Η δημοσίευση των ευρημάτων του Safer στο Intelligent Data Analysis αποτελεί μια μέθοδο ανάπτυξης μοντέλου .Τώρα οι αναλυτές από όλον τον κόσμο μπορούν να επωφεληθούν από αυτές τις μεθόδους για να εντοπίσουν τις ανώμαλες αποδόσεις τιμών μετοχών από συναλλαγές γνώστη και έτσι να προστατευτεί ο μικρός επενδυτής.

3.3 Μελέτη 3η – Εφαρμογή Κανόνων Συσχέτισης

«Άντληση κανόνων συσχέτισης από νόμιμες βάσεις δεδομένων»

1.Φάση κατανόησης έρευνας/επιχείρησης.

Οι ερευνητές, Sasha Ivkovic και John Yearwood από το πανεπιστήμιο του Ballarat, και ο Andrew Stranieri από το πανεπιστήμιο La Trobe της Αυστραλίας ενδιαφέρονται για το αν ενδιαφέροντες και ένακτοι κανόνες συσχετισμού μπορούν να αποκαλυφθούν σε ένα μεγάλο σετ δεδομένων το οποίο περιέχει πληροφορίες για νόμιμη βοήθεια χρηματοδοτούμενη από την κυβέρνηση στην Αυστραλία. Επειδή τα περισσότερα νόμιμα δεδομένα δεν είναι δομημένα με τέτοιο τρόπο ώστε να ταιριάζουν εύκολα στις περισσότερες τεχνικές εξόρυξης δεδομένων η εφαρμογή μεθόδων ανακάλυψης γνώσης στα νόμιμα δεδομένα δεν έχει αναπτυχτεί τόσο γρήγορα όσο σε άλλους τομείς. Σκοπός των ερευνητών είναι να βελτιώσουν την παράδοση νόμιμων υπηρεσιών και δίκαιων νομικών αποτελεσμάτων μέσα από τη βελτιωμένη χρήση των διαθέσιμων νόμιμων δεδομένων.

2. Φάση κατανόησης δεδομένων.

Τα δεδομένα παρέχονται από το Victoria Legal Aid (VLA), έναν διακυβερνητικό οργανισμό ο όποιος στοχεύει στο να παρέχει πιο αποτελεσματική νόμιμη βοήθεια σε μη προνομιούχους ανθρώπους στην Αυστραλία. Πάνω από 380.000 αιτήσεις για νομική βοήθεια μαζεύτηκαν από τα 11 περιφερειακά γραφεία του VLA από το 1997 μέχρι το 1999 συμπεριλαμβάνοντας πληροφορείς για πάνω από 300 μεταβλητές. Σε μια προσπάθεια να μειώσουν τον αριθμό των μεταβλητών οι ερευνητές στραφήκαν σε ειδικούς του χώρου για βοήθεια. Αυτοί οι ειδικοί μάζεψαν 7 από τις πιο σημαντικές μεταβλητές για συμπερίληψη στο σετ δεδομένων όπως το φύλο, η ηλικία, το επάγγελμα και ο λόγος για την άρνηση βοήθειας ,νομικός τύπος (για παράδειγμα αστικό δικαίου), απόφαση (για παράδειγμα βοήθεια που δίνεται η δε δίνεται) και τύπος αντιμετώπισης (για παράδειγμα εμφάνιση στο δικαστήριο).

3. Φάση προετοιμασίας δεδομένων.

Τα δεδομένα του VLA αποδείχτηκαν σχετικά καθαρά περιέχοντας πολύ λίγα αρχεία με άξιες ιδιοτήτων που είτε έλειπαν είτε ήταν λάθος. Αυτό οφείλεται εν μέρη στο σύστημα διαχείρισης της βάσης δεδομένων που χρησιμοποιείται από το VLA το οποίο διενεργεί ελέγχους

ποιότητας στα εισερχόμενα δεδομένα. Η μεταβλητή της ηλικίας διαχωρίστηκε σε ευδιάκριτα διαστήματα όπως κάτω των 18, πάνω των 50 κτλ.

4.Φάση μοντελοποίησης.

Οι κανόνες περιορίστηκαν στο να έχουν ένα μόνο προηγούμενο και ένα μόνο αποτέλεσμα. Πολλοί ενδιαφέροντες κανόνες συσχέτισης απεκαλύφθησαν μαζί με μη ενδιαφέροντες κανόνες το οποίο είναι το χαρακτηριστικό σενάριο της άντλησης των κανόνων συσχέτισης. Ένας τέτοιος ενδιαφέρον κανόνας ήταν : *If τόπος γέννησης= Βιετνάμ Then ο νομικός τύπος= ποινικό δίκαιο*, με ακρίβεια 90% .Οι ερευνητές προχώρησαν σε μια ακριβή υπόθεση ότι οι κανόνες συσχέτισης είναι ενδιαφέροντες εάν παράγουν ενδιαφέρουσες υποθέσεις .Μια συζήτηση ανάμεσα στους ερευνητές και τους ειδικούς για τους λόγους που βρίσκονται κάτω από των παραπάνω κανόνα συσχέτισης λήφθηκαν υπόψη οι παρακάτω υποθέσεις:

Υπόθεση Α: Οι βιετναμέζοι αιτώντας έκαναν αίτηση για υποστήριξη μόνο για το ποινικό δίκαιο και όχι για άλλους τύπους όπως το οικογενειακό και το αστικό δίκαιο.

Υπόθεση Β. Οι βιετναμέζοι αιτώντας έκαναν περισσότερα εγκλήματα από άλλες ομάδες.

Υπόθεση Γ. Υπάρχει μια δόλια μεταβλητή. Ίσως οι άντρες βιετναμέζοι είναι πιο πιθανό να ζητήσουν βοήθεια από ότι οι γυναίκες, και οι άντρες σχετίζονται περισσότερο με το ποινικό δίκαιο.

Υπόθεση Δ. Οι βιετναμέζοι δεν είχαν έτοιμη πρόσβαση στο διαφημιστικό υλικό του VLA .

Η ομάδα των ερευνητών και ειδικών κατέληξαν ανεπίσημα στο ότι η υπόθεση Α ήταν πιο πιθανή αν και επιπλέον έρευνα ίσως δικαιολογείται και κανένας αιτιακός συσχετισμός δεν μπορεί να υποτεθεί . Σημειώστε ωστόσο την έντονη ανθρωπινή διαπεραστικότητα καθόλη τη διάρκεια της διαδικασίας εξόρυξης δεδομένων. Χάρις την γνώση των ειδικών του χώρου και την εμπειρία τους τα αποτελέσματα εξόρυξης δεδομένων σε αυτή την περίπτωση δεν θα ήταν καρποφόρα.

5.Φάση αξιολόγησης.

Οι ερευνητές υιοθέτησαν μια μοναδική μέθοδο αξιολόγησης για αυτό το σχέδιο. Έφεραν 3 ειδικούς του χώρου και απέσπασαν από αυτούς τους υπολογισμούς των επίπεδων ακριβείας για κάθε έναν από τους 114 κανόνες συσχέτισης. Αυτά τα υπολογιζόμενα επίπεδα ακριβείας

συγκριθήκαν μετά με τα ίδια επίπεδα ακρίβειας των κανόνων συσχέτισης που αποκαλύφθηκαν στο σετ δεδομένων.

6.Φάση ανάπτυξης.

Μια χρήσιμη εφαρμογή με βάση το ιντερνέτ, WebAssociator ,αναπτύχτηκε έτσι ώστε και ένας μη ειδικός να μπορεί να εκμεταλλευτεί τη μηχανή κατασκευής κανόνων. Η χρήστες επιλέγουν το μοναδικό προηγούμενο και το μοναδικό αποτέλεσμα χρησιμοποιώντας έναν τύπο με βάση το ιντερνέτ. Οι ερευνητές προτείνουν ότι αυτή η εφαρμογή μπορεί να χρησιμοποιηθεί ως ένα σύστημα νομικής υποστήριξης ειδικά για την αναγνώριση άδικων διαδικασιών.

3.4 Μελέτη 4η – Εφαρμογή Κατηγοριοποίησης

«Πρόβλεψη εταιρικών χρεωκοπιών χρησιμοποιώντας δέντρα αποφάσεων»

1.Φάση κατανόησης έρευνας/επιχείρησης.

