ΤΕΙ ΠΑΤΡΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΥ ΣΧΕΔΙΑΣΜΟΥ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

« Αλγόριθμοι Ομαδοποίησης στην Εξόρυξη Δεδομένων (Data Mining): περιγραφή, αξιολόγηση και εφαρμογές σε τεχνητές και πραγματικές βάσεις δεδομένων»

Επιβλέπων καθηγητής: Νικόλαος Μαστρογιάννης

ΚΑΛΟΥΔΗ ΒΕΡΟΝΙΚΑ

ΣΦΥΡΟΓΙΑΝΝΑΚΗ ΕΙΡΗΝΗ

ΤΣΟΥΡΔΙΟΥ ΑΝΑΣΤΑΣΙΑ

ΠΑΤΡΑ 2009

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία αναφέρεται στη διαδικασία εξόρυξης γνώσης από βάσεις δεδομένων περιγράφοντας, αξιολογώντας και εφαρμόζοντας αλγορίθμους ομαδοποίησης τόσο σε τεχνητές όσο και σε πραγματικές βάσεις δεδομένων.

Συγκεκριμένα αποτελείται από τις ακόλουθες τέσσερις ενότητες:

Ενότητα Α: Στο μέρος αυτό γίνεται μία πρώτη προσέγγιση της έννοιας της εξόρυξης δεδομένων. Παρουσιάζονται σχετικοί ορισμοί, ιστορικές αναφορές, μέθοδοι ανακάλυψης γνώσης από βάσεις δεδομένων (KDD), καθώς επίσης εφαρμογές και τεχνικές εξόρυξης γνώσης.

Ενότητα Β: Στο σημείο αυτό αναφερόμαστε σε ένα σύνολο αλγορίθμων που ανήκουν στην τεχνική της ομαδοποίησης, η οποία και αποτελεί αναπόσπαστο κομμάτι της διαδικασίας εξόρυξης γνώσης από βάσεις δεδομένων.

Ενότητα Γ: Η ενότητα αυτή είναι άρρηκτα συνδεδεμένη με την προηγούμενη καθώς ταξινομούνται οι αλγόριθμοι ομαδοποίησης σε τέσσερις βασικές κατηγορίες ώστε να συγκριθούν μεταξύ τους και να προκύψουν συμπεράσματα για εξόρυξη χρήσιμης γνώσης από βάσεις δεδομένων.

Ενότητα Δ: Ολοκληρώνοντας την εργασία προχωρήσαμε στην εφαρμογή τριών σημαντικών αλγορίθμων της τεχνικής της ομαδοποίησης, τους DBSCAN, k-Means και k-Modes πάνω σε συγκεκριμένες βάσεις δεδομένων. Σκοπός της εφαρμογής αυτής είναι η ανακάλυψη γνώσης από τεχνητές και πραγματικές βάσεις δεδομένων.

ΕΝΟΤΗΤΑ Α΄

1.1 Η ΕΠΟΧΗ ΤΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Βρισκόμαστε σε μία εποχή όπου καλούμαστε καθημερινά να επεξεργαστούμε και να αναλύσουμε χιλιάδες δεδομένα από το γύρω περιβάλλον μας ανεξάρτητα από το επιστημονικό πεδίο που ανήκει ο καθένας από εμάς.

Κυρίαρχο ρόλο στον σκοπό αυτό, έπαιξε η τεχνολογία στα ηλεκτρονικά συστήματα η οποία άρχισε να εξελίσσεται ραγδαία με αποτέλεσμα σήμερα οι βάσεις δεδομένων που αποθηκεύουν πλήθος στοιχείων να έχουν αναπτυχθεί ευρέως. Με την ανάπτυξη των κατάλληλων μέσων και εργαλείων μπορούμε να μετατρέπουμε εύκολα, τα όποια δεδομένα λαμβάνουμε από οποιαδήποτε βάση δεδομένων, σε χρήσιμη πληροφορία και γνώση, αυτόματα και ευφυέστατα.

Όσον αφορά το μέγεθος των δεδομένων που παράγονται ετησίως, τόσο από εταιρίες όσο και από πανεπιστήμια αντιστοιχεί σ’ ένα τεράστιο όγκο δεδομένων της τάξεως των terabytes. Το πρόβλημα λοιπόν που παρουσιάζεται, είναι πώς θα τα επιλέξουμε, πώς θα τα ομαδοποιήσουμε και θα τα ταξινομήσουμε και τέλος ,πώς θα τα συσχετίσουμε μεταξύ τους.

Τη λύση στο πρόβλημα ήρθε να δώσει η αποθήκη βάσεων δεδομένων(data warehouse) και η εξόρυξή τους από αυτή, προκειμένου να μπορούμε εύκολα να εξάγουμε γνώση.

Αναφερόμενοι στην εξόρυξη δεδομένων (data mining),εννοούμε την εξαγωγή κρυμμένης πληροφορίας από μεγάλες βάσεις δεδομένων, όπου είναι μια νέα τεχνολογία με μεγάλη

δυναμική που βοηθά τις επιχειρήσεις να επικεντρωθούν στη σημαντική πληροφορία που βρίσκεται μέσα στις αποθήκες δεδομένων τους (data warehouses).

Γι' αυτό το λόγο αποτελεί πλέον μια περιοχή έρευνας με ολοένα αυξανόμενη σημασία. Είναι μια συνολική συνεργασία μεταξύ ανθρώπων και υπολογιστών όπου οι άνθρωποι σχεδιάζουν τεράστιες βάσεις δεδομένων περιγράφοντας προβλήματα και θέτοντας στόχους και οι υπολογιστές εξετάζουν προσεχτικά τα δεδομένα αναζητώντας πρότυπα που ταιριάζουν στους στόχους των ανθρώπων.

Οι τεχνικές Εξόρυξης Δεδομένων προβλέπουν μελλοντικές τάσεις και συμπεριφορές επιτρέποντας στις επιχειρήσεις να πάρουν σημαντικές αποφάσεις, καθοδηγούμενες από τη γνώση. Επίσης, μπορούν να απαντήσουν σε επιχειρηματικά ερωτήματα που παλαιότερα κατανάλωναν πολύ χρόνο για να επιλυθούν. Ψάχνουν δηλαδή λεπτομερειακά και γρήγορα βάσεις δεδομένων για την αναζήτηση κρυμμένων προτύπων (patterns), ανακαλύπτοντας πληροφορίες που οι ειδικοί μπορεί να χάσουν καθώς δεν είναι στις προσδοκίες τους.

Οι τεχνολογίες για τη δημιουργία και τη συλλογή δεδομένων αναπτύσσονται ταχύτατα την τελευταία δεκαετία. Έρευνες, συνέδρια, και διατριβές εμφανίζονται στο προσκήνιο κάθε χρόνο ολοένα και περισσότερο.

Μόλις πριν από 7 χρόνια, μόνο 50 ερευνητές συμμετείχαν στο συνέδριο του « Knowledge Discovery and Data Mining Workshop». Σήμερα το γνωστό ηλεκτρονικό περιοδικό έχει περισσότερους από 4000 αναγνώστες. Σύμφωνα με μία έκθεση της Gartner group η Εξόρυξη Δεδομένων και η Τεχνητή Νοημοσύνη βρίσκονται στην κορυφή των 5 σπουδαιότερων τεχνολογιών-κλειδιά που υπολογίζεται να έχουν σημαντικό αντίκτυπο σε ένα ευρύ φάσμα βιομηχανιών τα επόμενα 3-5 χρόνια (Weiss, 1997).

1.1.1 **Προσέγγιση της έννοιας «Εξόρυξη Δεδομένων»**

Πολλοί οργανισμοί, ιδιωτικές επιχειρήσεις μέχρι και κυβερνητικοί φορείς έχουν επενδύσει τεράστια ποσά σε έρευνες για την κατασκευή και τη συντήρηση μεγάλων βάσεων δεδομένων . Συχνά όμως τα δεδομένα είναι δύσκολο να αναλυθούν μέσω των στατιστικών απλών μεθόδων είτε γιατί έχουν χαθεί πολλές εγγραφές δεδομένων λόγω του τεράστιου όγκου αυτών είτε διότι τα δεδομένα είναι περισσότερο ποιοτικά παρά ποσοτικά.

Κάποιες φορές, τυχαίνει οι βάσεις δεδομένων να είναι τόσο μεγάλες που ακόμα και οι διαχειριστές αυτών δε γνωρίζουν τι είδους πληροφορία εμπεριέχεται στις βάσεις αυτές ή πόσο σχετική αυτή μπορεί να είναι με αυτά που ψάχνουν λόγω του ότι τα δεδομένα είναι δύσκολο να προσπελαστούν και να αναλυθούν.

Για τους παραπάνω λόγους λοιπόν χρησιμοποιούμε μια ποικιλία μεθοδολογιών εξόρυξης δεδομένων που χρησιμοποιούνται με σκοπό να αναλυθούν οι τεράστιες βάσεις δεδομένων ώστε να ανακαλυφθούν νέες τάσεις και πρότυπα (Braxton, 1998).

1.1.2 **Ορισμοί Εξόρυξης Δεδομένων**

«Εξόρυξη Δεδομένων είναι η διαδικασία επεξεργασίας και ανάλυσης δεδομένων με στόχο την εύρεση υπονοούμενης, αλλά ενδεχομένως χρήσιμης γνώσης, που χρησιμοποιεί έναν αριθμό από διαφορετικές τεχνικές, όπως: ομαδοποίηση, ταξινόμηση, εύρεση εξαρτημένων δικτύων, ανάλυση αλλαγών και εύρεση ανωμαλιών. Περιλαμβάνει δηλαδή τη συλλογή, την εξέταση και τη μοντελοποίηση μεγάλων ποσοτήτων δεδομένων για την αποκάλυψη αγνώστων προτύπων και σε τελευταία ανάλυση κατανοητής πληροφόρησης από μεγάλες βάσεις δεδομένων» (Frawley, Piatetsky-Shapiro, Matheus).

« Εξόρυξη Δεδομένων είναι η αναζήτηση σχέσεων και γενικών προτύπων που υπάρχουν σε μεγάλες βάσεις δεδομένων αλλά είναι κρυμμένες ανάμεσα σε τεράστιες ποσότητες δεδομένων όπως για παράδειγμα η σχέση ανάμεσα στα δεδομένα των ασθενών και στις ιατρικές διαγνώσεις τους. Αυτές οι σχέσεις αντιπροσωπεύουν πολύτιμη γνώση για τα βάση δεδομένων καθώς επίσης και των αντικειμένων σε αυτή» (Holshemier, Siebes).

Αυτή είναι λοιπόν μια γενική ιδέα για το τι είναι και με τι ασχολείται η επιστήμη της εξόρυξης δεδομένων μέσα από βάσεις δεδομένων. Πιο συγκεκριμένα, η εξόρυξη δεδομένων αφορά σύνολα ανόμοιων καταστάσεων στα οποία μπορούμε να συμπεριλάβουμε μαθηματικές ή στατιστικές αναλύσεις λαμβάνοντας υπ' όψη συγκεκριμένες υποθέσεις ώστε να είμαστε ικανοί να εξετάσουμε αναλυτικά σενάρια που θα μας οδηγήσουν σε μία ενδιαφέρουσα έκβαση.

Πρόκειται για μια διαδικασία επαναληπτική μέσα στην οποία η πρόοδος καθορίζεται από την ανακάλυψη προτύπων σε τεράστιες βάσεις δεδομένων. Με τον όρο πρότυπα αναφερόμαστε σε ένα τύπο επαναλαμβανόμενων γεγονότων ή αντικειμένων που αποτελούν στοιχεία ενός συνόλου τα οποία αυτά στοιχεία επαναλαμβάνονται κατά τρόπο προβλέψιμο. Η εξαγωγή προτύπων (patterns) είναι ένα σημαντικό συστατικό για κάθε δραστηριότητα της Εξόρυξης δεδομένων και έχει να κάνει με σχέσεις μεταξύ υποσυνόλων των δεδομένων.

Συνήθως, αρχίζουμε τη διαδικασία με το να πάρουμε μια γενική εικόνα των διαθέσιμων στοιχείων. Αυτό συνεπάγεται μια σειρά βημάτων στην οποία τα υποσύνολα των στοιχείων διαμορφώνονται και αναλύονται. Με βάση την ανακάλυψη των προτύπων που μας ενδιαφέρουν μπορεί να υπάρξει επόμενη λήψη δείγματος ενός συνόλου στοιχείων μαζί με τη διατύπωση των νέων προτύπων με σκοπό να υπογραμμιστούν οι ιδιαίτερες πτυχές των στοιχείων και ούτω καθ' εξής.

Η διαδικασία επαναλαμβάνεται μέχρις ότου εξαντληθεί. Αφού εξαντληθεί μπορούμε να προχωρήσουμε σε νέα εξόρυξη δεδομένων με νέα πρότυπα. Γενικότερα όποια και αν είναι η μορφή της ανάλυσης , το κλειδί για την επιτυχία είναι η υιοθεσία μιας εύκαμπτης προσέγγισης που θα μας επιτρέπει να κάνουμε απροσδόκητες ανακαλύψεις πάνω στο πρόβλημα που εξετάζουμε πέρα από τα όρια των καθιερωμένων προσδοκιών.

Μόνο όταν εξετάζονται τα στοιχεία από πολλές διαφορετικές απόψεις μπορούν να γίνουν ενδιαφέρουσες ανακαλύψεις. Η εξόρυξη δεδομένων περιστοιχίζεται από μια ευρεία οικογένεια υπολογιστικών μεθόδων που περιλαμβάνουν τη στατιστική ανάλυση, τα δένδρα αποφάσεων, τα νευρωνικά δίκτυα, την επαγωγή κανόνων, και τη γραφική οπτικοποίηση.

Παρ’ όλο που οι τεχνικές εξόρυξης είναι διαθέσιμες εδώ και πολύ καιρό, οι πρόοδοι στο λογισμικό και στο υλικό τμήμα των υπολογιστών έχουν μετατρέψει την εξόρυξη δεδομένων σε μία πιο ελκυστική και πρακτική διαδικασία (Βαζιργιάννης,2003).

1.1.3 **Τα θεμέλια της εξόρυξης δεδομένων**

Οι τεχνικές της εξόρυξης δεδομένων είναι αποτέλεσμα μιας μακράς διαδικασίας έρευνας και ανάπτυξης προϊόντων. Η εξέλιξη αυτή άρχισε όταν επιχειρηματικά δεδομένα εγκαταστάθηκαν στους υπολογιστές, συνεχίστηκε με βελτιώσεις στην προσπέλαση δεδομένων και πλέον με αυτοματοποιημένες διαδικασίες επιτρέπεται στους χρήστες των συστημάτων αυτών να εισάγουν τεράστιες ποσότητες δεδομένων και να τα επεξεργάζονται σε πραγματικό χρόνο (real time) [3]. Στη μετάβαση από τα επιχειρηματικά δεδομένα στην επιχειρηματική γνώση, κάθε καινούριο βήμα βασίστηκε στο προηγούμενο.

- 1960

Συλλογή Δεδομένων και Επεξεργασία αρχείων. Η χρησιμοποιούμενη τεχνολογία ήταν υπολογιστές, κασέτες και δίσκοι.

- 1970-1980

Πρόσβαση στα δεδομένα μέσω Ιεραρχικών και δικτυακών μοντέλων, σχεσιακών συστημάτων βάσεων δεδομένων, εργαλείων μοντελοποίησης, γλωσσών επερωτήσεων SQL, συναλλαγών, ανάκαμψη σφαλμάτων, χρήση OLAP(on-line analytical processing)

- 1980-1990

Data Warehousing & Decision Support. Εμφάνιση μεγάλων βάσεων δεδομένων. Χρησιμοποιούνται νέα μοντέλα (αντικειμενοσχεσιακό, εκτεταμένα σχεσιακά μοντέλα κλπ), νέες εφαρμογές και νέοι τύποι δεδομένων (χώρο-χρονικά, δεδομένα από αισθητήρες, συνεχή δεδομένα, κλπ)

- 1990-2009

Data Mining. Χρήση εξελιγμένων αλγορίθμων, παράλληλη επεξεργασία δεδομένων και τεράστιες βάσεις δεδομένων (Dunham,2004).

1.2 Δεδομένα

Τα δεδομένα τα οποία εξετάζουμε μπορεί να είναι γεγονότα, αριθμοί ή κείμενο που μπορούν να υποβληθούν σε επεξεργασία από ένα υπολογιστή. Σήμερα οι οργανώσεις συσσωρεύουν τα συνεχώς αυξανόμενα ποσά στοιχείων με διαφορετικά σχήματα σε διαφορετικές βάσεις δεδομένων.

Αυτό περιλαμβάνει μεν, λειτουργικά δεδομένα (ή δεδομένα συναλλαγών) όπως είναι οι πωλήσεις, το κόστος, η μισθοδοτική κατάσταση ,τα λογιστικά και γενικότερα τέτοιες λειτουργίες αλλά και μη-λειτουργικά δεδομένα όπως επίσης και μετα-δεδομένα.

Με τον όρο μη-λειτουργικά δεδομένα, αναφερόμαστε στις βιομηχανικές πωλήσεις, και διάφορα προβλεπόμενα μακροοικονομικά στοιχεία.

Αναφερόμενοι στα μετά-δεδομένα, εννοούμε δεδομένα για τα ίδια τα δεδομένα όπως για παράδειγμα οι λογικοί ορισμοί των λεξικών του σχεδίου βάσεων δεδομένων ή των στοιχείων.

Συνήθως τα στοιχεία που εξετάζουμε αφορούν:

- Τράπεζες(συναλλαγές πελατών, δάνεια, κλπ)
- Τηλεπικοινωνίες(ομάδες πελατών, χρεώσεις, προγράμματα)
- Επιχειρήσεις και οργανισμούς(πωλήσεις, μισθοδοσία, ζημιές, παραγγελίες προϊόντων κλπ)
- Επιστημονικούς τομείς(αστρονομία, βιολογία, ιατρική κλπ)
- Κείμενα
- Παγκόσμιο Ιστό

- E-εμπόριο

(Θεώνη Πιτουρά, διαθέσιμο: [http://www.cs.uoi.gr/pitoura/courses/dm)

1.2.1 **Πηγές Δεδομένων**

Στον πραγματικό κόσμο οι πηγές, από όπου μπορούμε να αντλήσουμε τα δεδομένα που χρειαζόμαστε, να τα αναλύσουμε και να τα επεξεργαστούμε ώστε να παράγουμε την απαραίτητη γνώση, είναι αμέτρητες. Αυτές χωρίζονται σε δύο κατηγορίες, τις εξωτερικές και τις εσωτερικές.

Όσον αφορά τις εξωτερικές πηγές, αναφερόμαστε σε κάθε είδους πληροφορία που βρίσκουμε από εξωτερικούς παράγοντες συνήθως ερευνώντας βάση κάποιου κριτηρίου ώστε να διαλέξουμε μόνο την απαραίτητη πληροφορία που μας χρειάζεται για τα πλαίσια κάποιας εργασίας που διεκπεραιώνουμε. Είναι πιθανό ωστόσο να συναντήσουμε προβλήματα νομικής φύσεως ως προς το είδος των πηγών που επιλέγουμε να συνδυάσουμε προς έρευνα επειδή κάποιες ίσως περιέχουν πληροφορίες ιδιωτικές και απόρρητες στο ευρύ κοινό.

Όσον αφορά τις εσωτερικές πηγές δεδομένων αναφερόμαστε στις αποθήκες δεδομένων (data warehouses), ένα σύνολο από ολοκληρωμένα δεδομένα ίδιας φύσεως που αποθηκεύονται χωρίς να διαγράφονται σε βάθος χρόνου σε βάσεις δεδομένων ώστε να μπορεί ο μελετητής να εξετάζει αυτά τα δεδομένα και να εξάγει την επιθυμητή προς αυτόν, κάθε φορά, γνώση (Inmon).

1.3 Ανακάλυψη γνώσης από βάσεις δεδομένων (KDD διαδικασία)

Είναι γεγονός πως υπάρχει αρκετή σύγχυση μεταξύ των ορισμών ανακάλυψη γνώσης από βάσεις δεδομένων και εξόρυξης γνώσης από αυτές. Ας ξεκαθαρίσουμε λοιπόν ότι η εξόρυξη δεδομένων αποτελεί αναπόσπαστο κομμάτι της διαδικασίας ανακάλυψης γνώσης από μεγάλες βάσεις δεδομένων.

1.3.1 **Ορισμός**

« *Η KDD διαδικασία είναι μια ντετερμινιστική επαναληπτική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και κατανοητών προτύπων στα δεδομένα που εξετάζουμε» (Frawley, Piatesky-Shapiro & Matheus 1991).*

1.3.2 **Διαδικασία KDD**

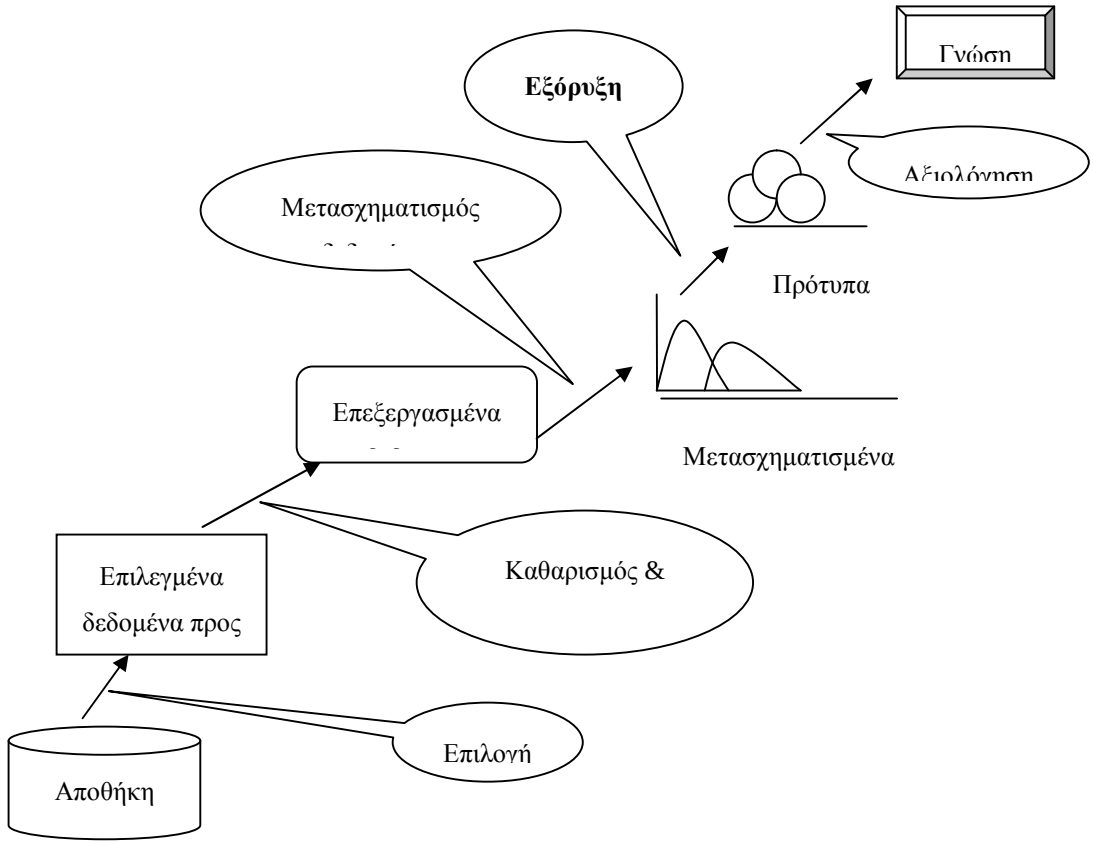
Μία διαδικασία KDD αποτελείται από πολλαπλά βήματα η οποία περιλαμβάνει την προ-επεξεργασία των δεδομένων, την αναζήτηση των προτύπων και την αξιολόγηση της εξαγόμενης γνώσης.

Με τον όρο **πρότυπα** αναφερόμαστε σε ένα τύπο επαναλαμβανόμενων γεγονότων ή αντικειμένων που αποτελούν στοιχεία ενός συνόλου τα οποία αυτά στοιχεία επαναλαμβάνονται κατά τρόπο προβλέψιμο. Η εγκυρότητα των προτύπων αφορά κατά πόσο αυτά τα πρότυπα είναι συνεπή σε νέα δεδομένα με βάση πάντα, κάποιου βαθμού βεβαιότητας. Τα εξαγόμενα πρότυπα αξιολογούνται βάση κάποιων συναρτήσεων χρησιμότητας διατηρώντας όσο το δυνατόν περισσότερη γνώση από τα αρχικά δεδομένα η οποία μπορεί να φανεί χρήσιμη για τη λήψη αποφάσεων.

Χαρακτηριστικό παράδειγμα εδώ είναι η βάση δεδομένων των τραπεζών που αφορά τα δάνεια που παρέχονται στους πελάτες. Εδώ, σαν συνάρτηση χρησιμότητας θα μπορούσε να θεωρηθεί η εξής:

«Εάν έσοδα< $(t), τότε ο πελάτης δεν μπορεί να δανειστεί.»

Ο στόχος της όλης KDD διαδικασίας είναι να προσδιοριστούν τα πρότυπα και να γίνουν κατανοητά ώστε να μπορούν να οδηγήσουν ειδικούς και μη, σε χρήσιμα συμπεράσματα και αποφάσεις (Βαζιργιάννης 2003).



Σχήμα 1: Τα βήματα της διαδικασίας KDD

Το παραπάνω σχήμα απεικονίζει τη διαδικασία ανακάλυψης χρήσιμης γνώσης από βάσεις δεδομένων και ακολούθως αναλύεται βήμα προς βήμα η όλη διαδικασία.

1.3.2.1 Καθορισμός του προβλήματος

Ξεκινώντας την διαδικασία μας εστιάζουμε στα ποιό σημαντικά βήματα τα οποία είναι ο καθορισμός του προβλήματος με το οποίο θα ασχοληθούμε ,η οριοθέτηση του και ο σωστός προγραμματισμός.

Βάση των στόχων μας οι οποίοι αφορούν στο τί θέλουμε να επιτύχουμε ,ποιοί είναι οι διαθέσιμοι πόροι και τι χρονικούς περιορισμούς έχουμε, πρέπει να γίνει ο ανάλογος έλεγχος προκειμένου οι στόχοι αυτοί να είναι ιδιαίτερα αξιοποιήσιμοι.

Σκοπός της διαδικασίας μας λοιπόν, είναι να προκύπτουν αποτελέσματα σε πρακτικό και όχι σε θεωρητικό επίπεδο. Χαρακτηριστικό παράδειγμα αποτελούν οι επιχειρήσεις οι οποίες επιδιώκουν την απόκτηση οικονομικών οφελών και όχι απλώς γνώσης μέσω μίας επιστημονικής έρευνας..

1.3.2.2 Διαδικασία επιλογής συνόλου δεδομένων

- Αποθήκη Δεδομένων

Η διαδικασία ξεκινά από την αποθήκη των δεδομένων από όπου και θα επιλέξουμε τα δεδομένα που μας ενδιαφέρουν και τα οποία πρόκειται να αναλύσουμε.

Υπάρχουν δύο ειδών αποθήκες πληροφοριών, τόσο οι εξωτερικές όσο και οι εσωτερικές. Όσον αφορά τις εξωτερικές πηγές είναι πιθανόν να παρουσιαστούν προβλήματα νομικής φύσεως, κατά την μεταφορά, στην χρησιμοποίηση και στο συνδυασμό πληροφοριών από διάφορες πηγές, παρά σε τεχνικής φύσης προβλήματα. Οι εσωτερικές πηγές δεδομένων είναι: οι Αποθήκες δεδομένων(Data Warehouse), οι Σχεσιακές Βάσεις Δεδομένων (Relational Databases), οι Προηγμένες Βάσεις Δεδομένων (Advanced Databases) και οι Αποθήκες Πληροφοριών (Information Repositories).

Έτσι, οι πολλαπλές πηγές δεδομένων μπορούν να συνδυαστούν καθορίζοντας το σύνολο στο οποίο τελικά πρόκειται να εφαρμοστεί η όλη διαδικασία.

Εφόσον λοιπόν επιλέξαμε τα δεδομένα προς ανάλυση συνεχίζουμε με το επόμενο βήμα της διαδικασίας KDD που αφορά τον καθαρισμό και την προ-επεξεργασία των στοιχείων (Β.Βουτσινάς)

1.3.2.3 Προ-επεξεργασία Δεδομένων

Στο βήμα της προ-επεξεργασίας των δεδομένων περιλαμβάνονται βασικές διαδικασίες οι οποίες περιλαμβάνουν την αφαίρεση του θορύβου και των outliers από τα δεδομένα, και τη συλλογή των απαραίτητων πληροφοριών για τη διαμόρφωση ή τη μέτρηση του θορύβου.

Είναι πολύ πιθανό η πληροφορία που διαθέτουμε από τα δεδομένα που συλλέγουμε να μην είναι πάντα έγκυρη και σωστά διαμορφωμένη. Αυτό πιθανότατα να οφείλεται στο ότι τα δεδομένα μπορεί να είναι *θορυβώδη*, να περιλαμβάνουν δηλαδή λάθη και ψευδή στοιχεία, *ελλιπή* να έχουν ελλείψεις στα χαρακτηριστικά των τιμών των δεδομένων που μας ενδιαφέρουν ή να υπάρχουν συναθροισμένα δεδομένα, και *ασυνεπή* να περιέχουν ασάφειες σε κωδικούς και ονόματα.

Θα πρέπει λοιπόν να επιλυθούν προβλήματα που αφορούν την αναπαράσταση, την κωδικοποίηση αλλά και τη διαμόρφωσή τους και να προ-επεξεργαστούμε αυτά τα δεδομένα ώστε να μπορέσουμε στη συνέχεια να επεξεργαστούμε τα καλά, οργανωμένα πλέον, δεδομένα μας και να βγάλουμε ορθά συμπεράσματα.

Η προ-επεξεργασία τους αποτελεί το 60% της διαδικασίας της εξόρυξης γνώσης διότι η ύπαρξη μη-ποιοτικών δεδομένων θέτει σε κίνδυνο την όλη διαδικασία εφ' όσον τα συμπεράσματα ,άρα και οι αποφάσεις, που εξάγονται δεν θα είναι ποιοτικές και άρα μη-ορθές, κάτι που μας οδηγεί σε πρόβλημα.

Με τον όρο *ποιοτικά* εννοούμε ότι τα δεδομένα πρέπει να χαρακτηρίζονται από συνέπεια, πληρότητα, ακρίβεια, αξιοπιστία, επικαιρότητα, να είναι προσπελάσιμα και κατανοητά. Τόσο οι αποφάσεις που λαμβάνουν τα διευθυντικά στελέχη όσο και η τεχνολογία της ψηφιακής αποθήκης δεδομένων (data warehouse) θα πρέπει να στηρίζονται σε ποιοτικά δεδομένα (Βαζιργιάννης 2003).

1.3.2.4 Μέθοδοι της προ-επεξεργασίας δεδομένων

1. *Καθαρισμός των δεδομένων.*

Πολλές φορές κάποια από τα δεδομένα που θέλουμε να εξετάσουμε μπορεί να μην υπάρχουν ή συμβαίνει ορισμένες από τις τιμές που συλλέγονται να είναι ασύμβατες με άλλες οπότε και διαγράφονται αυτόματα, ή ακόμη να μην έχουν καταγραφεί δεδομένα λόγω κακής συνεννόησης μεταξύ των υπευθύνων μιας επιχείρησης.

Αυτό μπορεί να συμβαίνει γιατί τα δεδομένα που συλλέγουμε δεν είναι πάντα διαθέσιμα, όπως για παράδειγμα, αν διεξάγουμε μια έρευνα για τους πελάτες μιας επιχείρησης υπάρχει πιθανότητα στο αρχείο πελατών της να μην αναγράφεται το εισόδημα του κάθε πελάτη.

Έτσι, δημιουργούνται κενά κελιά μεταξύ των δεδομένων μας τα οποία εμείς είτε χρησιμοποιώντας κάποιους σταθερούς όρους σε όλα, π.χ. γράφουμε «unknown» σε καθένα από αυτά είτε απλά τα διαγράφουμε σε περίπτωση που ο αριθμός τους είναι πολύ μικρός και δεν υπάρχει σημαντική διαφοροποίηση μεταξύ τους.

Επίσης ενδέχεται σε κάποια από τα δεδομένα μας να εντοπίζεται *θόρυβος* δηλαδή το τυχαίο λάθος ή η απόκλιση από μία μεταβλητή. Οι λανθασμένες τιμές που μπορεί να πάρουν τα δεδομένα μας οφείλονται είτε σε ονομαστική ασυνέπεια, είτε σε κάποιο τεχνικό πρόβλημα είτε σε ελαττωματικά όργανα συλλογής δεδομένων ή τέλος σε διάφορα προβλήματα κατά την εισαγωγή και τη μετάδοση τους.