Η πρόσφατη οικονομική κρίση στην ανατολική Ασία έχει δημιουργήσει ένα πρωτοφανές επίπεδο εταιρικών πτωχεύσεων σε εκείνη την περιοχή και σε όλο τον κόσμο. Ο στόχος των ερευνητών, Tae Kyung Sung από το πανεπιστήμιο Kyonggi, Namsik Chang από το πανεπιστήμιο της Σεούλ, και Gunhee Lee του πανεπιστημίου Sogang της Κορέα, είναι να αναπτύξουν τα πρότυπα για την πρόβλεψη των εταιρικών πτωχεύσεων που μεγιστοποιούν την επεξηγηματικότητα των αποτελεσμάτων. Θεώρησαν ότι η επεξηγηματικότητα ήταν σημαντική επειδή μια αρνητική πρόβλεψη πτώχευσης μπορούσε από μόνη της να ασκήσει καταστρεπτική επίδραση σε έναν οικονομικό οργανισμό, έτσι ώστε οι εταιρίες για τις οποίες προβλέφθηκε να χρεοκοπήσουν, απαιτείται ισχυρή και λογική σκέψη. Εάν η επιχείρησή κάποιου είναι σε κίνδυνο, και μια πρόβλεψη της πτώχευσης θα μπορούσε από μόνη της να συμβάλει στην τελική αποτυχία, αυτή η πρόβλεψη καλύτερα να υποστηρίζεται από «ανιχνεύσιμα» στοιχεία και όχι μια απλή απόφαση. Επομένως, οι ερευνητές επέλεξαν δέντρα απόφασης ως μέθοδο ανάλυσής τους, λόγω της διαφάνειας του αλγορίθμου και την επεξηγηματικότητα των αποτελεσμάτων.

2. Φάση κατανόησης δεδομένων.

Τα στοιχεία περιλαμβάνονται σε δύο ομάδες, κορεατικές εταιρίες που χρεοκόπησαν στη σχετικά σταθερή αναπτυξιακή περίοδο από 1991-1995, και κορεατικές εταιρίες που χρεοκόπησαν σε συνθήκες οικονομικής κρίσης κατά την περίοδο 1997-1998. Μετά από τις διάφορες διαδικασίες διαλογής, 29 εταιρίες εντοπίστηκαν, κυρίως στον κατασκευαστικό τομέα. Το οικονομικά δεδομένα συλλέχθηκαν απευθείας από το χρηματιστήριο αξιών της Κορέας και ελέγχθηκαν από την τράπεζα της Κορέας και την Korea Industrial Bank.

3. Φάση προετοιμασιών στοιχείων.

Πενήντα έξι χρηματοοικονομικοί δείκτες προσδιορίστηκαν από τους ερευνητές μέσω μιας αναζήτησης της βιβλιογραφίας στην πρόβλεψη πτώχευσης, 16 εκ των οποίων στη συνέχεια υποχώρησαν λόγω της επανάληψης. Έτσι παρέμειναν 40 χρηματοοικονομικοί δείκτες στο σύνολο στοιχείων, συμπεριλαμβανομένων των μέτρων για την ανάπτυξη, την αποδοτικότητα, την ασφάλεια/δύναμη, δραστηριότητα/αποδοτικότητα, και παραγωγικότητα.

4.Φάση μοντελοποίησης.

Ξεχωριστά μοντέλα δέντρων αποφάσεων εφαρμόστηκαν στα δεδομένα φυσιολογικών συνθηκών αλλά και σε αυτά των συνθηκών κρίσης . Τα μοντέλα δέντρων αποφάσεων μπορούν εύκολα να δημιουργήσουν σετ κανόνων.

Μερικοί από τους κανόνες αποκάλυψαν για τις φυσιολογικές συνθήκες τα παρακάτω :

- Εάν η παραγωγικότητα του κεφαλαίου είναι μεγαλύτερη από 19,65 τότε υπάρχει πρόβλεψη χρεωκοπίας με 86% ακρίβεια
- Εάν ο δείκτης ροής μετρητού στα συνολικά κεφάλαια είναι μεγαλύτερος από -5,65 η πρόβλεψη χρεωκοπίας έχει ακρίβεια 95%
- Αν η παραγωγικότητα τοθ κεφαλαίου είναι στο 19,65 η πιο κάτω και ο δείκτης ροής μετρητών στα συνολικά κεφάλαια είναι στο -5,65 η πιο κάτω η πρόβλεψη χρεωκοπίας έχει 84% ακρίβεια

Μερικοί κανόνες για τις συνθήκες κρίσης ήταν οι παρακάτω:

- ✓ εάν η παραγωγικότητα του κεφαλαίου είναι μεγαλύτερη από 20,61 τότε η ακρίβεια είναι 91%
- ✓ Αν ο δείκτης μετρητών στο παθητικό είναι μεγαλύτερος από 2,64 η ακρίβεια είναι στο 85%
- ✓ Εάν ο δείκτης καθορισμένων κεφαλαίων της ισότητας των κάτοχων μετοχών και του μακροπρόθεσμου παθητικού είναι μεγαλύτερος από 87,23 η ακρίβεια είναι στο 86%
- ✓ Εάν η παραγωγικότητα του κεφαλαίου είναι στο 20,60 και ο δείκτης της ροής μετρητών στο παθητικό είναι στο 2,64 η και κάτω κ ο δείκτης καθορισμένων κεφαλαίων της ισότητας κάτοχων μετοχών και το μακροπρόθεσμο παθητικό είναι στο 87,23 η και πιο κάτω τότε η ακρίβεια είναι στο 84%

Η ροη μετρητών και η παραγωγικότητα κεφαλαίου είναι σημαντικά άσχετα από τις οικονομικές συνθήκες ,ενώ, η ροη μετρητού είναι γνώστη στη φιλολογία ως πρόβλεψη χρεωκοπίας, η αναγνώριση της παραγωγικότητας κεφαλαίου ήταν σχετικά σπανία κάτι το οποίο απαιτούσε επιπλέον επαλήθευση.

5.Φάση αξιολόγησης.

Οι ερευνητές ανέθεσαν σε μια ομάδα ειδικών στα οικονομικά να επιλέξουν ομοφώνα την παραγωγικότητα κεφαλαίου ως τη πιο σημαντική ιδιότητα για την διαφοροποίηση των εταιριών που κινδυνεύουν να χρεοκοπήσουν από τις άλλες εταιρίες.

Άρα, τα αναπάντεχα αποτελέσματα που βγήκαν από το μοντέλο δέντρων αποφάσεων επαληθευτήκαν από τους ειδικούς. Για να διασφαλίσουν ότι το μοντέλο μπορούσε να γενικευθεί για όλες τις κορεάτικες κατασκευαστικές εταιρίες επιλέχτηκε ένα δείγμα ελέγχου και οι ιδιότητες του συγκριθήκαν με εκείνες στο σετ δεδομένων. Βρέθηκε ότι ο μέσος όρος κεφαλαίων του δείγματος έλεγχου και ο αριθμός των υπάλληλων ήταν μέσα στο 20% του δείγματος δεδομένων.

Τέλος, οι ερευνητές εφήρμοσαν πολλαπλή διακριτικά ανάλυση ως τη βάση απόδοσης. Πολλοί από τους 40 οικονομικούς δείκτες βρεθήκαν ότι ήταν σημαντικοί διαγνώστες χρεωκοπίας και η τελική διακριτική λειτουργιά περιλάμβανε μεταβλητές που αναγνωριστήκαν από το μοντέλο δέντρων αποφάσεων.

6.Φάση ανάπτυξης.

Δεν αναγνωρίστηκε καμία ανάπτυξη. Όπως αναφέρθηκε νωρίτερα η ανάπτυξη είναι στη διακριτική ευχέρεια των χρηστών. Ωστόσο, εξαιτίας της έρευνας αυτής οι οικονομικοί οργανισμοί στην Κορέα γνωρίζουν καλυτέρα τους διαγνωστες χρεωκοπίας σε συνθήκες κρίσης σε αντιδιαστολή με τις φυσιολογικές συνθήκες.

3.5 Μελέτη 5η – Εφαρμογή Συσταδοποίησης

«Δημιουργία προφίλ της τουριστικής αγοράς χρησιμοποιώντας K-Means ανάλυση ομάδων»

1.Φάση κατανόησης έρευνας/επιχείρησης.

Οι ερευνητές, Simon Hudson και Brent Ritchie, του πανεπιστημίου του Calgary, στην Αλμπέρτα, του Καναδά, ενδιαφέρονται να μελετήσουν τη συμπεριφορά των τουριστών στην περιοχή αυτή. Σκοπός τους είναι η δημιουργία σχεδιαγραμμάτων των εγχώριων τουριστών βασισμένα στη συμπεριφορά απόφασης τους. Ο γενικός στόχος της μελέτης είναι να σχηματίσουν μια ποσοτική βάση για την ανάπτυξη μιας εκστρατείας μάρκετινγκ που χρηματοδοτείται από το Travel Alberta. Προς αυτόν τον στόχο, οι κύριοι στόχοι ήταν να καθοριστούν ποιοι παράγοντες ήταν σημαντικοί για την επιλογή των προορισμών στην Αλμπέρτα, να αξιολογηθούν οι εσωτερικές αντιλήψεις για το «προϊόν διακοπών Αλμπέρτα,» και για να γίνει κατανοητή η διαδικασία λήψης απόφασης του ταξιδιού.

2. Φάση κατανόησης δεδομένων.