Έτσι μπορεί να παρατηρούμε κατά την απογραφή δεδομένων διπλές εγγραφές και ατελείς τιμές. Υπάρχουν οι εξής τρόποι να χειριστούμε το παραπάνω πρόβλημα:

i. ***Binning Method:*** Μέθοδος ταξινόμησης σε κουτιά

Αρχικά ταξινομούμε τα δεδομένα και στη συνέχεια τα διαμοιράζουμε σε ίσα κουτιά. Ύστερα εξομαλύνουμε, δηλαδή με τη χρήση συγκεκριμένων δεικτών εξετάζουμε την κλίση μιας τιμής κατά τη διάρκεια μιας χρονικής περιόδου, με το μέσο, το διάμεσο ή τα όρια του κουτιού (bin). Έχουμε δύο είδη διαμερισμού : το διαμερισμό ίσου πλάτους , το διαμερισμό ίσου βάθους.

- *Διαμερισμός ίσου πλάτους*

Μία άμεση μέθοδος διαχωρισμού της απόστασης σε N διαστήματα ίσου μεγέθους δημιουργώντας έτσι ένα ομοιόμορφο πλέγμα. Αν A και B οι χαμηλότερες και υψηλότερες τιμές του χαρακτηριστικού που εξετάζουμε τότε το πλάτος των διαστημάτων θα είναιW=(B-A)/N. Γενικότερα η μέθοδος αυτή χαρακτηρίζεται ως άμεση.

- *Διαμερισμός ίσου βάθους*

Η μέθοδος αυτή χωρίζει την απόσταση σε Ν διαστήματα καθένα από αυτά περιλαμβάνει τον ίδιο αριθμό δειγμάτων. Έτσι υπάρχει μία ομαλή κλιμάκωση των τιμών που πρόκειται να εξετάσουμε.

ii. *Ομαδοποίηση:*

Συγκέντρωση ενός συνόλου δεδομένων που έχουν όμοια χαρακτηριστικά

iii. *Παλινδρόμηση:*

Όταν αναφερόμαστε σε τιμές ψάχνουμε την πιο αντιπροσωπευτική συνάρτηση, $Y= \alpha+ \beta X$,για τα δεδομένα μας εξετάζοντας την γραμμικά σύμφωνα με τα κατάλληλα υποδείγματα. [περαιτέρω ανάλυση ξεφεύγει από τα όρια της εργασίας]

2. *Μείωση δεδομένων*

Το γεγονός ότι υπάρχουν μεγάλες βάσεις δεδομένων κάνει την ανάλυσή τους πολύπλοκη με αποτέλεσμα η εξόρυξή τους να απαιτήσει πολύ χρόνο. Έτσι λοιπόν μπορούμε να μειώσουμε τα δεδομένα διατηρώντας μειωμένες αναπαραστάσεις δεδομένων σε χωρητικότητα αλλά δημιουργούνται παρόμοια αποτελέσματα ανάλυσης. Οι μέθοδοι που χρησιμοποιούμε είναι:

- *Συνάθροιση Δεδομένων*

Συναθροίζουμε τα δεδομένα για μία ξεχωριστή οντότητα ενδιαφέροντος χρησιμοποιώντας τη λιγότερο δυνατή πληροφορία για την επίλυση του προβλήματος μας.

- *Μείωση διαστάσεων*

Μειώνουμε τα πρότυπα που βρίσκουμε ή επιλέγουμε ένα μικρό αριθμό δεδομένων για να εξάγουμε ισοδύναμα αποτελέσματα με αυτά που θα είχαμε αν είχαμε κρατήσει όλα τα δεδομένα για ανάλυση.

- *Συμπίεση δεδομένων*

Πρόκειται για μια τεχνική που εφαρμόζεται σε ένα διάστημα D ώστε να το μετασχηματίσει σε ένα αριθμητικά διαφορετικό διάνυσμα D’ ίδιου μήκους

- *Ανάλυση Κυρίων Συνιστωσών (Α.Κ.Σ)*

Με την μέθοδο αυτή βρίσκουμε τους κάθετους άξονες στο χώρο των δεδομένων οι οποίοι περιλαμβάνουν το σημαντικότερο μέρος του πληροφοριακού περιεχομένου. Η Α.Κ.Σ παρουσιάζει υψηλή χωρική και χρονική πολυπλοκότητα ειδικά σε πολυδιάστατους χώρους γι' αυτό και μας είναι χρήσιμη στην επίλυση OLAP συστημάτων. Η ανάλυση αυτή εφαρμόζεται μόνο για αριθμητικά δεδομένα και μόνο εάν ο αριθμός των διαστάσεων τους είναι μεγάλο (tsirakis@ceid.upatras.gr)

3. *Μετασχηματισμός δεδομένων*

Τα δεδομένα μας σε αυτό το σημείο μετασχηματίζονται και παγιώνονται σε μορφές οι οποίες είναι κατάλληλες ώστε αυτά να εξορυχτούν. Η τεχνική αυτή είναι αναγκαία για την ανάδειξη των ιδιαιτεροτήτων και των διαφορετικών γωνιών θέασης του συνόλου των δεδομένων. Συνήθως, μετασχηματίζουμε τις τιμές μίας ιδιότητας που έχει πολύ μεγάλες ή πολύ μικρές τιμές.

Χρησιμοποιούνται μέθοδοι όπως η εξομάλυνση και η κανονικοποίηση τιμών, η κατασκευή κύβου δεδομένων με χρήση OLAP και η δημιουργία νέων τιμών με χρήση παλαιότερων δεδομένων.

Με τον όρο *εξομάλυνση* εννοούμε την απομάκρυνση του θορύβου από τα δεδομένα.

Με την *κανονικοποίηση* καθορίζουμε τα δεδομένα μας κλιμακώνοντάς τα σε ένα μικρότερο εύρος τιμών. Αυτό μπορεί να επιτευχθεί με τους ακόλουθους τρόπους:

- Τυποποίηση z-score: Y= Y-Ỹ /σ
- Μετασχηματισμός min – max:Y=
- Y-Ymin / Ymax–Ymin(Y'max* Y'min) + Y'min
- Κβάντωση: Είναι ο μετασχηματισμός μίας συνεχούς τιμής σε διακριτή.
- Με χρήση δεκαδικής κλίμακας, όπου κάθε διάστημα αντιπροσωπεύεται από ισοδύναμο αριθμό σημείων

Ο μετασχηματισμός δεδομένων είναι αναγκαίος για την ανάδειξη των ιδιαιτεροτήτων και των διαφορετικών γωνιών θέασης του συνόλου των δεδομένων (http://gtziralis.googlepages.com, LogisticsCourse_LectureOnForecasting_Tziralis.pdf).

1.3.2.5 Μοντελοποίηση Εξόρυξης

Αφού μετασχηματίσαμε τα δεδομένα μας, ήρθε η ώρα να αποφασίσουμε τον στόχο της διαδικασίας KDD επιλέγοντας τα κατάλληλα μοντέλα και αλγόριθμους Εξόρυξης Δεδομένων,(Clustering, Classification, Association Rule) καθώς και τη μετατροπή των δεδομένων (εφόσον το μοντέλο το απαιτεί) που θα χρησιμοποιηθούν στην εξόρυξη των δεδομένων μας ώστε να προβούμε σε χρήσιμα συμπεράσματα και έγκυρες αναλύσεις.

Εξόρυξη δεδομένων

Εφαρμόζοντας κατάλληλες μεθόδους ψάχνουμε για τα πρότυπα που μας ενδιαφέρουν. Τα πρότυπα θα πρέπει να είναι συγκεκριμένης μορφής όπως κανόνες συσχέτισης, ταξινόμηση, δέντρα αποφάσεων, παλινδρόμηση, ομαδοποίηση κ.λ.π. Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα βήματα.

1.3.2.6 Αξιολόγηση των προτύπων

Τα εξαγόμενα πρότυπα αξιολογούνται με κάποια μέτρα, προκειμένου να προσδιοριστούν τα πρότυπα τα οποία αντιπροσωπεύουν τη γνώση, δηλαδή τα αληθινά ενδιαφέροντα πρότυπα. Γενικά, η αξιολόγηση του μοντέλου αναφέρεται και στην εγκυρότητα των προτύπων και στην αξιολόγηση της ακρίβειας, της χρησιμότητας και της δυνατότητας κατανόησης του μοντέλου.

1.3.2.7 Σταθεροποίηση και παρουσίαση της γνώσης

Στο βήμα αυτό η εξορυγμένη γνώση ενσωματώνεται στο σύστημα και χρησιμοποιούνται κάποιες τεχνικές αντιπροσώπευσης γνώσης για να παρουσιάσουν την εξορυγμένη γνώση στο

χρήστη. Επίσης ελέγχουμε για επίλυση τυχόν συγκρούσεων με προηγούμενη εξορυγμένη γνώση .

Τα αποτελέσματα που θα προκύψουν καθώς και η μεθοδολογία που χρησιμοποιήσαμε θα πρέπει να καταγράφονται λεπτομερειακά σε μια ξεχωριστή βάση γνώσεων για μελλοντική αναφορά σε αυτά (Βαζιργιάννης,2003).

1.4 ΕΦΑΡΜΟΓΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Ο λόγος που χρησιμοποιούμε την Εξόρυξη Δεδομένων είναι για να αναλύουμε βάσεις δεδομένων και να υποβοηθούμε στη λήψη αποφάσεων τους τομείς:

1.1 *Ανάλυση αγοράς και βελτίωση αγοραστικής εκστρατείας*

- Target marketing
- Customer relation Management
- Market basket analysis (supermarket)
- Cross selling
- Market segmentation

Εφαρμογή: π.χ. Η περίπτωση μιας αλυσίδας σουπερ-μαρκετ. Η παρατήρηση ότι οι πελάτες που αγοράζουν κατεψυγμένες πίτσες αγοράζουν και μπύρα έκανε τα καταστήματα να τοποθετήσουν αυτά τα είδη σχετικά κοντά, γνωρίζοντας ότι οι πελάτες θα κάνουν τη διαδρομή μεταξύ των ραφιών με τις πίτσες και αυτών με τις μπύρες. Τοποθετώντας ανάμεσά τους και πατατάκια αυξάνουν τις πωλήσεις και στα τρία είδη (Dunham,2004).

1.2 *Ανάλυση εταιρειών και διαχείριση ρίσκου*

- Προβλέψεις
- Διατήρηση πελατολογίου

- Βελτιωμένη χρηματοδότηση (π.χ. τράπεζες)
- Έλεγχος ποιότητας
- Ανάλυση ανταγωνιστικότητας

Εφαρμογή:

Σε μια τράπεζα η κατασκευή δένδρων αποφάσεων από ιστορικά στοιχεία τραπεζικών δανείων μέσα από τις βάσεις δεδομένων της τράπεζας, για την παραγωγή αλγορίθμων, ώστε να αποφασίζεται αν πρέπει ή όχι να δοθεί ένα δάνειο σε έναν υποψήφιο πελάτη.

Στον τομέα του e-επιχειρείν και στις αλυσίδες καταστημάτων όπως Πλαίσιο, Γερμανός, Praktiker, τα δεδομένα συγκεντρώνονται και εισάγονται σε αποθήκες δεδομένων προκειμένου να διευκολύνεται και το κατάστημα αλλά και ο πελάτης.

Η WalMart χρησιμοποίησε την τεχνική εξόρυξης δεδομένων μέσω υπολογιστή για να μετασχηματίσει τις σχέσεις προμηθευτών της. Η WalMart από τις συναλλαγές της έθεσε ένα συγκεκριμένο σημείο πώλησης κάποιων $ για τα 2.900 καταστήματα της σε 6 χώρες και διαβιβάζει συνεχώς αυτό το ογκώδες στοιχείο των 7,5 terrabyte στην Αποθήκη εμπορευμάτων στοιχείων Terradata. του υπολογιστή της.

Ο υπολογιστής επιτρέπει σε περισσότερους από 3.500 προμηθευτές, να έχουν πρόσβαση στα στοιχεία όσον αφορά τα προϊόντα τους και να εκτελούν τις αναλύσεις των στοιχείων αυτών.

Οι προμηθευτές χρησιμοποιούν αυτά τα στοιχεία για να προσδιορίσουν τα σχέδια αγοράς πελατών σε επίπεδο επίδειξης καταστημάτων. Χρησιμοποιούν αυτές τις πληροφορίες για να διαχειριστούν τον τοπικό κατάλογο καταστημάτων και να προσδιορίσουν τις νέες ευκαιρίες πώλησης. Το 1995, οι υπολογιστές WalMart επεξεργάστηκαν πάνω από 1 εκατομμύριο σύνθετες ερωτήσεις στοιχείων.

1.3 *Αναγνώριση απάτης*

- Τράπεζες: Μέσω της εξόρυξης μπορούμε να ελέγξουμε το ξέπλυμα βρώμικου χρήματος.

Πρόσφατα ανακαλύφθηκε ότι μια από τις μεθόδους που χρησιμοποιείται για το ξέπλυμα χρημάτων είναι η ανάμειξη των κεφαλαίων με τις νόμιμες εισπράξεις ενός εστιατορίου που οργανώνεται ως «μπροστινή επιχείρηση» ώστε να μη γίνεται αντιληπτή από τις αρχές αυτή παρανομία.. Υπάρχουν ειδικά ιδρύματα τα οποία συλλέγουν τους διάφορους τύπους πληροφοριών για τις συναλλαγές μετρητών από τις τράπεζες που είναι άνω των 10000$ συμπεριλαμβανομένων των ονομάτων, των διευθύνσεων, των αριθμών αναγνώρισης, των απολογισμών, και των σχετικών ποσών. Έτσι μπορεί εύκολα να ανιχνευθεί όποια παράνομη κίνηση των εγκληματιών αυτών.

§ Τηλεπικοινωνίες: Εδώ γίνεται λόγος για υποκλοπή πληροφοριών και προσωπικών δεδομένων.

Κάποιοι δηλαδή κλέβουν τις τηλεφωνικές γραμμές και κάνουν τηλεφωνήματα που έχουν κάποια επαναλαμβανόμενα σχέδια είτε προς μια κλειστή ομάδα ατόμων (κινητά) είτε κάποια συγκεκριμένη ώρα της ημέρας κλπ. Μέσω της εξόρυξης δεδομένων μπορούμε να εντοπίσουμε τα τηλεφωνήματα από τις βάσεις δεδομένων των πελατών και να βρούμε τα στοιχεία των υπόπτων όπως ακριβώς και με την μέθοδο εντοπισμού ξεπλύματος χρημάτων.

1.4 *Άλλες ιδιότητες εξόρυξης δεδομένων*

§ Εξέταση ιατρικών αρχείων:

Για παράδειγμα η αμερικάνικη κυβέρνηση κάποτε λόγω του ότι είχε παρατηρηθεί μια σχετικά υψηλή συχνότητα της φλυκταινώδους νόσου κοτόπουλου μεταξύ των νέων νεοσυλλέκτων μεταξύ των ηλικιών 17 και 19. Η φλυκταινώδης νόσος του κοτόπουλου στους ενηλίκους μπορεί να είναι ένα αρκετά σοβαρό θέμα. Εξέτασε τις βάσεις δεδομένων του στρατού ώστε να προσδιορισθεί ο αριθμός των προβληματικών υποομάδων μέσα στον στρατολογημένο πληθυσμό ώστε να καθοριστούν οι πολιτικές και υγειονομικές διαδικασίες που στοχεύουν στην ελαχιστοποίηση της απειλής αυτής.

§ Επιστημονικός τομέας:

Τα δεδομένα συλλέγονται και αποθηκεύονται σε τρομερές ταχύτητες χάρη στις επιστημονικές προσομοιώσεις που παράγουν terabytes δεδομένων. Η εξόρυξη δεδομένων

βοηθά τους επιστήμονες στην κατηγοριοποίηση και την τμηματοποίηση των δεδομένων αλλά και στη διατύπωση υποθέσεων πάνω σε αυτά.(Dunham,2004)

1.5 ΤΙ ΔΕΝ ΕΙΝΑΙ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη δεδομένων είναι μία διαδικασία που πραγματοποιείται μέσω των βάσεων δεδομένων με σκοπό την ανεύρεση και εξαγωγή χαμένων προτύπων. Δεν είναι η λειτουργική διαδικασία που εξυπηρετεί το σύστημα της γνώσης, όπως για παράδειγμα δεν είναι η εξαγωγή, το φιλτράρισμα και η ανάκτηση της όποιας πληροφορίας από μία μηχανή αναζήτησης όπως είναι το Google ή το Amazon ή κάποια αναζήτηση ενός αριθμού στον τηλεφωνικό κατάλογο. Δεν παρέχει υποστήριξη στη λήψη αποφάσεων, δεν είναι μέτρο πρόγνωσης και πρόβλεψης. Η Εξόρυξη Δεδομένων είναι η παραγωγή λειτουργικής γνώσης μέσω της ανάλυσης των δεδομένων από τις βάσεις δεδομένων και η οποία συνδυάζει μεθόδους από διάφορα επιστημονικά πεδία όπως είναι η Στατιστική η Τεχνητή Νοημοσύνη και οι Βάσεις Δεδομένων (Παλιούρας,1992).

1.6 ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ

Τα τελευταία έτη έχουν αναπτυχθεί διάφορες τεχνικές και μέθοδοι εξόρυξης δεδομένων. Εξ' αιτίας της ύπαρξης τεράστιων βάσεων δεδομένων χρησιμοποιούνται διαφορετικά κριτήρια κατηγοριοποίησης των μεθόδων και των συστημάτων της εξόρυξης ανάλογα πάντα με το είδος των βάσεων δεδομένων, το είδος της γνώσης που θέλουμε να εξάγουμε και τις τεχνικές που θα εφαρμόσουμε.

Γενικά ένα σύστημα εξόρυξης δεδομένων θα μπορούσε να κατηγοριοποιηθεί ανάλογα με τους διάφορους τύπους συστημάτων βάσεων δεδομένων, το επίπεδο ανακάλυψης γνώσης, αν δηλαδή αναφερόμαστε σε γενική ή πολυεπίπεδη γνώση, και το είδος των τεχνικών που θα χρησιμοποιήσουμε ώστε να εξορύξουμε τα δεδομένα. Οι βασικοί στόχοι της διαδικασίας εξόρυξης δεδομένων είναι η εφαρμογή τεχνικών περιγραφής και πρόβλεψης σε μεγάλα σύνολα δεδομένων.

Η πρόβλεψη στοχεύει στην μελέτη της συμπεριφοράς κάποιων μεταβλητών οι οποίες επηρεάζονται από άλλες μεταβλητές πάνω στις οποίες και βασίζονται.

Η περιγραφή αφορά την ανακάλυψη προτύπων τα οποία αναπαριστούν μία πολύπλοκη βάση δεδομένων με τρόπο κατανοητό και αξιόπιστο.

Έχει σημειωθεί πως η περιγραφή είναι πιο σημαντική από την πρόβλεψη ως προς τη διαδικασία εξόρυξης εφ' όσον μας ενδιαφέρει περισσότερο η εύρεση κατανοητών και χρήσιμων προτύπων με στόχο να εξάγουμε συμπεράσματα που θα μας βοηθήσουν στην λήψη αποφάσεων.

Γενικότερα υπάρχει μία πληθώρα μεθόδων εξόρυξης δεδομένων λόγω του ότι οι μέθοδοι αυτοί εφαρμόζονται σε πολλά διαφορετικά πεδία δεδομένων. Παρ' όλα αυτά όμως όλες εξυπηρετούν τον ίδιο σκοπό: την εξόρυξη δεδομένων από ένα σύνολο στόχων με σκοπό την ανακάλυψη, την περιγραφή και την αξιολόγηση χρήσιμων και κατανοητών προτύπων ώστε εμείς σαν ερευνητές να μπορέσουμε να λάβουμε αποφάσεις.

Λόγω του θέματος δεν θα αναφερθούμε εκτενώς σε όλες τις μεθόδους παρά μόνο σε ορισμένες από τις βασικές όπως: η ταξινόμηση, οι κανόνες συσχέτισης, η παλινδρόμηση και η ομαδοποίηση, η οποία καιν αφορά το βασικό κομμάτι της εργασίας μας. Οι μέθοδοι οι οποίοι πρόκειται να αναλύσουμε ικανοποιούν διαφορετικές απαιτήσεις εφαρμογών πάνω στην διαδικασία εξόρυξης των δεδομένων (Βαζιργιάννης, 2003)

1.6.1 **ΟΜΑΔΟΠΟΙΗΣΗ**

Ξεκινώντας με τη μέθοδο της ομαδοποίησης, αναφερόμαστε σε μια διεργασία στην διαδικασία εξόρυξης γνώσης ,βάση της οποίας αποκτούμε χρήσιμες πληροφορίες από τα υπό μελέτη δεδομένα μας. Η μέθοδος αυτή αποσκοπεί στην τμηματοποίηση μιάς ομάδας δεδομένων σε ομάδες οι οποίες θα εμπεριέχουν στοιχεία με περισσότερα κοινά χαρακτηριστικά μεταξύ τους απ'ότι είναι με τα στοιχεία των άλλων ομάδων.

Ακολούθως, στην ενότητα Β' θα παρουσιαστεί καλύτερα και εκτενέστερα η συμβολή της ομαδοποίησης στην εξόρυξη γνώσης μέσα από ένα σύνολο δεδομένων (Βαζιργιάννης,2003)

1.6.2 **ΤΑΞΙΝΟΜΗΣΗ**

Η μέθοδος της ταξινόμησης είναι μία κυρίαρχη διεργασία στη διαδικασία εξόρυξης γνώσης που καλύπτει πολλές διαφορετικές εφαρμογές. Συνδυάζει μεθόδους Στατιστικής και Μηχανικής Μάθησης και βασικός της στόχος είναι η δημιουργία ενός μοντέλου του οποίου τα στοιχεία αντιστοιχούν σε κατηγορίες οι οποίες έχουν ήδη προκαθοριστεί.

Τα δεδομένα είναι ένα σύνολο από καταχωρημένες εγγραφές. Κάθε εγγραφή χαρακτηρίζεται από μία συνάρτηση f(x,y) όπου *το x είναι ένα σύνολο ιδιοτήτων* και το *y* μία ειδική συγκεκριμένη ιδιότητα. η οποία ονομάζεται *ετικέτα κατηγορίας*.

Με λίγα λόγια η ταξινόμηση είναι η διαδικασία εκμάθησης μιας συνάρτησης-στόχου f που καθοδηγεί κάθε σύνολο ιδιοτήτων σε μία από τις προκαθορισμένες ετικέτες κατηγορίας.

Ένα μοντέλο ταξινόμησης χρησιμεύει στο *να προβλεφθεί* η ετικέτα κατηγορίας αγνώστων εγγραφών. Θα μπορούσαμε να παρομοιάσουμε τη διαδικασία ταξινόμησης σαν ένα μαύρο κουτί το οποίο αυτόματα αναθέτει μία ετικέτα κατηγορίας όταν παρουσιάζεται ένα σύνολο ιδιοτήτων αγνώστων εγγραφών.

Ακόμη η ταξινόμηση χρησιμοποιείται σαν ένα εργαλείο-βοήθημα στο διαχωρισμό μεταξύ αντικειμένων διαφορετικών κατηγοριών.

Τα βήματα που ακολουθούμε ώστε να ταξινομήσουμε τα δεδομένα είναι η εποπτευμένη μάθηση και η ταξινόμηση.

- Εποπτευμένη Μάθηση

Σε αυτό το στάδιο τα δεδομένα που έχουμε συλλέξει αναλύονται από έναν αλγόριθμο ταξινόμησης ώστε να κατασκευαστεί το επιθυμητό μοντέλο. Τα στοιχεία επιλέγονται τυχαία από ένα πληθυσμό δεδομένων. Το μοντέλο που ονομάζεται ταξινομητής αναλύεται βάση των κανόνων κατηγοριοποίησης, των δένδρων απόφασης ή διαφόρων μαθηματικών τύπων.

- Ταξινόμηση

Εδώ εκτιμάται, βάση των μεθόδων ταξινόμησης, η ακρίβεια του μοντέλου που έχουμε δημιουργήσει. Το μοντέλο ταξινομεί κάθε τυχαίο δείγμα που εμείς έχουμε δημιουργήσει και

συγκρίνει την κατηγορία, στην οποία ανήκουν τα δεδομένα μας, με την πρόβλεψη που έκανε το μοντέλο για αυτή την κατηγορία.

Με λίγα λόγια, η ακρίβεια του μοντέλου σε ένα καθορισμένο σύνολο δεδομένων προς δοκιμή ,είναι το ποσοστό των δειγμάτων που ταξινομήθηκαν σωστά από το μοντέλο που δημιουργήσαμε βάση της εποπτευμένης μάθησης. Εάν το μοντέλο αποδειχθεί ακριβές τότε χρησιμοποιείται για την ταξινόμηση μελλοντικών δειγμάτων των οποίων η ταξινόμηση είναι άγνωστη.

*Για παράδειγμα οι βιολόγοι θα μπορούσαν να έχουν ένα μοντέλο που περιλαμβάνει διάφορα ζώα και με τη μέθοδο της ταξινόμησης να τα διαχωρίσουν σε κατηγορίες όπως: θηλαστικά, ερπετά, πτηνά, ψάρια και αμφίβια

Όπως παρακάτω:

ΟΝΟΜΑ	**ΘΕΡΜΟΚΡΑΣΙΑ ΣΩΜΑΤΟΣ**	**ΥΔΡΟΒΙΑ ΧΑΡ/ΚΑ**	**ΑΕΡΙΑ ΧΑΡ/ΚΑ**	**ΕΤΙΚΕΤΑ ΚΑΤΗΓΟΡΙΑΣ**
Άνθρωπος	ζεστή	Όχι	όχι	θηλαστικό
Σκύλος	ζεστή	Όχι	όχι	θηλαστικό
Γάτα	ζεστή	Όχι	όχι	θηλαστικό
Σολωμός	κρύα	Ναι	όχι	ψάρι
Πύθωνας	κρύα	Όχι	όχι	ερπετό
Νυχτερίδα	ζεστή	Όχι	ναι	πτηνό
Χελώνα	κρύα	Και τα δυο	όχι	ερπετό
Σαλαμάνδρα	κρύα	Και τα δυο	όχι	αμφίβιο

Πιγκουίνος	ζεστή	Και τα δυο	όχι	πτηνό
Λεοπάρδαλη	ζεστή	Όχι	όχι	θηλαστικό
Περιστέρι	ζεστή	Όχι	ναι	πτηνό

Παράδειγμα 1. (Piatetsky-Shapiro, Smyth,Uthurusamy, 1994)

1.6.2.1 ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Οι δημοφιλέστερες μέθοδοι ταξινόμησης είναι η Bayesian ταξινόμηση, τα δένδρα απόφασης, τα νευρωνικά δίκτυα, η ταξινόμηση κοντινότερων γειτόνων και τα Support Vector Machines.

§ Bayesian ταξινόμηση

Στόχος είναι να ταξινομηθεί ένα δείγμα X σε μία από τις κατηγορίες C1, C2,…,Cn με τη χρήση ενός μοντέλου πιθανότητας _τύπου Bayes. Κάθε κατηγορία χαρακτηρίζεται από μία εκ των προτέρων πιθανότητα παρατήρησης της κλάσης Ci στην οποία κλάση ανήκει και το δείγμα που εξετάζουμε.

Ο απλούστερος ταξινομητής τύπου Bayes είναι ο Naïve Bayesian, ο οποίος υποθέτει πως η επίδραση ενός γνωρίσματος σε μία υποτιθέμενη κατηγορία είναι ανεξάρτητη από τις τιμές των άλλων γνωρισμάτων.

Ένας άλλος ταξινομητής είναι τα Bayesian Belief Networks, γραφικά μοντέλα που επιτρέπουν σχέσεις εξάρτησης μεταξύ των υποσυνόλων των γνωρισμάτων

§ Δένδρα Απόφασης

Ένα δένδρο απόφασης κατασκευάζεται βάση ενός συνόλου προ- κατηγοριοποιημένων δεδομένων.

Κάθε κόμβος του προσδιορίζει τον έλεγχο κάποιου ιδιαίτερου γνωρίσματος, της κατηγορίας που έχει οριστεί, και κάθε κλαδί του κόμβου παίρνει μια πιθανή τιμή για το συγκεκριμένο γνώρισμα.

Εκφράζουν κανόνες και ταξινομούν τον ανομοιογενή πληθυσμό σε μικρότερες ομοιογενείς ομάδες στη βάση μιας μεταβλητής-στόχου.

Π.χ. εάν ηλικία<25 και φύλο= άνδρας και χρήση πιστωτικής= όχι

Τότε αγοραστής= όχι

Ωστόσο υπάρχουν προβλήματα κατά την εφαρμογή τους διότι εφαρμόζονται μόνο σε ονοματικές μεταβλητές καθώς όταν τα δεδομένα είναι αριθμητικά τα δένδρα γίνονται πολύπλοκα και δύσκολα κατανοούνται. Επίσης περιορίζονται μόνο σε μία μεταβλητή-στόχο.

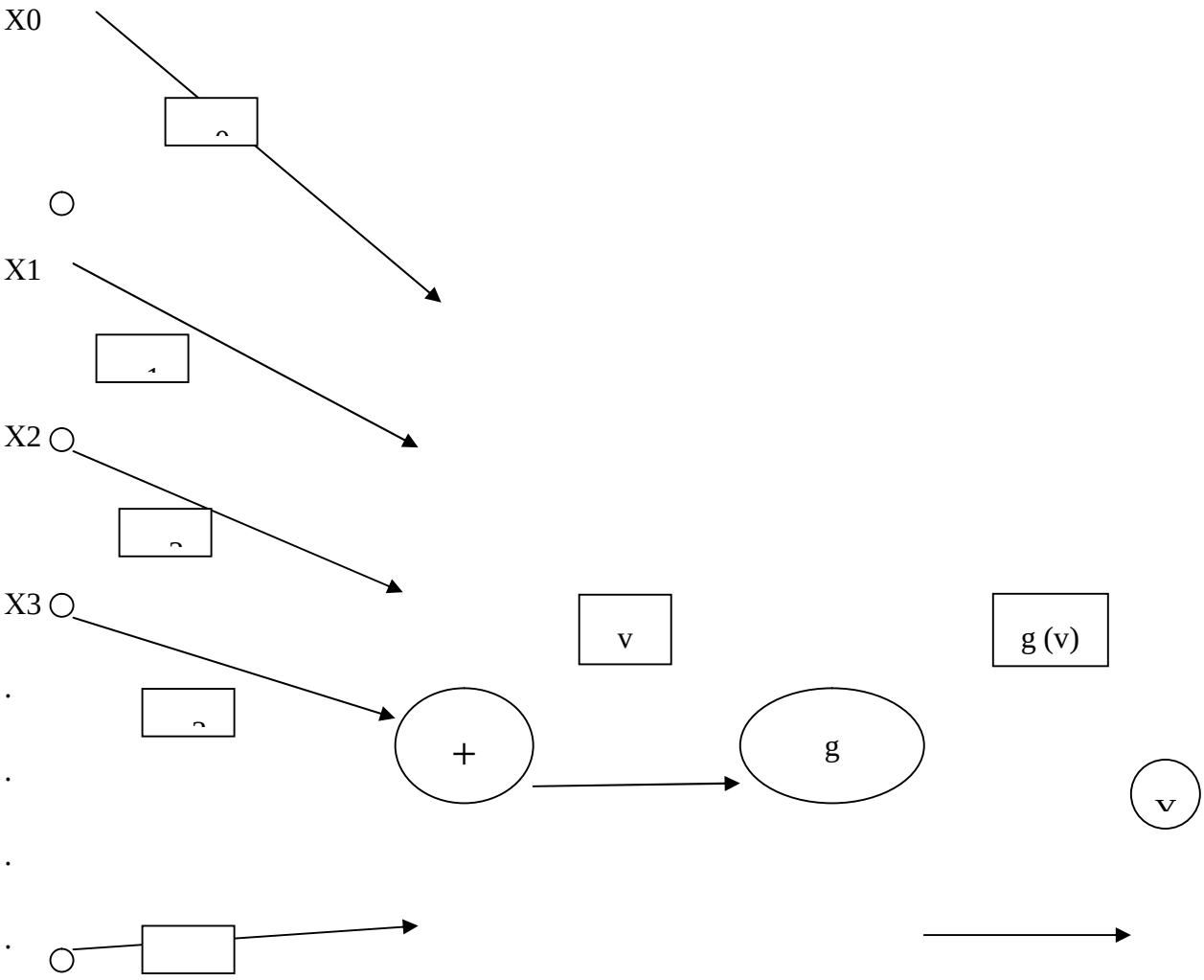
Παρ' όλα αυτά η κατανόηση τους και η γραφική τους απεικόνιση μας είναι ευκολονόητες. Βασίζονται σε διάφορους αλγόριθμους όπως: ο ID3, ο SLIQ, ο C4.5, ο SPRINT, ο Rain Forest κλπ ([18])

§ Νευρωνικά Δίκτυα

Πρόκειται για μία διαφορετική προσέγγιση της ταξινόμησης που αναγνωρίζοντας τα χαρακτηριστικά εισόδου και εξόδου κατασκευάζουμε ένα δίκτυο με την κατάλληλη τοπολογία ώστε να επιλεχθεί το σωστό σύνολο που θα εξετασθεί και στη συνέχεια να ελεγχθεί ώστε να παραχθεί ένα μοντέλο που θα προβλέπει τις τάξεις (έξοδοι) των μη-κατηγοριοποιημένων δειγμάτων (είσοδοι).