Τα στοιχεία συλλέχθηκαν στα τέλη του 1999 χρησιμοποιώντας μια τηλεφωνική έρευνα για 13.445 κατοίκους της Αλμπέρτα. Οι ερωτηθέντες ελέγχθηκαν σύμφωνα με εκείνους που ήταν άνω των 18 ετών και είχαν ταξιδέψει για λόγους αναψυχής τουλάχιστον 80 χιλιόμετρα για

τουλάχιστον μια νύχτα στην Αλμπέρτα κατά το προηγούμενο έτος. Μόνο 3.071 από τους 13.445 κατοίκους συμπλήρωσαν την έρευνα και ήταν επιλέξιμοι για το συνυπολογισμό στη μελέτη.

3. Φάση προετοιμασιών στοιχείων.

Ένα από τα ερωτήματα της έρευνας ήταν να δηλώσουν σε ποιο βαθμό καθένας από μια λίστα 13 παραγόντων επηρεάζουν περισσότερο τις αποφάσεις τους στα ταξίδια. Αυτά στη συνέχεια θεωρούνται ως μεταβλητές βάσει των οποίων διεξήχθη η ανάλυση συστάδων , και περιλαμβάνονται παράγοντες όπως η ποιότητα των καταλυμάτων, σχολικές διακοπές, και καιρικές συνθήκες.

4. Φάση διαμόρφωσης.

Η συσταδοποίηση είναι μια φυσική μέθοδος για τα σχεδιαγράμματα τμήματος. Οι ερευνητές επέλεξαν την K-means ομαδοποίηση, δεδομένου ότι εκείνος ο αλγόριθμος είναι γρήγορος και αποδοτικός εφόσον γνωρίζουμε τον αριθμό συστάδων που αναμένετε να βρεθούν. Διερεύνησαν μεταξύ δύο και έξι προτύπων συστάδων πριν καταλήξουν σε μια λύση πέντε-συστάδων ως η καλύτερη λύση που αντικατοπτρίζει την πραγματικότητα. Τα συνοπτικά σχεδιαγράμματα των συστάδων είναι τα ακόλουθα:

- Συστάδα 1: η νέα αστική υπαίθρια αγορά. Η νεώτερη όλων των συστάδων, που ισορροπούνται εξίσου με τα σχολικά προγράμματα και τους προϋπολογισμούς που κυριαρχούν στις αποφάσεις ταξιδιού τους.

- Συστάδα 2: η εσωτερική αγορά του ταξιδιώτη αναψυχής. Επόμενη νεώτερη συστάδα , συνήθως παντρεμένοι με παιδιά, με επισκέψεις στην οικογένεια και σε φίλους, ένας κύριος παράγοντας στα σχέδια ταξιδιού.

- Συστάδα 3: τα παιδιά-πρώτη αγορά. Περισσότεροι παντρεμένοι και περισσότερα παιδιά από οποιαδήποτε άλλη συστάδα, με τα σπορ των παιδιών και τα προγράμματα ανταγωνισμού να έχουν το μεγάλο βάρος στην απόφαση σε πιο μέρος της Αλμπέρτα να ταξιδεύουν.

- Συστάδα 4: δίκαιη-καιρικές συνθήκες. Δεύτερη παλαιότερη συστάδα, όπου συμμετέχουν περισσότεροι άνδρες ,με τις καιρικές συνθήκες να επηρεάζουν τις αποφάσεις ταξιδιού.

- Συστάδα 5: η παλαιότερη, κόστος-συνειδητή αγορά ταξιδιώτη. Η παλαιότερη των συστάδων, οι περισσότεροι επηρεάζονται από το κόστος/αξίας και το ασφαλές περιβάλλον κατά την απόφαση πραγματοποίησης ταξιδιού στην Αλμπέρτα.

5. Φάση αξιολόγησης.

Διαχωριστική ανάλυση χρησιμοποιήθηκε για να ελέγξει την «πραγματικότητα» των συστάδων κατηγοριοποίησης , σωστά ταξινομημένα περίπου το 93% των ατόμων στις σωστές συστάδες. Η διακριτική ανάλυση έδειξε, επίσης, ότι οι διαφορές μεταξύ των συστάδων ήταν στατιστικά σημαντικές.

6. Φάση ανάπτυξης.

Αυτά τα συμπεράσματα μελέτης οδήγησαν στην προώθηση μιας νέας εκστρατείας μάρκετινγκ, «Alberta, Made to Order» με βάση την προσαρμογή του μάρκετινγκ στους τύπους συστάδων που αποκαλύπτονται στην εξόρυξη δεδομένων. Περισσότερα από 80 προγράμματα προωθήθηκαν, μέσω μιας συνεταιριστικής ρύθμισης μεταξύ της κυβέρνησης και των επιχειρήσεων. Για την εκστρατεία αυτή έχουν ήδη προβληθεί τηλεοπτικές διαφημίσεις, περίπου 20 φορές πάνω από το 90% των ενηλίκων ηλικίας κάτω των 55. ToTravel Alberta αργότερα υπέστη αύξηση πάνω από 20% στον αριθμό των κατοίκων της Αλμπέρτα που έδειξαν την Αλμπέρτα ως ιδανικό προορισμό.

4ο Κεφάλαιο

Το πακέτο WEKA

4.1 Εισαγωγή

Είναι γεγονός ότι δεν υπάρχει κανένας αλγόριθμος εξόρυξης δεδομένων που να εφαρμόζεται σε όλες τις περιπτώσεις. Τα σύνολα δεδομένων που συναντώνται στην πράξη ποικίλουν ευρέως και για να εξάγει κανείς ακριβή μοντέλα απαιτείται τα χαρακτηριστικά του αλγορίθμου εκμάθησης να ταιριάζουν με τη δομή του πεδίου εφαρμογής. Άλλωστε η εξόρυξη γνώσης από δεδομένα αποτελεί μια πειραματική επιστήμη και στηρίζεται στην εφαρμογή της μεθόδου *trial and error*.

Το πακέτο Weka δημιουργήθηκε στο πανεπιστήμιο Waikato στη Νέα Ζηλανδία. Η πλήρης ονομασία του είναι Waikato Environment for Knowledge Analysis και αποτελεί μια συλλογή από τους πλέον σύγχρονους αλγορίθμους μηχανικής μάθησης. Μεταξύ άλλων περιέχει μεθόδους για:

- ✓ Προεπεξεργασία δεδομένων
- ✓ Οπτικοποίηση
- ✓ Ταξινόμηση
- ✓ Ομαδοποίηση
- ✓ Εύρεση κανόνων συσχέτισης
- ✓ Παλινδρόμηση

Το λογισμικό είναι γραμμένο σε java και είναι «ανοιχτής πηγής» (open source) και ελεύθερης διανομής.

4.2 Περιγραφή του περιβάλλοντος Weka

Κατά την έναρξη του λογισμικού, ο χρήστης καλείται να επιλέξει ένα από τα τέσσερα πιθανά περιβάλλοντα εργασίας (τα οποία θα αναλυθούν παρακάτω) που του παρέχει με τις εξής ονομασίες:

- Explorer
- Knowledge Flow
- Experimenter
- Simple CLI (Command Line Interface)