Τα νευρωνικά δίκτυα αποτελούνται από νευρώνες, με βάση τη νευρωνική δομή του ανθρώπινου εγκεφάλου. Καθώς επεξεργάζονται τα στοιχεία <<μαθαίνουν>> και έτσι τα λάθη από την αρχική ταξινόμηση της πρώτης εγγραφής ανατροφοδοτούνται στο δίκτυο και χρησιμοποιούνται για να τροποποιήσουν τον αλγόριθμο δικτύων την επόμενη φορά. Η διαδικασία συνεχίζεται επαναληπτικά:

○



Σχήμα-2.

Δομή νευρωνικού δικτύου:

Όπου χ: σύνολο εισερχόμενων τιμών χi

w: συσχετιζόμενα βάρη

g: συνάρτηση που αθροίζει τα βάρη (v)

y: έξοδος

§ Τεχνική των κοντινότερων γειτόνων

Σύμφωνα με αυτή τη μέθοδο κάθε νέο στοιχείο ταξινομείται βασιζόμενο σε προηγούμενες ταξινομήσεις στοιχείων παρόμοιων με αυτό. Έτσι παράγονται συνεχείς και επικαλυπτόμενες γειτνιάσεις. Το ποσοστό σφάλματος είναι ασύμπτωτο και δεν σχετίζεται με το μέτρο απόστασης που χρησιμοποιείται.

§ Support Vector Machines

Πρόκειται για τεχνική η οποία ελαχιστοποιεί τον εμπειρικό κίνδυνο και στοχεύει στην ελαχιστοποίηση του ανώτερου ορίου του σφάλματος γενίκευσης.

Κάθε παρατήρηση αποτελείται από ένα ζευγάρι της μορφής διανύσματος $\chi i \varepsilon R^n$ και από μία ετικέτα της συσχετιζόμενης κατηγορίας yj. Υποθέτουμε ακόμη ότι υπάρχει μία άγνωστη κατανομή πιθανότητας P(x, y) με βάση την οποία τα στοιχεία παράγονται.

Σκοπός είναι να βρούμε το σύνολο των παραμέτρων α της συνάρτησης f (x, α) έτσι ώστε η f να πραγματοποιεί την αντιστοιχία $xi \rightarrow yi$. Η ελαχιστοποίηση του ανώτερου σφάλματος του ορίου γενίκευσης επιτυγχάνεται με την εκμάθηση του α ώστε το όριο απόφασης που αντιπροσωπεύει η μηχανή εκπαίδευσης να έχει τη μέγιστη ή την ελάχιστη απόσταση από το πιο κοντινό σημείο που εξετάζεται (Βλαχαβάς, Κεφαλάς).

1.6.3 **ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ**

Οι κανόνες συσχέτισης είναι μία σχετικά νέα μέθοδος που ανακαλύφθηκε στις αρχές του 1990 και προήλθε από την ανάγκη των υπεραγορών (supermarkets) να καταχωρήσουν τις συναλλαγές κάθε πελάτη ηλεκτρονικά αναλύοντας το «καλάθι αγοράς» του (market basket analysis). Οι υπεραγορές αυτές με τη χρήση της μεθόδου αυτής συγκεντρώνουν ένα τεράστιο όγκο πληροφοριών σχετικά με τις αγορές των πελατών τους.

Έτσι, με τη χρήση των κανόνων συσχέτισης μπορούμε να εκφράσουμε το αποτέλεσμα ανάλυσης των χιλιάδων καλαθιών αγοράς των πελατών συσχετίζοντας τα αντικείμενα μεταξύ τους με σχέσεις εξάρτησης (if..then..). Οι κανόνες αυτοί εφαρμόζονται στην προώθηση

προϊόντων, στην τοποθέτηση προϊόντων στα ράφια των καταστημάτων και στη διαχείριση των αποθεμάτων.

Παράδειγμα:

Οι πελάτες που αγοράζουν μπύρες αγοράζουν και πατατάκια σε ποσοστό 65%.

Αυτό γράφεται σε συντομία «μπύρες πατατάκια, (65%)»

Η πρόταση αυτή συνδέει ένα αίτιο (μπύρες) με ένα αιτιατό (πατατάκια) καθώς επίσης παρουσιάζει μία ένδειξη για το πόσο πιθανό είναι αυτό να συμβαίνει ανάλογα με το ποσοστό που δίνεται.

Οι κανόνες συσχέτισης λοιπόν ανακαλύπτουν κρυμμένες συσχετίσεις μεταξύ των γνωρισμάτων ενός συνόλου δεδομένων, οι οποίες είναι της μορφής $A \rightarrow B$,όπου Α, Β είναι τα σύνολα στοιχείων που υπάρχουν στα προς ανάλυση δεδομένα.

Βέβαια ενδέχεται ο αριθμός των κανόνων που προκύπτουν να είναι πολύ μεγάλος λόγω των πολλών δεδομένων που έχουμε προς εξέταση. Γι' αυτό το λόγο εφαρμόζονται κάποια κριτήρια για τη σημαντικότητα των κανόνων αυτών.

Έτσι λοιπόν παράγεται ένα ελάχιστο σύνολο κανόνων συσχέτισης από το οποίο προκύπτουν οι υπόλοιποι κανόνες και το οποίο σύνολο ονομάζεται Αντιπροσωπευτικοί Κανόνες Συσχέτισης. Χρησιμοποιούμε δηλαδή ένα τελεστή κάλυψης ώστε από ένα αρχικό κανόνα να προκύψουν άλλοι βασισμένοι σ' αυτόν.

Υποθέτουμε δηλαδή ότι έχουμε ένα σύνολο συναλλαγών S={S1, S2,…,Sn}, όπου κάθε συναλλαγή S είναι ένα υποσύνολο ενός συνόλου δεδομένων A={A1, A2,…,Aκ} όπου Κ, οι ιδιότητες του Α.

Για ένα δεδομένο σύνολο Δ, υποσύνολο του Α καθορίζουμε την υποστήριξη *s* του Δ. Αν είναι μεγαλύτερη από ένα καθορισμένο από το χρήστη *κατώτατο όριο υποστήριξης T*, τότε καλούμε το Δ ως *συχνό σύνολο*.

Αυτό σημαίνει με απλά λόγια πως οι συναλλαγές στο σύνολο δεδομένων που περιέχουν τις ιδιότητες του Α τείνουν να περιέχουν και τις ιδιότητες του Δ.

Μόλις βρεθούν τα συχνά σύνολα το πρόβλημα του υπολογισμού των κανόνων συσχέτισης απλοποιείται και με τη χρήση αποδοτικών αλγορίθμων μπορούμε να βελτιώσουμε σύνολα δεδομένων που έχουν πολύπλοκα και μεγάλα συχνά σύνολα. (Βαζιργιάννης, Τσιράκης: tsirakis@ceid.upatras.gr)

1.6.4 **ΠΑΛΙΝΔΡΟΜΗΣΗ**

Η παλινδρόμηση είναι μια μέθοδος της εξόρυξης δεδομένων η οποία λειτουργεί σαν μία συνάρτηση πρόβλεψης μέσα στην οποία έχουν καταχωρηθεί τα προς εξέταση δεδομένα και η οποία προβλέπει ένα πραγματικό αριθμό.

Με τη χρήση της μεθόδου αυτής μπορούμε να προβλέψουμε τα κέρδη μιας επιχείρησης, τις πωλήσεις, τις τιμές των ακινήτων, τη θερμοκρασία, το τετραγωνικό μήκος σε πόδια, την απόσταση κλπ

Η διαδικασία εύρεσης ενός προτύπου βάση της παλινδρόμησης ξεκινά από ένα σύνολο δεδομένων μέσα στο οποίο οι μεταβλητές- στόχοι είναι ήδη γνωστές και συνιστούν τα ιστορικά δεδομένα εκ των οποίων τα μισά χρησιμοποιούνται για την πρόβλεψη και τα υπόλοιπα για τη δοκιμή του προτύπου.

Για παράδειγμα, ένα πρότυπο παλινδρόμησης θα μπορούσε να χρησιμοποιηθεί για να προβλέψει την αξία ενός σπιτιού βασισμένο στη θέση, τον αριθμό δωματίων, και άλλους παράγοντες. Το πρότυπο αυτό θα μπορούσε να αναπτυχθεί βασισμένο σε δεδομένα που έχουμε συλλέξει για πολλά παρόμοια σπίτια σε μία συγκεκριμένη χρονική περίοδο.

Επιπλέον στη συνάρτηση που θα βγάζαμε θα συμπεριλαμβάναμε και άλλες μεταβλητές εκτός από την αξία όπως το τετραγωνικό μήκος, τους φόρους, το parking, την πρόσβαση σε εμπορικά κέντρα και σχολεία. Η αξία δηλαδή θα ήταν ο στόχος και οι υπόλοιπες μεταβλητές θα χρησιμοποιούνταν σαν στοιχεία - μέτρα πρόβλεψης.

Κατά τη διάρκεια της δημιουργίας του προτύπου μέσω της παλινδρόμησης χρησιμοποιείται ένας αλγόριθμος παλινδρόμησης που εκτιμά την αξία της συνάρτησης-στόχου, η οποία επηρεάζεται από τις μεταβλητές κάθε φορά για κάθε περίπτωση.

Οι σχέσεις μεταξύ παραμέτρων και στόχου συμπεριλαμβάνονται στο πρότυπο που δημιουργήσαμε το οποίο μπορεί να εφαρμοστεί σε διαφορετικό σύνολο δεδομένων του οποίου οι μεταβλητές – στόχοι είναι άγνωστες. Τα πρότυπα παλινδρόμησης εξετάζονται από πολλές στατιστικές μεθόδους που μετράνε τη διαφορά μεταξύ προβλεπόμενων και αναμενόμενων τιμών (Βαζιργιάννης,2003)

ΕΝΟΤΗΤΑ Β΄

2.1 ΟΜΑΔΟΠΟΙΗΣΗ

Η ομαδοποίηση μπορεί να θωρηθεί ως μια από τις πλέον σημαντικότερες διεργασίες στη διαδικασία της μη-εποπτευμένης μάθησης, βάση της οποίας μπορούν να προσδιοριστούν κατανομές ή πρότυπα που παρουσιάζουν ενδιαφέρον στα υπό μελέτη δεδομένα μας.

Η διαδικασία της ομαδοποίησης αρχίζει με τον διαχωρισμό ενός συνόλου δεδομένων σε (k) ομάδες έτσι ώστε τα στοιχεία του συνόλου που ανήκουν σε κάποια ομάδα να είναι περισσότερο όμοια μεταξύ τους από ότι είναι με εκείνα των άλλων ομάδων.

Το μέτρο ομοιότητας μεταξύ των στοιχείων καθορίζεται από μία συνάρτηση απόστασης που τη συμβολίζουμε ως D (distance).

Αρχικά το πιο γνωστό μέτρο ομοιότητας που χρησιμοποιείται είναι η Ευκλείδια απόσταση, η οποία ορίζεται ως εξής:

$$d(i,j) = \sqrt{(|Xi1 - Xj1|^2 + |Xi2 - Xj2|^2 + \ldots + |Xip - Xjp|^2)}$$

Όπου ισχύουν τα παρακάτω:

$d(i,j) = 0$

$d(i,j) = d(j,i)$

$d(i,j) \leq d(i,k) + d(k,j)$

Βέβαια, ένα μειονέκτημα που παρουσιάζει η παραπάνω μέθοδος είναι πως όταν έχουμε να κάνουμε με πολλές διαστάσεις, το χαρακτηριστικό το οποίο παρουσιάζει μεγαλύτερη διαφοροποίηση κυριαρχεί από τα άλλα και έτσι κάνει τα αποτελέσματα που εξάγουμε μη-έγκυρα (Han and Kamber).

Για το λόγο αυτό χρησιμοποιούμε ένα άλλο μέτρο απόστασης που ονομάζεται Minkowski distance η οποίοα απόσταση ορίζεται ως εξής:

$$d(i,j) = \sqrt[n]{(|Xi1 - Xj1|^n + |Xi2 - Xj2|^n + \ldots + |Xip - Xjp|^n)}$$

όπου i=(xi1, xi2,…,xip) και j= (xj1, xj2,…,xjp) είναι δύο p-διάστατα στοιχεία και n ένας θετικός ακέραιος

Η διαδικασία ομαδοποίησης ενός συνόλου δεδομένων περιλαμβάνει τα εξής βήματα:

§ Επιλογή χαρακτηριστικών γνωρισμάτων:

Θα πρέπει να επιλεγούν κατάλληλα τα γνωρίσματα στα οποία πρόκειται να εφαρμοστεί η ομαδοποίηση ώστε να μπορεί να κωδικοποιηθεί όσο το δυνατόν περισσότερη πληροφορία.

§ Επιλογή αλγορίθμου ομαδοποίησης:

Υπάρχουν πολλοί αλγόριθμοι οι οποίοι χρησιμοποιούνται στην ομαδοποίηση των δεδομένων. Φυσικά αυτό εξαρτάται από τον τύπο τω δεδομένων, αν δηλαδή αναφερόμαστε σε κατηγορικά ή αριθμητικά δεδομένα.

Ανάλογα λοιπόν με τη φύση των στοιχείων προς εξέταση θα επιλέξουμε τον κατάλληλο αλγόριθμο ο οποίος με τη σειρά του καθορίζεται από *το μέτρο γειτνίασης*, το οποίο προσδιορίζει κατά πόσο είναι όμοια δύο στοιχεία και *το κριτήριο ομαδοποίησης* που θα χρησιμοποιηθούν.

**Μέτρο γειτνίασης*

Προσδιορίζει κατά πόσο είναι όμοια δύο στοιχεία

***Κριτήριο ομαδοποίησης*

Αναφερόμαστε εδώ σε κάποια συνάρτηση κόστους ή κάποιον άλλο τύπο κανόνων αναλόγως πάντα τον εκάστοτε αλγόριθμο που χρησιμοποιούμε κάθε φορά.

§ Έλεγχος και ερμηνεία αποτελεσμάτων:

Σε αυτό το σημείο θα αξιολογήσουμε κατά πόσον τα αποτελέσματα που μας έδωσε ο αλγόριθμος είναι έγκυρα. Η ακρίβεια του κάθε αλγορίθμου ομαδοποίησης εξακριβώνεται βάση καταλλήλων κριτηρίων και εφαρμογών.

Η ομαδοποίηση δεδομένων εφαρμόζεται σε διάφορα πεδία, κάποια από τα οποία είναι τα εξής:

§ Λήψη αποφάσεων

Εδώ μπορούμε να χρησιμοποιήσουμε την τεχνική ομαδοποίησης για την ανακάλυψη σημαντικών πελατειακών ομάδων οι οποίες ομαδοποιούνται σύμφωνα με τα καταναλωτικά πρότυπα. Χρησιμοποιώντας λοιπόν την τεχνική αυτή μπορούμε να εφαρμόσουμε στρατηγικές για την εξυπηρέτηση των ομάδων αυτών.

§ Spatial Data Mining

Με τον παραπάνω όρο αναφερόμαστε στην ομαδοποίηση σε χωρικά δεδομένα όπως γεωγραφικό μήκος και πλάτος, ταχυδρομικό κώδικα, διευθύνσεις. Οι αλγόριθμοι που χρησιμοποιούμε εδώ εντοπίζουν τα στοιχεία στο χώρο και τα εξετάζουν.

Η διαφορά των αλγορίθμων που χρησιμοποιούνται στα χωρικά δεδομένα από εκείνους που χρησιμοποιούνται σε κανονικά δεδομένα, είναι ότι οι δεύτεροι δουλεύουν χρησιμοποιώντας κέντρα βάρους και απλές μετρήσεις απόστασης και δεν είναι σε θέση να αναγνωρίσουν ασυνήθιστα σχήματα όπως αυτά που περιλαμβάνουν χωρικά δεδομένα.

§ Web Mining

Αναφερόμαστε σε ομαδοποίηση των δεδομένων που συλλέγουμε μέσω παγκόσμιου ιστού ώστε να βρούμε ομάδες χρηστών που προσπελαύνουν την κάθε φορά εξεταζόμενη ιστοσελίδα.

Παράδειγμα:

Έστω ότι έχουμε μία βάση δεδομένων σε ένα σουπερ-μάρκετ που περιλαμβάνει εγγραφές προϊόντων τα οποία καταναλώνονται. Μέσω της διαδικασίας της ομαδοποίησης θα μπορούσαμε να ομαδοποιήσουμε τους πελάτες σύμφωνα με τις αγοραστικές τους προτιμήσεις

με τέτοιο τρόπο ώστε οι πελάτες που παρουσιάζουν όμοια καταναλωτικά πρότυπα να ανήκουν στην ίδια ομάδα καταναλωτών.

Βασική μας προτεραιότητα είναι να οργανώσουμε τα πρότυπα σε λογικές ομάδες οι οποίες θα μας επιτρέψουν, με τη βοήθεια κάποιων αλγορίθμων, να ανακαλύψουμε ομοιότητες και διαφορές ώστε να προβούμε σε χρήσιμα συμπεράσματα.

Παρακάτω παραθέτουμε τους αλγόριθμους ομαδοποίησης οι οποίοι ταξινομούνται ανάλογα με:

- τον τύπο δεδομένων που εισάγονται στον εκάστοτε αλγόριθμο
- τη μέθοδο που καθορίζει την ομαδοποίηση του συνόλου των δεδομένων
- τη θεωρία και τις θεμελιώδεις έννοιες όπου βασίζονται οι τεχνικές ανάλυσης της ομάδας (Βαζιργιάννης 2003, Παπαδάκης 2002, Τσιράκης ιστοσελίδα)

2.2 ΕΙΔΗ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ:

Οι αλγόριθμοι ομαδοποίησης μπορούν να κατηγοριοποιηθούν ανάλογα με τον τύπο των δεδομένων που περιέχουν και το είδος της μεθόδου ομαδοποίησης που χρησιμοποιούν για τον καθορισμό των επιθυμητών ομάδων (Hsien Liao, 1995-2002).

2.2.1 **Διαιρετική Ομαδοποίηση**

Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη για την συμπίεση και την απεικόνιση μεγάλων βάσεων δεδομένων. Κυρίως βασίζεται στην άμεση αποσύνθεση του συνόλου των δεδομένων που πρόκειται να εξετάσουμε σε ένα σύνολο ομάδων οι οποίες δεν θα συσχετίζονται μεταξύ τους.

Οι αλγόριθμοι που ανήκουν σε αυτή την κατηγορία προσπαθούν να ελαχιστοποιήσουν τη συνάρτηση ανομοιότητας μεταξύ των δειγμάτων κάθε ομάδας και να μεγιστοποιήσουν τη συνάρτηση ανομοιότητας μεταξύ των διαφορετικών ομάδων που δημιουργούνται.

ΔΙΑΙΡΕΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ

§ *K-Means*

Ο αλγόριθμος αυτός ακολουθεί μία απλή επαναληπτική μέθοδο ώστε να χωρίσει ένα δεδομένο σύνολο δεδομένων σε ένα καθορισμένο αριθμό k ομάδων.

Η εφαρμογή του k-means ξεκινά τη λειτουργία της σε ένα σύνολο διανυσμάτων διάστασης d, όπου D= {xi, i=1,2,..,N}. Κάθε xi παριστάνει τον αριθμό των σημείων των δεδομένων προς εξέταση Κατόπιν ο αλγόριθμος επιλέγει k σημεία από το d .σύνολο διανυσμάτων τα οποία σημεία ορίζουμε ως αρχικούς αντιπροσώπους των k ομάδων, προκειμένου να ελαχιστοποιήσει την παρακάτω αντικειμενική συνάρτηση.

Η αντικειμενική συνάρτηση που προσπαθεί να ελαχιστοποιήσει ο αλγόριθμος είναι η μέση τετραγωνική απόσταση των σημείων από τα πλησιέστερα κέντρα k των ομάδων.:

$E = \Sigma \Sigma d(x, mi)$, $x \in c$.

Στην εξίσωση αυτή το mi είναι το κέντρο της ομάδας Ci, ενώ d (x, mi) είναι η απόσταση μεταξύ ενός στοιχείου x και του κέντρου mi.

Η διαδικασία που ακολουθεί ο k-means αποτελείται από δύο αρχικά βήματα:

Πρώτα ορίζει τα k κέντρα των C ομάδων όπου αναθέτει κάθε στοιχείο του συνόλου δεδομένων στην ομάδα της οποίας το κέντρο που έχει οριστεί είναι πιο κοντά στο στοιχείο αυτό. Με αυτόν τον τρόπο γίνεται ένας διαχωρισμός των στοιχείων. Αν η σειρά των δεδομένων δεν έχει κάποια ιδιαίτερη σημασία τότε παίρνουμε τις πρώτες k- εγγραφές προς εξέταση, αλλιώς επιλέγουμε τους k- αντιπροσώπους για τις ομάδες που εξετάσουμε.

Στη συνέχεια υπολογίζεται η απόσταση κάθε στοιχείου του συνόλου δεδομένων από το κέντρο της κάθε ομάδας, σύμφωνα με την παρακάτω απόσταση: $d(ki) = (x(k) – v(i))^2$,όπου k=1,2,..,n και i=1,2,..,c

Κάθε στοιχείο Xk αντιστοιχίζεται στην ομάδα για την οποία ισχύει ότι:

$\text{Min}_{k,i} (d_{ik})$, για κάθε i,k.

Έπειτα γίνεται ένας επαναπροσδιορισμός των μέσων, των κέντρων δηλαδή των ομάδων. Δηλαδή κάθε k- αντιπρόσωπος επαναπροσδιορίζεται ώστε να υπολογιστούν τα νέα κέντρα ομάδων με τη χρήση του μέσου όρου των σημείων που ήδη υπάρχουν σε αυτές..

Έτσι για άλλη μία φορά αντιστοιχίζεται κάθε σημείο στην ομάδα της οποίας το κέντρο είναι πιο κοντά στο σημείο αυτό.

Ο υπολογισμός των νέων κέντρων των σημείων προκύπτει από το μέσο:

$m_i\,\hat{}\,(r+1) = \Sigma\, x_k / n_i$,όπου n_i ο αριθμός των στοιχείων που ανήκουν στην i ομάδα μέχρι στιγμής.

Σύμφωνα με την παραπάνω σχέση η οποία βασίζεται στο μέσο, θα πρέπει να ισχύει: $\| m_i\,\hat{}\,(r) - m_i\,\hat{}\,(r+1) \| > \varepsilon$, όπου r = r+1.

Η διαδικασία ανάθεσης των στοιχείων στα πλησιέστερα κέντρα ομάδων επαναλαμβάνεται μέχρι τα όρια των ομάδων να σταματήσουν να μεταβάλλονται, δηλαδή μέχρι να τείνουν στο μηδέν οι αποκλίσεις μεταξύ των κέντρων των ομάδων τα οποία προέκυψαν από την τελευταία επανάληψη (Βαζιργιάννης 2003, [21])

ΠΑΡΑΛΛΑΓΕΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ K-Means

Στην προηγούμενη ενότητα μιλήσαμε για τον αλγόριθμο K-Means ο οποίος αποτελεί μία καθιερωμένη πλέον μέθοδο ομαδοποίησης με αποτελεσματικές τεχνικές για διάφορα πεδία ορισμού των , κάθε φορά , υπό ανάλυση δεδομένων.

Ωστόσο υπάρχουν διάφορες παραλλαγές του αλγορίθμου αυτού, καθεμία από τις οποίες διαφέρει ως προς τον τρόπο επιλογής των αρχικών κέντρων των ομάδων, και ως προς τη γενικότερη στρατηγική τους. Εμείς θα αναφερθούμε στις παρακάτω παραλλαγές του k-means:

§ *Αλγόριθμος ISODATA*

(Iterative Self - Organizing Data Analysis Technique Algorithm)

Ο αλγόριθμος αυτός αποτελεί επέκταση του k-means και ειδικεύεται στο να επιλέγει αυτόματα το πλήθος των κλάσεων των δειγμάτων δεδομένων που εμπεριέχονται σε ομάδες, βάση κάποιου κόστους εκτέλεσης.

Εδώ, ο χρήστης επιλέγει

το ελάχιστο πλήθος δειγμάτων ανά ομάδα

το επιθυμητό πλήθος ομάδων

τη μέγιστη διακύμανση για τον διαχωρισμό των ομάδων ($\sigma^2 s$)

τη μέγιστη απόσταση μεταξύ των ομάδων

το μέγιστο πλήθος ομάδων που μπορούν να ενωθούν

Κατά την αρχική εκτέλεση το αλγορίθμου, εφαρμόζεται ομαδοποίηση όμοια με αυτήν του k-means ώστε να δημιουργηθούν οι ομάδες προς εξέταση. Στη συνέχεια ο Isodata διασπά τις ομάδες που έχουν αρκετά ανόμοια δεδομένα ενώ ταυτόχρονα ενώνει εκείνες που έχουν αρκετά όμοια.

Έπειτα εκτελούνται από την αρχή τα παραπάνω βήματα μέχρι να ολοκληρωθεί η διαδικασία του αλγορίθμου για κάθε ομάδα.

Πλεονεκτήματα Isodata

Το βασικό πλεονέκτημα του αλγορίθμου αυτού είναι ότι διαθέτει δυνατότητες αυτό-οργάνωσης. Δηλαδή, μπορεί να καταργεί τις ομάδες που περιέχουν λίγα δείγματα, να διαιρεί αυτές που έχουν ανόμοια δείγματα και να ενώνει εκείνες με τα κοινά στοιχεία.

Μειονεκτήματα Isodata

Το σημαντικό μειονέκτημα αυτού του αλγορίθμου είναι ότι δεν μπορεί να χρησιμοποιηθεί για μεγάλα σετ δεδομένων καθώς είναι δύσκολο να καθοριστούν οι παράμετροι που πρέπει να λάβει υπ’ όψη του ο Isodata. Επίσης για να εφαρμοστεί ο αλγόριθμος με επιτυχία θα πρέπει τα δεδομένα να διαχωρίζονται γραμμικά.

Ωστόσο, κατά την εφαρμογή του σε διάφορες παραμέτρους χρησιμοποιείται ο συνδυασμός δεδομένων με το μικρότερο τετραγωνικό σφάλμα (http://www.icsd.aegean.gr/).

§ *Ασαφής ομαδοποίηση*

Οι αλγόριθμοι ασαφούς ομαδοποίησης επεκτείνουν τις αρχές των κλασικών τεχνικών ομαδοποίησης προτύπων, εκμεταλλευόμενοι τα ελκυστικά χαρακτηριστικά της θεωρίας των ασαφών συνόλων.

Κατά τα τελευταία χρόνια, η έρευνα έχει εστιαστεί στην ανάπτυξη αλγορίθμων επιβλεπόμενης ασαφούς ομαδοποίησης όπου χρησιμοποιούνται διάφοροι αλγόριθμοι για την βελτιστοποίηση των παραμέτρων.

Συγκεκριμένα, οι αλγόριθμοι χρησιμοποιούνται για την τοποθέτηση των ομάδων στον χώρο των χαρακτηριστικών, θεωρώντας διαφορετικές μεθοδολογίες ασαφούς ομαδοποίησης, όπως fuzzy c-means και fuzzy k-NN.

Αλγόριθμοι πολύ-παραγοντικής βελτιστοποίησης έχουν επίσης προταθεί με στόχο την διαδοχική βελτιστοποίηση πολλαπλών κριτηρίων αξιολόγησης, σε μια προσπάθεια να αντιμετωπισθεί το πρόβλημα του διαμερισμού του χώρου χαρακτηριστικών μεταξύ των ομάδων.

Στην περιοχή των ασαφών συστημάτων τα συστήματα βασισμένα σε ασαφείς κανόνες (fuzzy rule based systems, FRBS) παρέχουν μία πιο διαισθητική παράσταση γνώσης. Τα συστήματα FRBS έχουν τύχει εκτεταμένης εφαρμογής σε μια μεγάλη ποικιλία εφαρμογών, δεδομένου ότι προσδίδουν στους αναλυτές σημαντική ποιοτική πληροφορία του υπό εξέταση συστήματος (Dunham).

§ *Fuzzy c-means*

Ο αλγόριθμος αυτός είναι μία τεχνική ομαδοποίησης στην οποία κάθε σύνολο δεδομένων ομαδοποιείται σε ν ομάδες όπου κάθε στοιχείο του κάθε συνόλου δεδομένων αντιστοιχείται σε κάποια ομάδα βάσει κάποιου μέτρου εγγύτητας.

Για παράδειγμα, έστω ότι έχουμε ένα συγκεκριμένο στοιχείο που πρόκειται να εξετάσουμε. Αν το στοιχείο αυτό βρίσκεται κοντά σε ένα κέντρο μιας ομάδας τότε το στοιχείο αυτό θα έχει υψηλό βαθμό εγκυρότητας σε αυτήν την ομάδα.

Αντίστοιχα, εάν κάποιο στοιχείο υπό εξέταση βρίσκεται μακριά από το κέντρο μιας κάποιας ομάδας ,τότε θα παρουσιάζει χαμηλό βαθμό εγγύτητας για αυτήν την ομάδα.

Η δομή του αλγορίθμου έχει ως εξής:

Αρχικά ορίζονται κάποια υποθετικά κέντρα ομάδων έτσι ώστε να σημειωθεί η κεντρική τοποθεσία κάθε ομάδας, κάτι το οποίο δεν είναι απόλυτα σωστό.

Στη συνέχεια ο fuzzy c-means αντιστοιχεί για κάθε στοιχείο ένα μέτρο εγγύτητας ως προς την κάθε ομάδα. Ανανεώνοντας τα κέντρα των ομάδων και τον βαθμό εγγύτητας για κάθε σημείο, ο αλγόριθμος προσαρμόζει τα κέντρα των ομάδων στη σωστή τοποθεσία σε κάθε σύνολο δεδομένων.

Όλη αυτή η διαδικασία βασίζεται στην ελαχιστοποίηση μιας αντικειμενικής συνάρτησης, η οποία αναπαριστά την απόσταση από κάθε δεδομένο σημείο προς το κέντρο μιας ομάδας εξαρτώμενο πάντα από το βαθμό εγγύτητας του κάθε σημείου.

Η αντικειμενική συνάρτηση που πρόκειται να ελαχιστοποιήσει ο αλγόριθμος είναι η εξής:

$$J_m = \sum_{i=1}\sum_{j=1} u_{ij}^{m} \| x_i - c_j \|^2 \quad , 1 \leq m < \infty$$

Όπου m, είναι κάθε πραγματικός αριθμός μεγαλύτερος από τη μονάδα.

$U_{ij,}$ είναι το κατά πόσο αντιστοιχεί κάθε σημείο Xi στην εκάστοτε ομάδα j

$C_{j,}$ είναι το ν-διάστατο κέντρο της κάθε ομάδας

Η τετραγωνική ρίζα στην οποία είναι υψωμένη η απόλυτη τιμή των σημείων και των κέντρων των ομάδων συμβολίζει την ομοιότητα μεταξύ του κάθε σημείου και του κάθε κέντρου, τα οποία προφανώς και σχετίζονται (David Heckerman 1997, Osarech, Mirmehdi, Markham)

§ *Αλγόριθμος K-Prototypes*

Ο αλγόριθμος αυτός σχεδιάστηκε για να ομαδοποιεί μεγάλα σύνολα δεδομένων τα οποία περιέχουν αριθμητικές και λεκτικές τιμές.

Κατά την εφαρμογή του ορίζεται ένα μέτρο ανομοιότητας το οποίο υπολογίζει στοιχεία τόσο αριθμητικών όσο και λεκτικών τιμών.

Θεωρεί πως το μέτρο ανομοιότητας για αριθμητικές τιμές εκφράζεται με S_n ενώ εκείνο για τις λεκτικές ως S_c, το οποίο αφορά τις αταίριαστες κατηγορίες μεταξύ δύο αντικειμένων.

Το μέτρο ανομοιότητας μεταξύ των δυο αντικειμένων ορίζεται ως: $S_n + \gamma S_c$, όπου «γ» είναι ένα βάρος που θέτουμε για να εξισορροπήσουμε τα δυο ανόμοια μέρη.

Η διαδικασία που ακολουθεί ο K-Prototypes είναι ανάλογη με αυτή του αλγορίθμου k-means με τη μόνη διαφορά ότι όπως προαναφέραμε ο πρώτος χρησιμοποιεί τη μέθοδο για την ενημέρωση των λεκτικών τιμών των κέντρων των ομάδων.

Το μόνο πρόβλημα που υπάρχει κατά την εφαρμογή του αλγορίθμου αυτού είναι η επιλογή του κατάλληλου βάρους που θα χρησιμοποιήσουμε. Συνήθως χρησιμοποιείται η τυπική απόκλιση των αριθμητικών γνωρισμάτων (Βαζιργιάννης, 2003).

§ *Αλγόριθμος K-modes*

Ο αλγόριθμος αυτός αποτελεί μία πιο απλουστευμένη έκδοση του K-Prototypes, καθώς λαμβάνει υπ' όψη μόνο δεδομένα που αφορούν λεκτικές τιμές.

Δεν υπολογίζει δηλαδή το βάρος «γ», εφόσον δεν υπάρχει ο παράγοντας S_n.