Εικόνα 18: Περιβάλλοντα εργασίας Weka

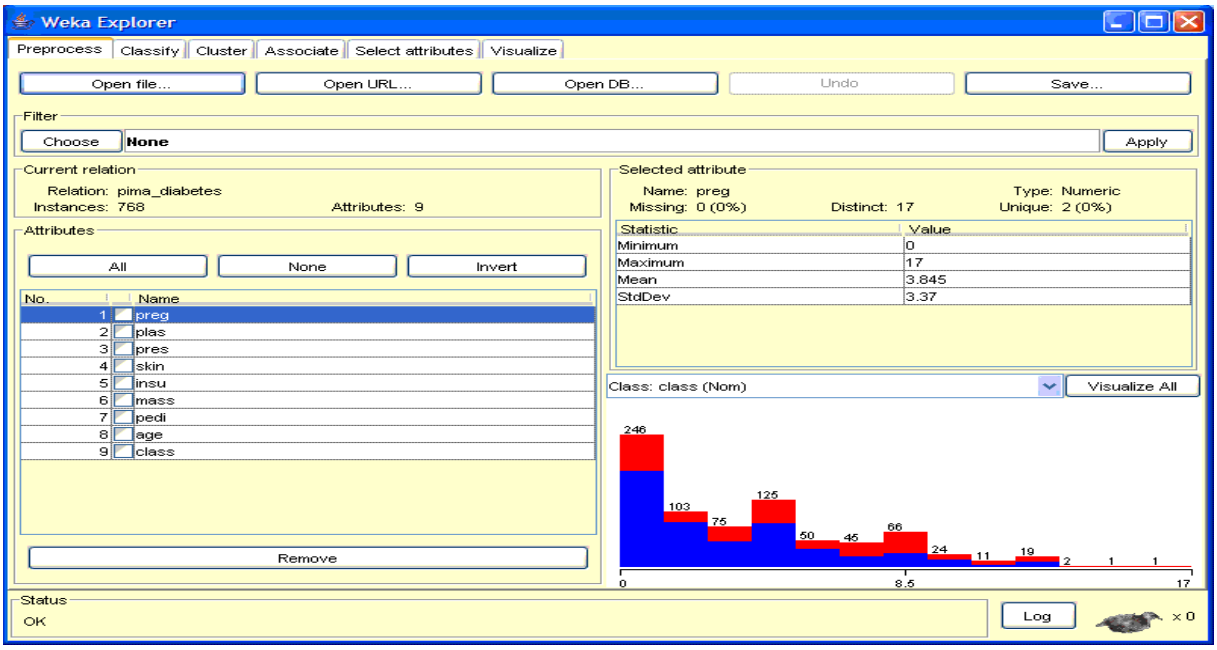
4.2.1 Explorer

Ο πιο εύχρηστος τρόπος για να χρησιμοποιήσει κανείς το Weka είναι μέσω αυτού του γραφικού περιβάλλοντος. Παρέχει πρόσβαση σε όλες τις δυνατότητες που έχει το Weka, παρουσιάζοντας του τις επιλογές του μέσα από κατάλληλα οργανωμένες λίστες. Επιλογές που δεν είναι συμβατές με την κάθε διαδικασία που ακολουθεί ο χρήστης, παρουσιάζονται μαρκαρισμένες. Επίσης χρησιμοποιούνται λογικές προεπιλεγμένες τιμές σε διάφορες επιλογές έτσι ώστε ο χρήστης να είναι σε θέση να έχει κάποια αποτελέσματα με την ελάχιστη δυνατή προσπάθεια. Μέσα από τον Explorer μπορεί κανείς να εξετάσει ότι αποτελέσματα

έχουν προκύψει από την εφαρμογή των διάφορων αλγορίθμων, να αξιολογήσει και να συγκρίνει διαφορετικά μοντέλα που έχει δημιουργήσει από διάφορα σύνολα δεδομένων, και να αστικοποιήσει τόσο τα μοντέλα όσο και τα σύνολα δεδομένων αυτών.

Ο Explorer είναι οργανωμένος σε έξι μεγάλες κατηγορίες λειτουργιών με τις αντίστοιχες ονομασίες:

- Preprocess. Περιέχει εργαλεία και αλγόριθμους που αφορούν την επιλογή ή την τροποποίηση του συνόλου δεδομένων που επεξεργάζονται.
- Classify. Περιέχει αλγόριθμους κατάλληλους για προβλήματα ταξινόμησης ή παλινδρόμησης.
- Cluster. Περιέχει αλγόριθμους που χρησιμοποιούνται για την εύρεση υποομάδων μέσα από το σύνολο δεδομένων.
- Associate. Περιέχει αλγόριθμους κατάλληλους για την εύρεση κανόνων συσχέτισης μέσα στο σύνολο δεδομένων και την αξιολόγηση τους.
- Select Attributes. Περιέχει αλγόριθμους που χρησιμοποιούνται στην επιλογή των πιο σχετικών χαρακτηριστικών μέσα από το σύνολο δεδομένων.
- Visualize. Περιέχει εργαλεία για την οπτικοποίηση των δεδομένων ή των μοντέλων που δημιουργεί σε δισδιάστατα γραφήματα.

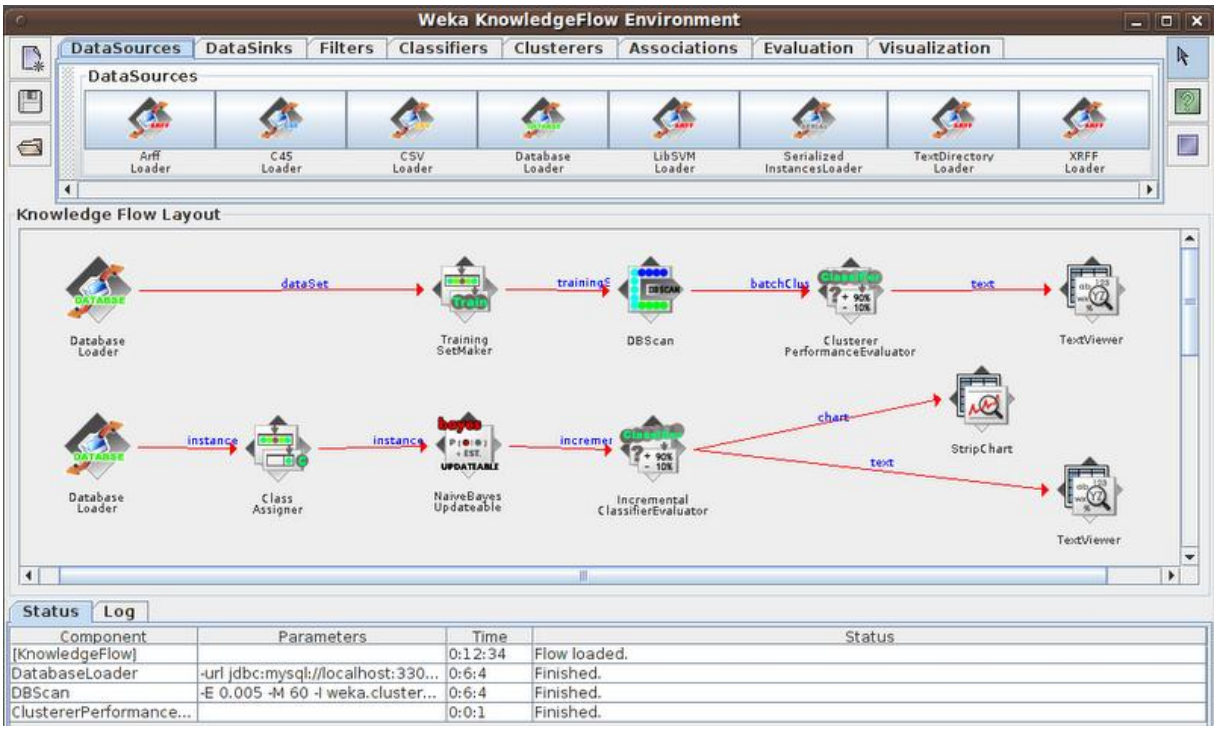


Εικόνα 19: Περιβάλλον Explorer

Φορτώνοντας ένα σύνολο δεδομένων στο πρόγραμμα, εμφανίζονται τα δεδομένα με τη μορφή γραφικών για καθένα από τα γνωρίσματα ξεχωριστά, καθώς και στατιστικές πληροφορίες για καθένα από αυτά. Επίσης, αν στο σύνολο δεδομένων δίνεται και κάποια κλάση στην οποία ταξινομούνται, τα δεδομένα που βρίσκονται στην ίδια κλάση ταξινομούνται με το ίδιο χρώμα.

4.2.2. Knowledge Flow

Το περιβάλλον αυτό απευθύνεται σε πιο προχωρημένους χρήστες , δηλαδή, σε όσους θέλουν να έχουν επίγνωση του πως τα δεδομένα και οι πληροφορίες που παράγονται από αυτά «κυλούν» μέσα στο σύστημα. Ο χρήστης επιλέγει τα διάφορα συστατικά κομμάτια του Weka , από μία μπάρα εργαλείων , τα τοποθετεί σε έναν πίνακα και τα συνδέει σε ένα κατευθυνόμενο γράφημα που υποδεικνύει πως γίνεται η ανάλυση και η επεξεργασία δεδομένων.



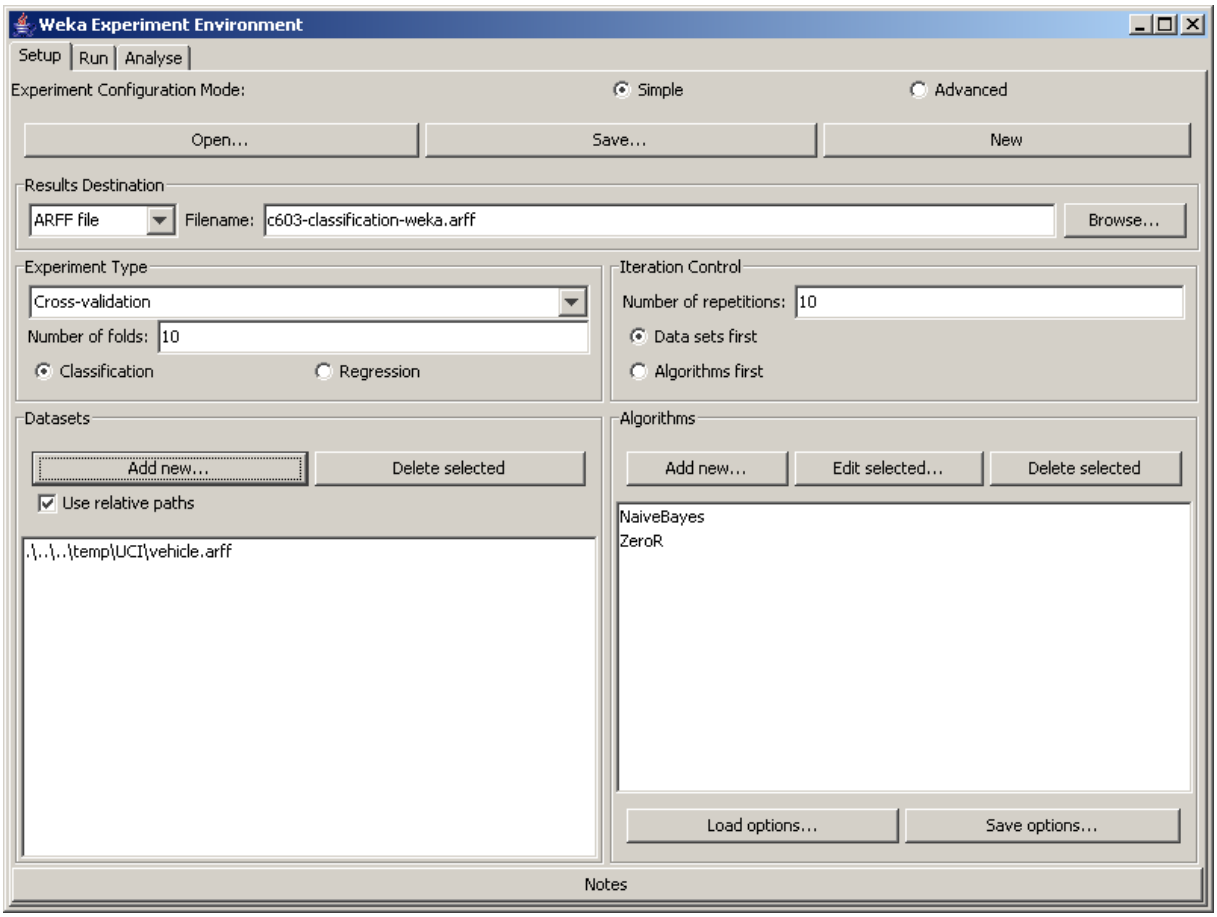
Εικόνα 20: Περιβάλλον Knowledge Flow

Λειτουργικά, το Knowledge Flow περιβάλλον μοιάζει πολύ με τον Explorer ,δηλαδή, μπορεί κανείς να εκτελέσει αντίστοιχες εργασίες και στα δύο. Η διαφορά τους είναι ότι το Knowledge Flow παρέχει μια παραπάνω ευελιξία από την άποψη ότι μπορείς να εξετάσεις όλη τη διαδικασία λεπτομερώς και όχι μόνο το αποτέλεσμα που προκύπτει. Ωστόσο, το στοιχείο που το ξεχωρίζει από τον Explorer και το κάνει να υπερέχει είναι η δυνατότητα για αυξητική λειτουργία.

Αν όλα τα στοιχεία που έχουν συνδεθεί στον πίνακα έχουν τη δυνατότητα να λειτουργήσουν αυξητικά, τότε έτσι λειτουργεί ολόκληρο το μαθησιακό σχήμα. Δεν διαβάζει ολόκληρο το σετ δεδομένων που του δίνεται σαν input πριν αρχίσει η «μάθηση», όπως θα έκανε ο Explorer, αλλά διαβάζει κάθε υπόδειγμα ξεχωριστά και το προωθεί στην διαδικασία που έχει σχηματιστεί στο Knowledge Flow πριν πάει στο επόμενο. Επομένως, μια τέτοια διάταξη μπορεί να επεξεργαστεί αρχεία οποιουδήποτε μεγέθους , ακόμα και μεγαλύτερου της κύριας μνήμης του συστήματος , καθώς δεν χρειάζεται να τα αποθηκεύσει εσωτερικά για να ξεκινήσει την διαδικασία.

4.2.3. Experimenter

Τα περιβάλλοντα Explorer και Knowledge Flow εξυπηρετούν τους χρήστες που θέλουν να διαπιστώσουν πόσο καλή απόδοση έχουν τα μαθησιακά σχήματα σε συγκεκριμένα σετ δεδομένων. Παρόλα αυτά, πιο σοβαρές ερευνητικές εργασίες απαιτούν πειράματα μεγαλύτερου εύρους, κάτι που μεταφράζεται στην εφαρμογή διαφόρων μαθησιακών σχημάτων σε πολλά διαφορετικά σετ δεδομένων και συχνά με διαφορετικές παραμέτρους , τέτοιες εργασίες περιέχει το περιβάλλον Experimenter.



Εικόνα 21: Περιβάλλον Experimenter

Ο Experimenter αυτοματοποιεί την πειραματική διαδικασία. Οι πληροφορίες που προκύπτουν για τα διάφορα μαθησιακά σχήματα και τα διάφορα σετ δεδομένων μπορούν να αποθηκευτούν καθώς και να αποτελέσουν αντικείμενο περεταίρω μελέτης και εφαρμογής εξόρυξης δεδομένων. Επιπρόσθετα, ο Experimenter έχει ένα χαρακτηριστικό υπεροχής ανάλογο αυτού που έχει το Knowledge Flow.

Η βασική τους διαφορά είναι ότι, ενώ το Knowledge Flow ξεπερνούσε τους περιορισμούς σχετικά με το μέγεθος του αρχείου, εξετάζοντας κάθε υπόδειγμα από το σετ δεδομένων χωριστά χωρίς να χρειάζεται να φορτώσει ολόκληρο το σετ δεδομένων, ο Experimenter ξεπερνά τους χρονικούς περιορισμούς. Περιέχει υποδομές για προχωρημένους χρήστες ώστε να διαμοιράσουν το υπολογιστικό φορτίο που απαιτείται από μεγάλου εύρους πειράματα σε διάφορους υπολογιστές .

4.2.4. Command Line Interface

Είναι το τελευταίο περιβάλλον εργασίας που μπορεί να συναντήσει κανείς στο Weka. Επιλέγοντας το, έρχεται στην επιφάνεια ένας κενός χώρος με μια γραμμή εισαγωγής εντολών στο κάτω μέρος . Είναι το πιο απλό και χωρίς γραφικά βοηθήματα περιβάλλον εργασίας και απευθύνεται σε χρήστες που γνωρίζουν εις βάθος το Weka και τις εντολές του.

4.3 Φόρτωση Δεδομένων στο Weka

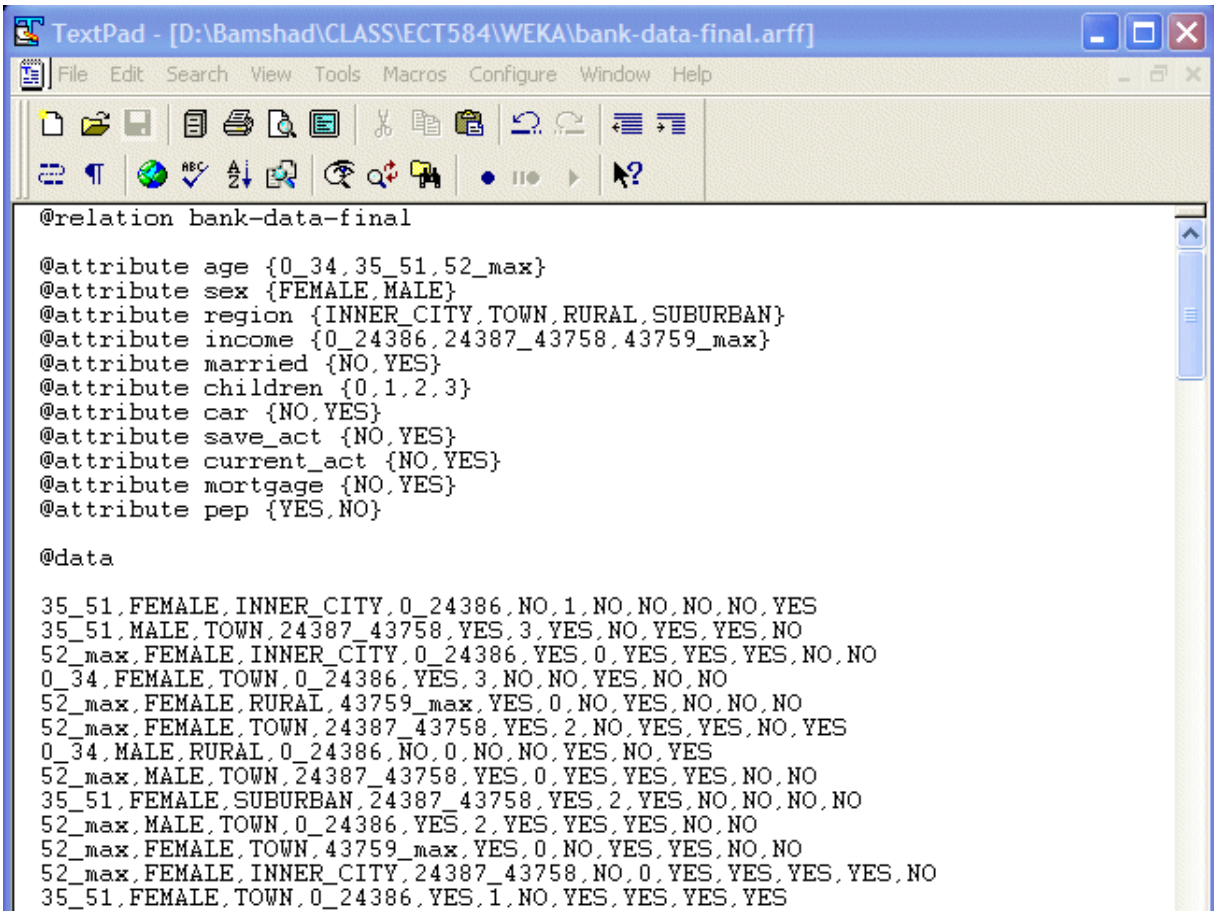
Η φόρτωση δεδομένων στο Weka μπορεί να γίνει με πολλούς τρόπους. Στον Explorer, που είναι το περιβάλλον που χρησιμοποιούμε πιο συχνά, παρατηρούμε ότι υπάρχουν στην καρτέλα preprocess οι επιλογές :