Ο αλγόριθμος αυτός έχει σαν βάση τον k-means με κάποιες τροποποιήσεις. Δηλαδή, χρησιμοποιεί διαφορετικά μέτρα ανομοιότητας έτσι ώστε να μπορούν να εφαρμοστούν στις λεκτικές τιμές, αντικαθιστά τα k-κέντρα με τα k-modes και χρησιμοποιεί μεθόδους που βασίζονται στη συχνότητα εμφάνισης των τιμών προκειμένου να ενημερώνονται τα κέντρα των ομάδων, δηλαδή τα k-modes.

Βήματα k-modes:

Έστω ότι έχουμε τις ομάδες $\{S_1, S_2, S_3,\ldots..,S_k\}$ ενός συνόλου X, όπου $S_i \neq \text{ø}$ για $1 \leq l \leq k$ και $\{Q_1, Q_2, Q_3,\ldots. Q_k\}$ τα modes του $\{S_1, S_2, S_3,\ldots..,S_k\}$.

Επιλέγουμε k αρχικά modes, ένα για κάθε ομάδα.

Στη συνέχεια αναθέτουμε ένα στοιχείο στην ομάδα της οποίας το κέντρο (mode) είναι πιο κοντά στο κατ' επιλογήν μας στοιχείο με κριτήριο την απόσταση:

$D(x,y) = \Sigma_{j=1} \delta(x_j y_j)$ όπου $\delta (x_j\ y_j)$ παίρνει τιμές {0, 1} Όταν πάρει την τιμή 0 ισχύει: $x_j = y_j$,

ενώ όταν πάρει την τιμή 1 ισχύει: $x_j \neq y_j$.

Έπειτα ενημερώνεται το mode της ομάδας μετά από κάθε ανάθεση στοιχείου σύμφωνα με το παρακάτω θεώρημα:

Η συνάρτηση D(Q,X) ελαχιστοποιείται εάν και μόνο εάν:

$f_r (A_{J=} q_j| X) \geq f_r (A_{J=} c_{k,j}| X)$,για $q_{j=} c_{k,j}$ για όλα τα j=1,2,3,…m.

Το παραπάνω θεώρημα μας χρησιμεύει στο να ανακαλύψουμε ένα τρόπο ώστε να βρούμε ένα Q για ένα στοιχείο X.

Εφ' όσον λοιπόν έχουμε τοποθετήσει όλα τα αντικείμενα σε ομάδες επανεξετάζεται η ανομοιότητα των στοιχείων ως προς τα τρέχοντα modes. Σε περίπτωση που ένα στοιχείο βρίσκεται πιο κοντά στο mode μιας άλλης ομάδας από ότι στο mode της τρέχουσας ομάδας επανατοποθετείται στο στοιχείο της άλλης ομάδας και ενημερώνονται αναλόγως τα modes των ομάδων.

Η όλη διαδικασία επαναλαμβάνεται μέχρις ότου κανένα από τα στοιχεία να μην αλλάζει ομάδες μετά από τον πλήρη έλεγχο του συνόλου των δεδομένων.

§ *Αλγόριθμος **PAM***

Ο αλγόριθμος αυτός αποτελεί μία από τις πιο γνωστές μεθόδους ομαδοποίησης και ανήκει και αυτός στην κατηγορία των διαιρετικών αλγορίθμων.

Στόχος είναι να βρεθούν k ομάδες με την προϋπόθεση ότι θα καθοριστεί ένα στοιχείο αντιπρόσωπος για την κάθε ομάδα. Τα στοιχεία αυτά ονομάζονται medoids και είναι εκείνα που βρίσκονται πιο κοντά στα κέντρα των ομάδων που έχουν δημιουργηθεί.

Μόλις επιλεχθούν τα medoids κάθε στοιχείο το οποίο δεν έχει επιλεγεί ομαδοποιείται στην ομάδα εκείνη της οποίας το medoid μοιάζει περισσότερο με αυτό.

Να συμπληρώσουμε εδώ πως το αρχικό σύνολο των medoids επιλέγεται τυχαία.

Γενικά, ισχύει πως εάν Oj είναι ένα μη επιλεγμένο στοιχείο και Oi είναι ένα medoid τότε το Oj θα ανήκει στην ομάδα που αντιπροσωπεύεται από το Oi εάν,

d(Oj, Oi) = minOc d(Oj, Oc).

Όπου min Oc δηλώνει το ελάχιστο μεταξύ όλων των medoids Oc και το d(Oj, Oc) την απόσταση μεταξύ των στοιχείων.

Οι τιμές των αποστάσεων των στοιχείων δίνονται σαν είσοδο στην εκκίνηση του PAM. Η ποιότητα της ομαδοποίησης υπολογίζεται με βάση τη μέση διαφοροποίηση ανάμεσα σε ένα στοιχείο και το medoid της ομάδας στην οποία ανήκει.

Ο αλγόριθμος PAM χρησιμοποιείται για μικρά σύνολα δεδομένων διότι λόγω της μεγάλης πολυπλοκότητάς του είναι ανεπαρκής ώστε να χρησιμοποιηθεί για δεδομένα μεσαίου ή μεγάλου μεγέθους.

Ο αλγόριθμος λοιπόν ξεκινά βρίσκοντας τα k-medoids επιλέγοντας τυχαία στοιχεία του συνόλου δεδομένων. Έπειτα σε κάθε ανταλλαγή ανάμεσα σε ένα επιλεγμένο στοιχείο και σε ένα μη επιλεγμένο μέχρι να βελτιωθεί η ποιότητα της ομάδοποίησης. Για να υπολογιστούν τα αποτελέσματα αυτής της ανταλλαγής ο PAM υπολογίζει το κόστος Cjih για όλα τα μη επιλεγμένα Oj στοιχεία. Αυτό μπορεί να επιτευχθεί με τους εξής τρόπους:

Όταν το Oj ανήκει στην ομάδα που αντιπροσωπεύεται από το Oi και συγκεκριμένα βρίσκεται πιο κοντά στο Oj2 (το κοντινότερο medoid του Oj) από ότι στο Oh: d(Oj, Oh)≥ d(Oj, Oj,2) .

Αντικαθιστώντας λοιπόν το Oi με το Oh, το Oj θα ανήκει στην ομάδα που αντιπροσωπεύεται από το Oj2. Η ανταλλαγή αυτή πραγματοποιείται σύμφωνα με την εξίσωση:

Cijh= d(Oj, Oj,2) – d(Oj, Oi)

Από την παραπάνω εξίσωση προκύπτει πως το αποτέλεσμα βγαίνει πάντα θετικό.

Όταν το Oj ανήκει πάλι στην ομάδα που αντιπροσωπεύεται από το Oi ,με τη μόνη διαφορά πως εδώ το Oj είναι λιγότερο κοντά στο στοιχείο Oj2 σε σχέση με το Oh. Δηλαδή, d(Oj, Oh)≤ d(Oj, Oj,2).

Αντικαθιστώντας τώρα το Oi με το Oh σαν medoid, το Oj θα ανήκει στην ομάδα που αντιπροσωπεύεται από το Oh. Οπότε θα έχουμε την παρακάτω συνάρτηση κόστους:

Cijh= d(Oj, Oh) – d(Oj, Oi).

Το αποτέλεσμα της εξίσωσης μπορεί να πάρει και αρνητικές αλλά και θετικές τιμές, ανάλογα με το αν το Oj προσεγγίζει περισσότερο το Oi ή το Oh.

Αυτή τη φορά ας υποθέσουμε πως το Oj ανήκει σε μια διαφορετική ομάδα από αυτή που αντιπροσωπεύεται από το Oi. Οπότε ας θεωρήσουμε ότι το Oj2 είναι το medoid της ομάδας και ότι το Oj προσεγγίζει περισσότερο το Oi από ότι το Oh. Έτσι έχουμε: Cijh=0

Τέλος, στην περίπτωση όπου το Oj ανήκει στην ομάδα που αντιπροσωπεύεται από το Oj2 και το Oj είναι πιο κοντά στο Oh από ότι στο Oj2 τότε αντικαθιστούμε το Oi με το Oh θεωρώντας πως το Oh είναι το νέο medoid ,το Oj θα μετακινηθεί στην ομάδα που αντιπροσωπεύεται από το Oh. Οπότε θα έχουμε: Cijh= d(Oj, Oh) – d(Oj, Oj,2).

Το κόστος εδώ θα είναι πάντα αρνητικό.

Το συνολικό κόστος αντικατάστασης του στοιχείου Oi με το Oh προκύπτει από τον τύπο: TCih= Σj Cjih.

Βήματα PAM:

Επιλογή k αντιπροσώπων (medoids) για τις ομάδες.

Υπολογισμός συνολικού κόστους TCih για όλα τα ζεύγη αντικειμένων Oi, Oh όπου το πρώτο είναι το επιλεγμένο στοιχείο και το δεύτερο το μη επιλεγμένο.

Επιλογή ζεύγους Oi, Oh που αντιστοιχεί στο min Oi, Oh TCih. Σε περίπτωση που το κόστος βγει αρνητικό αντικαθιστούμε το Oi με το Oh και ξανακάνουμε το δεύτερο βήμα. Αν βγει θετικό τότε για κάθε μη επιλεγμένο στοιχείο βρίσκουμε το medoid που προσεγγίζει περισσότερο (Βαζιργιάννης http://mmlab.ceid.upatras.gr/).

§ *Αλγόριθμος CLARA*

Ο αλγόριθμος αυτός δημιουργήθηκε ώστε να εφαρμόζεται σε μεγάλα σύνολα δεδομένων.

Ο CLARA αντίθετα με τον PAM δεν βρίσκει στοιχεία αντιπροσώπους για ολόκληρο το σύνολο των δεδομένων αλλά βασίζεται στην τυχαία δειγματοληψία από το σύνολο των δεδομένων, και εφαρμόζοντας στο επιλεγμένο δείγμα τον αλγόριθμο PAM βρίσκει τα medoids του δείγματος.

Αν το δείγμα είναι εντελώς τυχαίο τότε αναπαριστά ολόκληρο το σύνολο των δεδομένων σε ένα πολύ ικανοποιητικό βαθμό, και έτσι τα medoids του δείγματος θα προσεγγίζουν τα medoids ολόκληρου του συνόλου.

§ *Αλγόριθμος CLARANS*

Ο αλγόριθμος αυτός συνδυάζει τους αλγορίθμους PAM και CLARA αναζητώντας κάθε φορά στοιχεία μόνο σε ένα υποσύνολο του συνόλου δεδομένων χωρίς να χρησιμοποιεί, όπως ο CLARA δειγματοληψία από το σύνολο.

Η διαφορά με τον CLARA είναι ότι ενώ ο δεύτερος παίρνει ένα καθορισμένο δείγμα κάθε φορά, ο CLARANS δημιουργεί ένα τυχαίο δείγμα σε κάθε αναζήτηση. Με λίγα λόγια, η ομαδοποίηση εδώ αναπαριστάται ως ένα γράφημα όπου κάθε κόμβος είναι μία πιθανή λύση , ένα σύνολο από k-medoids.

Μόλις ολοκληρωθεί η ομαδοποίηση μετά την αντικατάσταση ενός medoid, καλείται γείτονας (neighbor) της τρέχουσας ομαδοποίησης κ. ο. κ

Ωστόσο υπάρχει μία παράμετρος η οποία περιορίζει τον αριθμό των γειτόνων που ενδέχεται να υπάρχουν, η οποία καλείται maxneighbor.

Σε περίπτωση που βρεθεί κάποιος καλύτερος γείτονας ο αλγόριθμος μας μετακινείται στον γειτονικό αυτό κόμβο και ξαναρχίζει τη διαδικασία από τον κόμβο αυτόν. Διαφορετικά, από την τρέχουσα ομαδοποίηση παράγει ένα τοπικό βέλτιστο.

Τότε ο CLARANS ξεκινά πάλι αναζήτηση ενός νέου τοπικού βελτίστου σε ένα νέο, τυχαία επιλεγμένο, κόμβο.

Ο αριθμός των τοπικών βελτίστων που θα βρεθούν περιορίζεται από μία παράμετρο που λέγεται numlocal. Η υπολογιστική πολυπλοκότητα του αλγορίθμου για κάθε επανάληψη εξαρτάται από τον αριθμό των στοιχείων που βρίσκονται στο υποσύνολο που εξετάζεται κάθε φορά $O(n^2)$.

Ωστόσο, λόγω της τυχαίας επιλογής στοιχείων στην οποία βασίζεται ο CLARANS , δεν είναι κατάλληλος ώστε να εφαρμοστεί σε μεγάλες τιμές διότι η ποιότητα των αποτελεσμάτων δεν θα είναι εγγυημένα ορθή.

Βήματα CLARANS:

Ορισμός των παραμέτρων numlocal και maxneighbor που θα εξεταστούν θέτοντας ως ελάχιστο κόστος mincost ένα μεγάλο αριθμό για i = 1.

Καθορίζεται ο κόμβος προς εξέταση ο οποίος θα αναφέρεται σε έναν αρχικό κόμβο G n,k και ο οποίος θα ονομάζεται current.

Θέτουμε j = 1. Θεωρούμε ωστόσο έναν S τυχαίο γείτονα του τρέχοντος κόμβου και υπολογίζουμε το κόστος αντικατάστασης του τρέχοντος κόμβου από τον γειτονικό. Σε περίπτωση που S έχει μικρότερο κόστος θέτουμε ως current κόμβο τον S και επαναλαμβάνουμε το 3^o βήμα , διαφορετικά αυξάνουμε το j κατά μία μονάδα.

Εάν $j \leq$ maxneighbor επαναλαμβάνουμε το 4^o βήμα. Εάν $j \geq$ maxneighbor συγκρίνουμε το κόστος του τρέχοντος κόμβου (current) με το ελάχιστο κόστος (mincost). Εάν είναι μικρότερο από το mincost τότε θέτουμε ως mincost το κόστος του τρέχοντος κόμβου και ορίζουμε αυτόν ως καλύτερο κόμβο.

Τέλος, αυξάνουμε το i κατά μία μονάδα. Εάν i> numlocal εξάγουμε τον καλύτερο κόμβο και τερματίζεται η διαδικασία. Εάν όχι, τότε επαναλαμβάνουμε το $2^{ο}$ βήμα μέχρι να καταλήξουμε στην παραπάνω παράμετρο (Τασουλής,2007)

2.2.2 **Ιεραρχική ομαδοποίηση**

Σκοπός της μεθόδου αυτής είναι είτε να συγχωνευτούν οι μικρότερες ομάδες συνόλων δεδομένων σε μεγαλύτερες, ή να διαχωριστούν οι πολύ μεγάλες ομάδες σε μικρότερες.

Βέβαια το ζητούμενο εδώ είναι να διαπιστωθεί ποιες από αυτές ήταν μεγάλες και διασπάστηκαν και ποιες μικρές και άρα συγχωνεύτηκαν.

Το τελικό αποτέλεσμα του αλγορίθμου θα είναι ένα δένδρο από ομάδες, ένα δενδροδιάγραμμα το οποίο θα απεικονίζει τις σχέσεις μεταξύ των ομάδων. Εάν κόψουμε το δεδροδιάγραμμα σε κάποιο επίπεδο που επιθυμούμε μπορούμε να έχουμε ένα αποτέλεσμα ομαδοποίησης δεδομένων, των οποίων οι ομάδες δε σχετίζονται μεταξύ τους.

Υπάρχουν δυο είδη ιεραρχικών αλγορίθμων:

<u>Συσσωρευτικοί</u>

Οι αλγόριθμοι αυτοί έχουν αποδειχθεί αποτελεσματικοί σε πολλά πεδία όπως στην αναγνώριση οπτικών χαρακτήρων, στην ομαδοποίηση εγγράφων, στην εικόνα ιατρικής γνωμάτευσης.

Σε αυτούς τους αλγόριθμους κάθε σημείο δεδομένο είναι αρχικά ορισμένο στην ομάδα του. Αργότερα ο συσσωρευτικός αλγόριθμος βάση μιας συγκεκριμένης αντικειμενικής συνάρτησης συγχωνεύει διαδοχικά ζεύγη ομάδων μέχρις ότου όλα τα εναπομείναντα σημεία να ανήκουν στην ίδια ομάδα. Έτσι, παράγεται μία ακολουθία σχημάτων ομαδοποίησης και καθώς αυξάνεται αυτή η ακολουθία μειώνεται ο αριθμός των ομάδων.

Για να βρει ο αλγόριθμος που χρησιμοποιούμε την ομοιότητα ανάμεσα σε δύο ομάδες βασίζεται στην ελάχιστη ή μέγιστη ή μέση απόσταση μεταξύ των σημείων των ομάδων.