- Open file. Χρησιμοποιείται για εύρεση του αρχείου στον υπολογιστή μας, το οποίο θέλουμε να φορτώσουμε.
- Open URL. Χρησιμοποιείται για τη φόρτωση δεδομένων κατευθείαν από κάποια ιστοσελίδα του διαδικτύου.
- Open DB. Χρησιμοποιείται για να φορτώσουμε όποια δεδομένα θέλουμε από μια βάση δεδομένων.
- Generate. Χρησιμοποιείται για την δημιουργία τυχαίων δεδομένων μέσα από διάφορους αλγορίθμους, σε περίπτωση που δεν έχουμε δεδομένα διαθέσιμα και θέλουμε να πειραματιστούμε με το Weka.

Βέβαια, τα δεδομένα μπορεί να βρίσκονται σε διάφορες μορφές και τύπους αρχείων. Το Weka περιέχει ενσωματωμένους μετατροπείς για τους πιο κοινούς τύπους αρχείων για να μετρτρέψει στην τυποποίηση με την οποία μπορεί να τα χειριστεί, την τυποποίηση .arff.

4.3.1. Η Τυποποίηση .arff

Η τυποποίηση .arff αποτελεί τη φυσική μέθοδο αποθήκευσης δεδομένων του Weka. Υποστηρίζει τόσο αριθμητικά όσο και ονομαστικά χαρακτηριστικά. Τα δεδομένα αρκετές φορές βρίσκονται σε υπολογιστικά φύλλα ‘η σε βάσεις δεδομένων. Τα προγράμματα που τα χειρίζονται συνήθως επιτρέπουν την εξαγωγή των δεδομένων σε τυποποίηση .csv. Το Weka έχει ενσωματωμένο μετατροπέα αρχείων από .csv σε .arff . Παρόλα αυτά η διαδικασία μετατροπής είναι αρκετά απλή όπως φαίνεται παρακάτω.



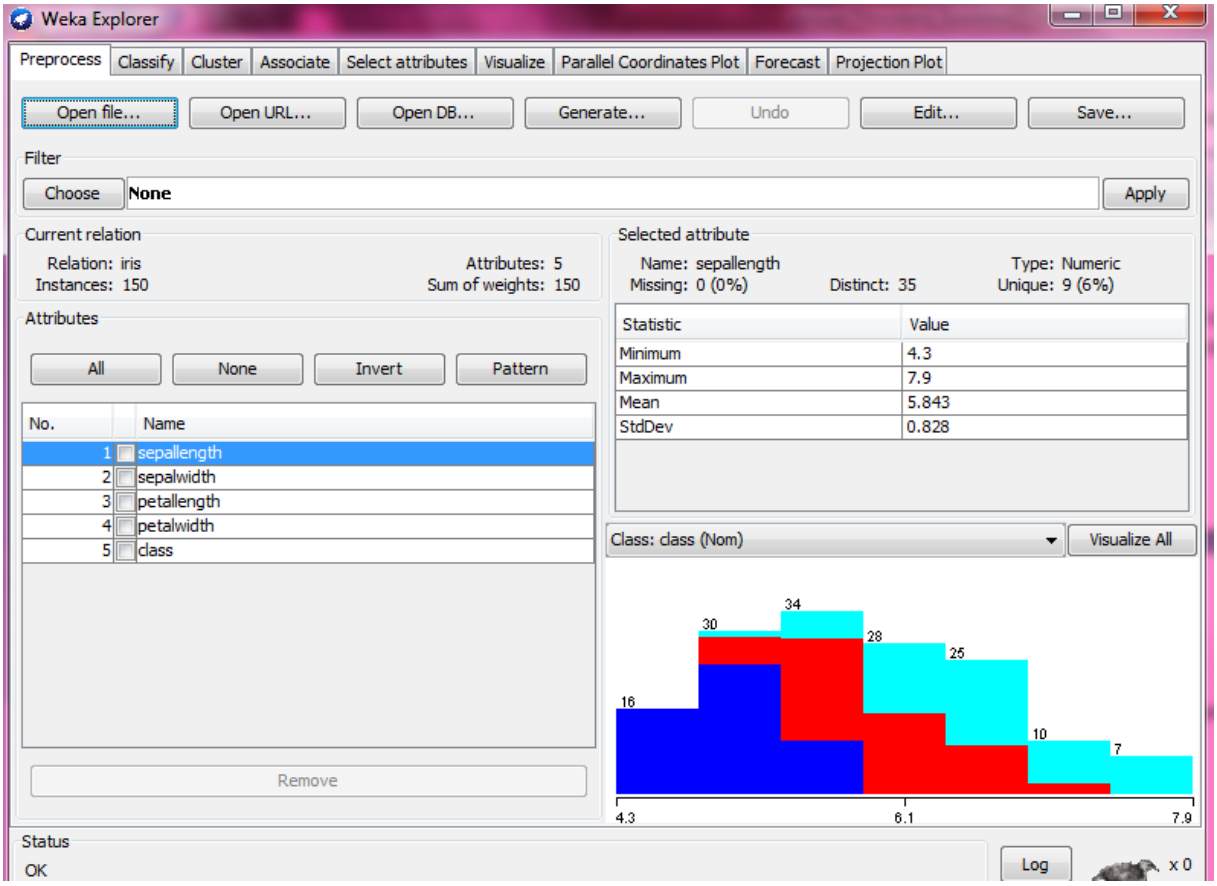
Εικόνα 22: Τυποποίηση .arff

Το πρώτο βήμα είναι να ανοίξουμε το αρχείο .csv που θέλουμε να μετατρέψουμε, με έναν επεξεργαστή κειμένου. Επόμενο βήμα είναι να προσθέσουμε το όνομα του σετ δεδομένων σε μια γραμμή στην αρχή η οποία θα αρχίζει με την έκφραση @relation, τις πληροφορίες για το κάθε χαρακτηριστικό χρησιμοποιώντας την έκφραση @attribute (βάζοντας κάθε χαρακτηριστικό σε νέα γραμμή) και μια σειρά με την έκφραση @data, που δείχνει ότι από εκεί και κάτω βρίσκονται τα δεδομένα. Τέλος, κάνουμε αποθήκευση του αρχείου στην μορφή .arff. Στο παράρτημα στο τέλος της εργασίας παρατηρούμε ένα αρχείο .arff όπως αυτό προκύπτει από τον μετατροπέα του Weka.

4.4 Εφαρμογή στο Weka

Χρησιμοποιούμε, αρχικά, μια βάση δεδομένων από τις έτοιμες που μας δίνει το πρόγραμμα Weka (iris). Η βάση αυτή περιέχει 150 στοιχεία και 5 χαρακτηριστικά (attributes).

Αφού εισάγουμε τα δεδομένα εκπαίδευσης (training set) στο πρόγραμμα στην καρτέλα Preprocess εμφανίζονται τα παρακάτω στατιστικά δεδομένα:



Εικόνα 23: Περιβάλλον Preprocess στο Weka

Μετά την προεπεξεργασία των δεδομένων εκπαίδευσης μας δίνονται τα εξής αποτελέσματα:

- ✓ Minimum (ελάχιστο). Είναι η ελάχιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων.

  Minimum = 4,3

- ✓ Maximum (μέγιστο). Είναι η μέγιστη τιμή που εντοπίστηκε στο σύνολο των δεδομένων.

  Maximum = 7,9

- ✓ Mean (μέση τιμή). Είναι η μέση τιμή του συνόλου των δεδομένων.

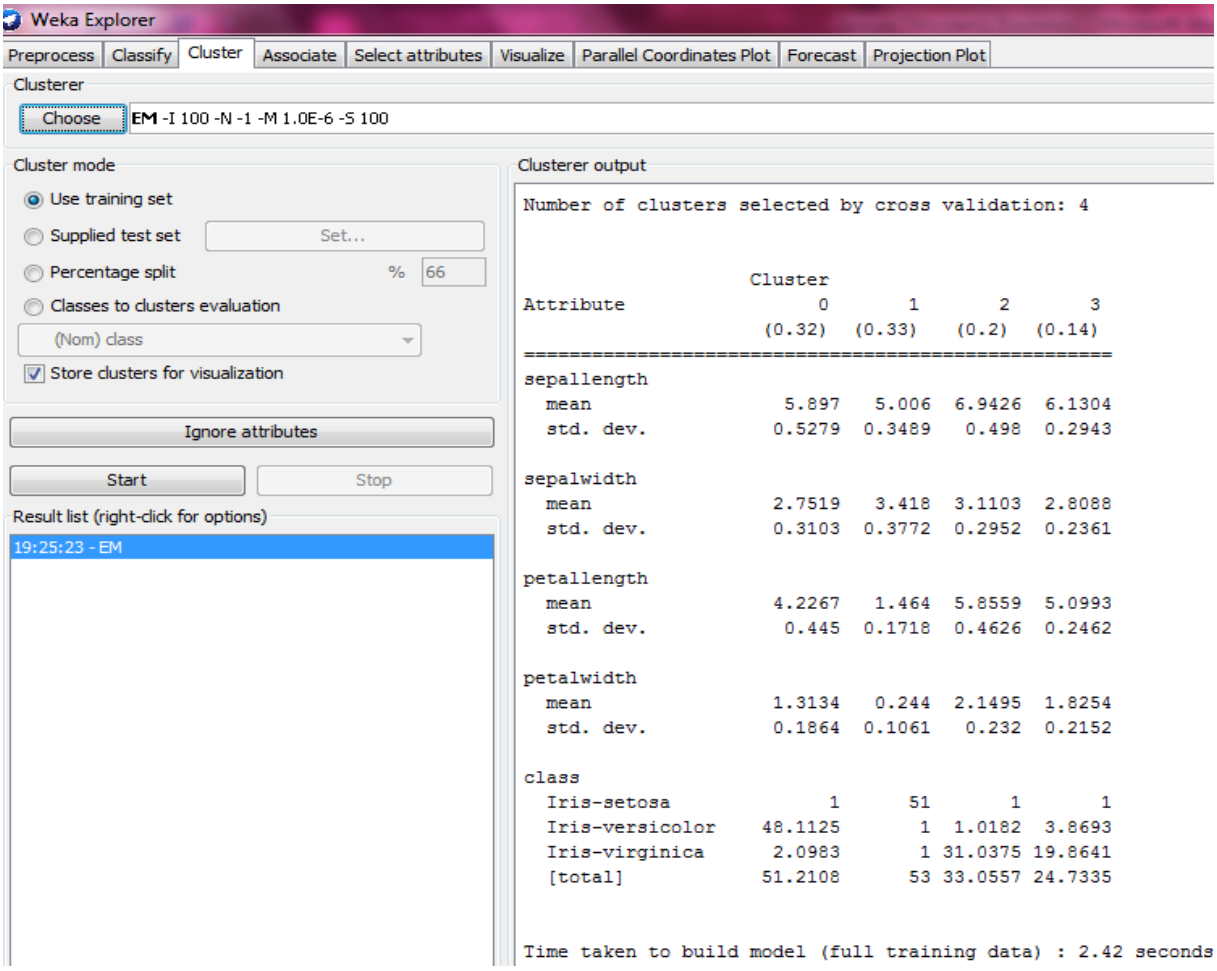
Mean = 5,843

- ✓ StdDev (τυπική απόκλιση). Είναι η τυπική απόκλιση του συνόλου των δεδομένων.

StdDev = 0,828

Επιπλέον, γνωρίζουμε ότι τα δεδομένα που βρίσκονται στην ίδια κλάση ταξινομούνται με το ίδιο χρώμα. Στη συγκεκριμένη περίπτωση τα δεδομένα που έχουμε ταξινομούνται σε τρείς κλάσεις, όπως φαίνεται και στο διάγραμμα.

Στη συνέχεια, επιλέγουμε Cluster για να εφαρμόσουμε ομαδοποίηση στα δεδομένα εκπαίδευσης. Αφήνοντας την επιλογή *use training set* και πατώντας *start* ξεκινάει η διαδικασία.



Εικόνα 24: Καρτέλα Cluster στο Weka

Όπως φαίνεται και παραπάνω, οι κλάσεις που δημιουργήθηκαν για κάθε μια από τις πέντε μεταβλητές (attributes) είναι τέσσερις, καθώς επίσης, φαίνεται η μέση τιμή και η τυπική απόκλιση αντίστοιχα.

5ο Κεφάλαιο

Συμπεράσματα

Είναι γεγονός ότι ζούμε στην κοινωνία της πληροφορίας καθώς τεράστιος όγκος δεδομένων έχει αποθηκευτεί σε βάσεις δεδομένων. Το γεγονός αυτό οφείλεται στη διαρκή εξέλιξη της τεχνολογίας, και έτσι προβάλει επιτακτική η ανάγκη της μετατροπής των δεδομένων σε πληροφορία, πράγμα το οποίο αποτελεί προαπαιτούμενο βήμα για την μετατροπή της πληροφορίας σε γνώση. Η ανάγκη αυτή οδήγησε στην διαδικασία της Εξόρυξης Δεδομένων (Data Mining). Η εξόρυξη δεδομένων αποτελεί ένα μεμονωμένο και ταχέως αναπτυσσόμενο πεδίο. Η εξόρυξη γνώσης από βάσεις δεδομένων επικεντρώνεται στην εξαγωγή ενδιαφέρουσας και μη προφανούς πληροφορίας με σκοπό την κατανόηση της συμπεριφοράς των καταναλωτών ή και άλλων παραμέτρων που ενδιαφέρουν τον εκάστοτε ερευνητή. Αυτό επιτυγχάνεται με τη χρήση των αλγορίθμων που αναλύθηκαν στο Κεφάλαιο 2 οι οποίοι εντάσσονται σε τέσσερις βασικές τεχνικές.

Η Κατηγοριοποίηση αποτελεί μια από τις βασικότερες εργασίες εξόρυξης δεδομένων. Γίνεται εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένου), το οποίο, με βάση τα χαρακτηριστικά αυτά, αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαριστάνονται γενικά από τις εγγραφές της βάσης δεδομένων κι έτσι, η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κατηγορίες. Οι βασικές μέθοδοι που χρησιμοποιούνται είναι: Μέθοδος Bayers, Δέντρα Αποφάσεων, Αλγόριθμοι διανυσμάτων Υποστήριξης , Νευρωνικά Δίκτυα.

Η Συσταδοποίηση είναι η εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων. Στην κατηγοριοποίηση, ο πληθυσμός διαιρείται σε κατηγορίες, αναθέτοντας κάθε στοιχείο ή εγγραφή σε μια προκαθορισμένη κατηγορία, με βάση ένα μοντέλο που αναπτύσσεται μέσω της εκπαίδευσής του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων. Αντιθέτως, στη συσταδοποίηση, οι εγγραφές ομαδοποιούνται σε σύνολα, με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Αξίζει να σημειωθεί πως μπορεί να χρησιμοποιηθεί και σαν εισαγωγή σε κάποια άλλη διαδικασία εξόρυξης γνώσης ή μοντελοποίησης. Αντιπροσωπευτικοί αλγόριθμοι: K- Means και παραλλαγές, PAM, DBSCAN, COBWEB.

Οι Κανόνες Συσχέτισης θεωρούνται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων, καθώς παρέχουν ένα συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους τελικούς χρήστες. Πιο συγκεκριμένα, οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Οι συσχετισμοί αυτοί έχουν τη μορφή « If *A* then *B*», όπου το *A* και το *B* αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα. Τέλος, η ανάλυση συσχέτισης είναι γνωστή στον επιχειρηματικό κόσμο και ως *ανάλυση συνάφειας* με πολλές εφαρμογές.

Η Παλινδρόμηση είναι η παλαιότερη και η πιο γνωστή στατιστική τεχνική που υλοποιείται εντός των πλαισίων της εξόρυξης δεδομένων. Χρησιμοποιώντας μια βάση δεδομένων, αναπτύσσεται μια μαθηματική σχέση που ταιριάζει στα δεδομένα αυτά, η οποία βοηθά στην πρόβλεψη μελλοντικής συμπεριφοράς, εφαρμόζοντας σε αυτή νέα αριθμητικά δεδομένα. Η συγκεκριμένη τεχνική εφαρμόζεται καλά μόνο σε συνεχή ποσοτικά δεδομένα (ζήτηση, δαπάνες διαφήμισης), ενώ δεν λειτουργεί καλά με κατηγορικά δεδομένα.
Τα οφέλη της εξόρυξης δεδομένων γίνονται περισσότερο κατανοητά από την εφαρμογή της σε επιχειρησιακά προβλήματα- πραγματικά δεδομένα όπως αναλύθηκε στο Κεφάλαιο 3.Για την ανάλυση χρησιμοποιήθηκε το εργαλείο CRISP-DM το οποίο αποτελείται από έξι φάσεις: Φάση κατανόησης έρευνας/επιχείρησης, Φάση κατανόησης δεδομένων, Φάση προετοιμασίας δεδομένων, Φάση μοντελοποίησης, Φάση αξιολόγησης, Φάση ανάπτυξης.

Τέλος, στο Κεφάλαιο 4 έγινε εκτενής ανάλυση του λογισμικού Weka. Από έρευνες προέκυψε ότι δεν υπάρχει κανένας αλγόριθμος εξόρυξης δεδομένων που να εφαρμόζεται σε όλες τις περιπτώσεις. Τα σύνολα δεδομένων που συναντώνται στην πράξη ποικίλουν ευρέως και για να εξάγει κανείς ακριβή μοντέλα απαιτείται τα χαρακτηριστικά του αλγορίθμου εκμάθησης να ταιριάζουν με τη δομή του πεδίου εφαρμογής. Άλλωστε η εξόρυξη γνώσης από δεδομένα αποτελεί μια πειραματική επιστήμη και στηρίζεται στην εφαρμογή της μεθόδου *trial and error*. Το πακέτο Weka αποτελεί μια συλλογή από τους πλέον σύγχρονους αλγορίθμους μηχανικής μάθησης. Μεταξύ άλλων περιέχει μεθόδους για: Προεπεξεργασία δεδομένων, Οπτικοποίηση, Ταξινόμηση, Ομαδοποίηση, Εύρεση κανόνων συσχέτισης, Παλινδρόμηση. Για αυτό το λόγο το πακέτο Weka θεωρείτε ένα πολύ ισχυρό εργαλείο το οποίο μπορεί να επεξεργάζεται μεγάλες βάσεις δεδομένων, να αναπαριστά γραφικά τα σύνολα των δεδομένων και να εξάγει χρήσιμα συμπεράσματα με βάσει τα κριτήρια του εκάστοτε χρήστη.

Ευρετήριο Εικόνων

Ελληνική Βιβλιογραφία:

1. *Γουρδούλης Ι, Πάτρα, 2009, Αλγόριθμοι Εξόρυξης Δεδομένων για Χειρισμό Πολλαπλών Υποστηρίξεων και Αρνητικών Συσχετίσεων, Διπλωματική Εργασία.*
2. *Βαρσάμη Ε, Πάτρα, 2010, Εφαρμογή Μεθόδων Εξόρυξης Δεδομένων σε Βαρομετρικούς Χάρτες, Διπλωματική Εργασία.*
3. *Ντάλλα Μ, Πάτρα, 2009, Εφαρμογή Αλγορίθμων Επαγωγικού Λογικού Προγραμματισμού στη Σχεδίαση Εξόρυξης Δεδομένων, Μεταπτυχιακή Εργασία.*
4. *Κουρής Ν, Πάτρα, 2006, Εφαρμογή Τεχνικών Data Mining σε Συστήματα Ηλεκτρονικού Εμπορίου, Διδακτορική Διατριβή.*
5. *Μεττούρης Χ, Πάτρα, 2008, Υλοποίηση Εφαρμογής Εξόρυξης Δεδομένων σε Αποτελέσματα Εντοπισμού της Θέσης Κινητού Χρήστη και Αξιοποίηση της Πληροφορίας σε M-commerce Εφαρμογές, Διπλωματική Εργασία.*
6. *Μαστρογιάννης Ν, Πάτρα, 2009, Μεθοδολογικό Πλαίσιο Υποστήριξης της Εξόρυξης Γνώσης από Δεδομένα με την Χρήση Αρχών της Πολυκριτήριας Ανάλυσης Αποφάσεων, Διδακτορική Διατριβή.*
7. *Καραολής Μ, Κύπρος, 2010, Εξόρυξη Γνώσης με Εξαγωγή Κανόνων σε Καρδιαγγειακές Βάσεις Δεδομένων, Διδακτορική Διατριβή, Διαθέσιμο: http://www.medinfo.cs.ucy.ac.cy/doc/Publications/PhD/MKaraolis/PhD_Mina_Karaolis.pdf.*
8. *Γολέμη Ε, Πάτρα, 2010, Κρυπτογραφία και Εξόρυξη Δεδομένων, Μεταπτυχιακή Εργασία.*
9. *Ουγιάρογλου Σ, Θεσσαλονίκη, 2006, Κατηγοριοποίηση με Βάση Δυναμικό Αριθμό Κοντινότερων Γειτόνων, Διπλωματική Εργασία, Διαθέσιμο: http://users.sch.gr/stoug/papers/final_work_msc.pdf*
10. *Θεοδωρίδης Ι, Αθήνα, 1996, Χωρικές Δομές Δεδομένων: Αναλυτικά Μοντέλα και Αποδοτικοί Αλγόριθμοι, Διδακτορική Διατριβή, Διαθέσιμο: http://www.dbnet.ece.ntua.gr/pubs/uploads/PHD-1996-1.pdf.*
11. *Τσεκούρας Γ, Αθήνα, 2004, Εφαρμογή της Μεθόδου των Τεχνιτών Νευρωνικών Δικτύων σε Θέματα Σηράγγων, Διπλωματική Εργασία, Διαθέσιμο: http://artemis-new.cslab.ece.ntua.gr:8080/jspui/handle/123456789/4294?mode=full*

12. *Δημητρακοπούλου Κ, Πάτρα, 2007, Αναγνώριση Λειτουργικών Υπο-δομών στο Πρωτεϊνικό Δίκτυο του Saccharomyces Cerevisiae Συνδυάζοντας Δεδομένα Έκφρασης Γονιδίων και Αλληλεπίδρασης Πρωτεϊνών, Διπλωματική Εργασία.*
13. *Παγουρόπουλος Α, Πάτρα, 2006, Data Mining στην Χρηματοοικονομική Ανάλυση, Διπλωματική Εργασία.*
14. *Τσαρακτσίδης Γ, Θεσσαλονίκη, 2008, Εξόρυξη Γνώσης από Βάση Δεδομένων Ηλεκτρονικών Δημοπρασιών από τον Δικτυακό Τόπο e-Bay, Διπλωματική Εργασία, Διαθέσιμο: http://vivliothmmy.ee.auth.gr/76/*
15. *Παναγιωτάκος Θ, Αθήνα, 2012, Τεχνικές Εξόρυξης Δεδομένων Data Mining, Διπλωματική Εργασία,Διαθέσιμο: http://dspace.lib.ntua.gr/handle/123456789/6493?mode=full.*
16. *Μάσσου Λ, Αθήνα, 2008, Αλγόριθμοι Εξόρυξης Πληροφορίας (Εφαρμογή στο Weka), Εργασία,Διαθέσιμο: http://datamininggntua.files.wordpress.com/2008/05/cea4ce95ce9bce99ce9ace97_ce95cea1ce93ce91cea3ce99ce911.pdf.*
17. *Καμπυλαυκά Ι, Θεσσαλονίκη, 2011, Εφαρμογή Τεχνικών Εξόρυξης Δεδομένων για την Ομαδοποίηση και τον Χαρακτηρισμό Καταναλωτών Ηλεκτρικής Ενέργειας, Διπλωματική Εργασία, Διαθέσιμο: http://vivliothmmy.ee.auth.gr/1232/.*
18. *Αντζουλάτος Γ, Πάτρα, 2005, Εφαρμογές αλγορίθμων και έλεγχοι αξιοπιστίας ομαδοποίησης στην αναγνώριση προτύπων και στον καθορισμό δεδομένων. Διπλωματική εργασία..*
19. *Βαζιργιάννης Μ, Χαλκίδη Μ, Αθήνα, 2003, Εξόρυξης Γνώσης από Βάσεις Δεδομένων. Τυπωθήτω – Γιώργος Δάρδανος.*

Διεθνής Βιβλιογραφία:

20. *P. Tan, M. Steinbach, V. Kumar, 2005, Introduction to Data Mining. Addison Wesley.*
21. *M. Kantardzic, 2002, Data Mining: Concepts, Models, Methods, and Algorithms, Wiley.*
22. *Ian H. Witten, Eibe Frank, Mark A. Hall, 2011, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann.*

23. *Osmar R. Zaïane, Alberta, 1999, Principles of Knowledge Discovery in Databases, Διαθέσιμο: http://www.exinfm.com/pdffiles/intro_dm.pdf.*
24. *Markus Hegland, 2003, Data Mining – Challenges, Models, Methods and Algorithms. Διαθέσιμο: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.4513.*
25. *Jeffrey W. Seifert, 2004, Data Mining: An Overview, CRS Report for Congress, Διαθέσιμο: http://www.fas.org/irp/crs/RL31798.pdf.*
26. *Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, 1996, From Data Mining to Knowledge Discovery in Databases, Διαθέσιμο: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.*
27. *Daniel t. Larose, 2005, Discovering Knowledge in Data an Introduction to Data Mining, Διαθέσιμο: http://www.dss.dpem.tuc.gr/pdf/Interscience%20Discovering%20Knowledge%20in%20Data%20-%20An%20Introduction%20to%20Data%20Mining.pdf.*