<u>Διαιρετικοί</u>

Σε αντίθεση με τους συσσωρευτικούς, οι διαιρετικοί αλγόριθμοι παράγουν μια ακολουθία σχημάτων ομαδοποίησης και καθώς η ακολουθία συνεχίζεται σε κάθε βήμα, αυξάνεται και ο αριθμός των ομάδων.

Η μέθοδος αυτή συνίσταται στο διαχωρισμό μιας ομάδας κατ' επανάληψη σε δύο υποομάδες ξεκινώντας από τη βασική ομάδα δεδομένων.

Χρησιμοποιώντας ένα διαιρετικό αλγόριθμο οι φορές που θα διαχωριστεί μια ομάδα σε υποομάδες μπορεί να είναι αμέτρητες. Το ζητούμενο σε αυτή τη μέθοδο είναι να βρεθεί ποια ομάδα θα επιλεγεί και με ποιο τρόπο θα διαχωριστεί σε δυο υποομάδες.

Οι διαιρετικοί αλγόριθμοι μπορεί να είναι πολύ αποτελεσματικοί στη δημιουργία μικρότερων μοντέλων τάξεων που θα είναι ομαδοποιημένα ιεραρχικά (Βαζιργιάννης).

Ιεραρχικοί Αλγόριθμοι:

§ *BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)*

Ο αλγόριθμος αυτός ομαδοποιεί βάση της χρήσης ιεραρχιών. Εφαρμόζεται μόνο για αριθμητικά δεδομένα και η δομή του είναι ιεραρχική και αυξητική. Ουσιαστικά πρόκειται για ένα Clustering Feature tree, CF-tree, που χρησιμεύει στην τμηματοποίηση των στοιχείων του συνόλου των δεδομένων με έναν αυξητικό και ιεραρχικό τρόπο.

Ο αλγόριθμος ξεκινά με το να εντοπίσει ένα στοιχείο μέσα σε μία ομάδα.

Μετά ομαδοποιεί τα κοντινότερα στοιχεία σε ομάδες ξεχωριστές μεταξύ τους και συνεχίζει την ίδια διαδικασία μέχρις ότου απομείνει μία μόνο ομάδα (η οποία στο εσωτερικό της σίγουρα θα περιέχει μικρότερες υποομάδες)

Ο Birch χρησιμοποιεί μία κύρια μνήμη δομής δεδομένων περιορισμένου μεγέθους που λέγεται Clustering Feature tree, CF-tree. Κάθε κόμβος περιέχει τις πραγματικές ομάδες και το μέγεθος καθεμίας από αυτές δεν θα πρέπει να είναι μεγαλύτερο από R.

Ο αλγόριθμος ανιχνεύει το σύνολο των στοιχείων και ορίζει τις εισερχόμενες αποστάσεις των δεδομένων μία προς μία..

Η εισαγωγή ενός στοιχείου στο δένδρο πραγματοποιείται διασχίζοντας το δένδρο από πάνω προς τα κάτω (top down) από τη ρίζα σύμφωνα με μία συνάρτηση που ομαδοποιεί την απόσταση των σημείων του δένδρου. Τέλος, κάθε στοιχείο εισάγεται στην πλησιέστερη υποομάδα κάτω από ένα κόμβο του CF-tree μας.

Σε περίπτωση που η εισαγωγή ενός στοιχείου σε μία υποομάδα προκαλέσει τη διάμετρο της υποομάδας να υπερβεί το κατώτατο όριο, μία νέα υποομάδα δημιουργείται. Η καινούρια υποομάδα που δημιουργείται, όπου ουσιαστικά πρόκειται για ένα νέο κόμβο στο δένδρο μας, μπορεί να περιέχει περισσότερα «παιδιά» από τον παράγοντα διακλάδωσης και το ανώτατο-κατώτατο όριο της διαμέτρου.

Για να διαχωρίσουμε ένα κόμβο πρέπει πρώτα να προσδιορίσουμε το ζευγάρι των υποομάδων κάτω από τον κόμβο των οποίων η εσωτερική απόσταση είναι η μεγαλύτερη. Αφού ο κόμβος διαχωριστεί σε δυο άλλους κόμβους οι εναπομείναντες υποομάδες αποστέλλονται σ' αυτούς σύμφωνα με το πόσο ταιριαστές είναι στο ζευγάρι υποομάδων που προσδιορίσαμε προηγουμένως.

Όταν ο διαχωρισμός των κόμβων τερματιστεί σε ένα κόμβο, η συγχώνευση των γειτονικότερου ζευγαριού κόμβων γίνεται για όσο αυτοί οι δύο κόμβοι δεν επηρεάζονται από την τελευταία διάσπαση.

Η διαδικασία της συγχώνευσης μπορεί να οδηγήσει σε έναν απευθείας διαχωρισμό, εάν ο κόμβος ο οποίος συγχωνεύεται περιέχει πολλά «παιδιά-κόμβους»(Han and Kamber).

CURE

Ο αλγόριθμος αυτός προσδιορίζει τις ομάδες που έχουν μη-καθορισμένο σχήμα και μεγάλες διαφορές στο μέγεθος.

Κάθε ομάδα αντιπροσωπεύεται από ένα συγκεκριμένο αριθμό k σημείων, τους αντιπροσώπους, οι οποίοι προκύπτουν από την επιλογή των πιο διάσπαρτων στοιχείων της ομάδας τα οποία ο αλγόριθμος «σπρώχνει» προς το κέντρο της ομάδας σε μια καθορισμένη κατεύθυνση κατά ένα ποσοστό α.

Η απόσταση μεταξύ των ομάδων είναι η απόσταση μεταξύ των πιο κοντινών αντιπροσώπων των δυο ομάδων. Δηλαδή μόνο τα στοιχεία- αντιπρόσωποι χρησιμοποιούνται για να υπολογιστεί η απόσταση μίας ομάδας από μια άλλη.

Η ύπαρξη k αντιπροσώπων στην ομάδα βοηθά τον αλγόριθμο να προσαρμόσει καλά τη γεωμετρική ανομοιομορφία των σημείων της ομάδας. Επιπλέον, η μετακίνηση των διάσπαρτων σημείων προς το κέντρο της ομάδας απομακρύνει τον υπάρχων θόρυβο και μετριάζει τις επιδράσεις των outliers (*αντικείμενα δεδομένων τα οποία δεν ακολουθούν τη γενική συμπεριφορά των δεδομένων και που συχνά τα θεωρούμε θόρυβο*) .

Αυτό συμβαίνει γιατί τυπικά τα outliers βρίσκονται μακριά από το κέντρο της ομάδας και έτσι η συρρίκνωση θα τα κάνει να κινηθούν περισσότερο προς το κέντρο ενώ οι αντιπρόσωποι που απομένουν υπόκεινται σε ελάχιστη μετακίνηση. Οι μεγάλες μετακινήσεις στα outliers μειώνουν την πιθανότητα λάθους συγχώνευσης των ομάδων και η παράμετρος α χρησιμοποιείται για τον έλεγχο του σχήματος των ομάδων.

Όταν το α παίρνει μικρές τιμές ,αυτό σημαίνει πως τα διάσπαρτα σημεία συρρικνώνονται πολύ λίγο και οι μη-σφαιρικές ομάδες ενισχύονται. Αντίθετα, οι μεγάλες τιμές για το α έχουν σαν αποτέλεσμα να δημιουργούνται συμπαγείς ομάδες καθώς τα διάσπαρτα σημεία μετακινούνται πιο κοντά στο μέσο της ομάδας.

Για να χειριστεί ο CURE μεγάλες βάσεις δεδομένων συνδυάζει τυχαία δειγματοληψία και τμηματοποίηση των στοιχείων μαζί. Δηλαδή, ένα τυχαίο δείγμα συλλέγεται από τα δεδομένα έτσι ώστε να είναι άκρως αντιπροσωπευτικό του συνόλου ώστε κατά την εφαρμογή του αλγορίθμου να μην παραληφθούν συγκεκριμένες ομάδες ή να μην προσδιοριστούν ομάδες οι οποίες δεν ανταποκρίνονται στις πραγματικές.

Καθώς όμως ο διαχωρισμός των ομάδων μειώνεται και οι ομάδες γίνονται λιγότερο πυκνές απαιτούνται δείγματα μεγάλου μεγέθους έτσι ώστε να είναι επιτυχής η ομαδοποίηση που θα επιτευχθεί.

Καθώς το μέγεθος του συνόλου των δεδομένων αυξάνεται η πολυπλοκότητα για τον αλγόριθμο ομαδοποίησης αυξάνεται σημαντικά. Για το λόγο αυτόν το δείγμα του συνόλου

μας διαιρείται σε ομάδες. Έπειτα βάση των ομάδων που έχουν προσδιοριστεί εφαρμόζεται ο αλγόριθμος για την εύρεση των ομάδων του συνόλου των δεδομένων.

Ωστόσο μπορούμε να διακόψουμε την συγχώνευση των ομάδων σε ένα τμήμα ένα η απόσταση μεταξύ των πλησιέστερων ομάδων που πρόκειται να συγχωνευτούν στο επόμενο βήμα ξεπερνά ένα συγκεκριμένο όριο (Βαζιργιάννης,).

§ *ROCK (RObust Clustering using linKs)*

Ο αλγόριθμος αυτός χρησιμοποιεί την έννοια των συνδέσεων (links) μεταξύ των ζευγών των σημείων προς εξέταση ενός τυχαία επιλεγμένου δείγματος.. Αυτό επιτυγχάνεται βάση μιας συνάρτησης ομοιότητας μεταξύ δύο γειτονικών σημείων.

Για κάθε ομάδα i διατηρούνται ταξινομημένα οι ομάδες j των οποίων ο αριθμός των συνδέσεων με την ομάδα i είναι διάφορος του μηδενός. Η τιμή τα ομοιότητας για τα ζεύγη των σημείων μπορεί να μετρηθεί με τη χρήση τόσο μετρικών όσο και μη-μετρικών συναρτήσεων.

Το πλήθος των συνδέσεων μεταξύ ενός ζεύγους σημείων ορίζεται από το πλήθος των κοινών γειτόνων των συγκεκριμένων σημείων. Η συνάρτηση που εξετάζουμε είναι η παρακάτω:

$$g(\mathrm{Ci}, \mathrm{Cj}) = \frac{link[Ci, Cj]}{(\mathrm{ni} + \mathrm{nj})^{1+2f(\theta)} - \mathrm{ni}^{1+2f(\theta)} - nj^{1+2f(\theta)}}$$

Όπου ο αριθμητής υποδηλώνει τον αριθμό των συνδέσεων που υπάρχουν σε μία ομάδα. Όσο μεγαλύτερος είναι ο αριθμός τόσο μεγαλύτερη συνοχή έχει η ομάδα.

Ενώ ο παρανομαστής αφορά τον αναμενόμενο αριθμό συνδέσεων σε μία ομάδα και εξομαλύνει τον αριθμητή.

Όσο μεγαλύτερη είναι η ποσότητα g(Ci,Cj) τόσο πιο αποτελεσματική είναι η συγχώνευση των ομάδων.

Έτσι ο αλγόριθμος προχωράει συσσωρευτικά στη συνένωση των ομάδων ξεκινώντας από τα γειτονικά σημεία με βάση πάντα το μέτρο ποιότητας της συγχώνευσης (http://www.ted.unipi.gr/)

2.2.3 ***Ομαδοποίηση βασισμένη σε γράφους:***

Η μέθοδος αυτή έγκειται στο γεγονός πως ένας αλγόριθμος ομαδοποίησης βασισμένης σε γράφους όταν εφαρμοστεί σε ένα γράφο κρατάει μόνο τις γειτονικές συνδέσεις ενός στοιχείου με τα κοντινότερά του και καθορίζει ένα μέτρο ομοιότητας για τα στοιχεία το οποίο βασίζεται στον αριθμό των κοντινότερων γειτόνων τους. Επίσης καθορίζει τα στοιχεία-πυρήνες που μας ενδιαφέρουν περισσότερο και κατασκευάζει ομάδες γύρω από αυτά και τέλος χρησιμοποιεί τις πληροφορίες που συνθέτουν το γράφο ώστε να αξιολογήσει αν δύο ομάδες πρέπει να συγχωνευθούν (Βαζιργιάννης, 2003).

Αλγόριθμοι ιεραρχικοί και βασισμένοι σε γράφους

Σε αυτή την κατηγορία ανήκουν αλγόριθμοι οι οποίοι συνδυάζουν γνωρίσματα τόσο της ιεραρχικής όσο και της βασισμένης σε γράφους ομαδοποίησης.

§ *CHAMELEON*

Ένας αλγόριθμος της παραπάνω κατηγορίας είναι ο chameleon ο οποίος μετρά την ομοιότητα δύο ομάδων. Ουσιαστικά επειδή τις ομάδες αυτές τις ανακαλύπτει σε δύο φάσεις χρησιμοποιώντας διαφορετικούς αλγορίθμους σε κάθε μια, θα λέγαμε ότι ο chameleon αποτελείται από δύο αλγορίθμους οι οποίοι συνδυάζονται.

Στην πρώτη φάση ο Chameleon χρησιμοποιεί έναν αλγόριθμο ομαδοποίησης βασισμένο σε γράφους ώστε να τμηματοποιήσει τα δεδομένα σε πολλές μικρές υποομάδες.

Για την αναπαράσταση των δεδομένων χρησιμοποιείται ο πλησιέστερος γράφος γειτνίασης. Δηλαδή, τα αντικείμενα του συνόλου δεδομένων είναι οι κορυφές του πλησιέστερου γράφου. Επίσης κάθε κόμβος συνδέεται με έναν άλλο με ακμές οι οποίες καθορίζουν πόσο κοντά και σε ποιους γειτονικούς γράφους των κόμβων βρίσκονται τα αντικείμενά μας.

Έπειτα ο Chameleon βάσει του αλγορίθμου τμηματοποίησης βρίσκει τις αρχικές υποομάδες ώστε να κατανεμηθεί ο k-πλησιέστερος γράφος γειτνίασης του συνόλου δεδομένων σε ένα μεγάλο αριθμό τμημάτων.

Στη δεύτερη φάση ο Chameleon με τη βοήθεια ενός ιεραρχικού συσσωρευτικού αλγόριθμου ο οποίος συνδυάζει τις υποομάδες που δημιουργούνται από τις συνεχείς επαναλήψεις

συνδυασμών μεταξύ τους που προέκυψαν από την πρώτη φάση. Η συγχώνευση μεταξύ των υποομάδων καθορίζεται με το να ελέγχουμε την ομοιότητά τους ως προς τη σύνδεση τους και το κατά πόσο έγκυρες είναι αυτές οι υποομάδες.

Αν τα ζευγάρια των υποομάδων των οποίων η σχετική μεταξύ τους σύνδεση και η εγκυρότητά τους είναι πάνω από το όριο που οι χρήστες του αλγορίθμου έχουν ορίσει, τότε αυτές συγχωνεύονται αυτόματα (http://mmlab.ceid.upatras.gr , Βαζιργιάννης).

§ *C²P*

Ο αλγόριθμος αυτός χρησιμοποιείται στις μεγάλες βάσεις δεδομένων και εκμεταλλεύεται τις δομές ευρετηρίων και την επεξεργασία ερωτήσεων του κοντινότερου ζευγαριού (Closest Pair Queries, CPQ).

Οι κύριοι στόχοι του είναι η αποδοτικότητα, η συγκέντρωση δηλαδή της ποιότητας των ομάδων, και η αποτελεσματικότητά τους. Ο C²P οργανώνει το αποτέλεσμα του CPQ πάνω σε ένα R-tree, μία δομή ενός γράφου, και αποτελείται και αυτός από δυο φάσεις.

Στην πρώτη φάση, ο αλγόριθμος παράγει διάφορες υποομάδες οι οποίες αντιπροσωπεύουν τις τελικές ομάδες. Πρόκειται για μία διαδικασία επαναληπτική στη διάρκεια της οποίας συγχωνεύονται διάφορες ομάδες.

Με τη διαδικασία CPQ ο αλγόριθμος βρίσκει τα ζευγάρια των σημείων που ανήκουν σε ένα σύνολο δεδομένων S . Με τη βοήθεια μιας γραφικής παράστασης οργανώνονται οι πληροφορίες εγγύτητας που έχουμε συλλέξει μέσω της διαδικασίας CPQ και έτσι ορίζονται οι ομάδες ως συστατικά του γράφου για την αναπαράσταση των οποίων χρησιμοποιούνται τα κέντρα τους.

Έπειτα ο C²P εφαρμόζει τον αλγόριθμο depth-First Search στο γράφο ο οποίος περιλαμβάνει τις υποομάδες του S για να βρει τα συνδεδεμένα στοιχεία του γράφου. Έτσι για παράδειγμα, σημεία που ανήκουν στο ίδιο συνδεδεμένο στοιχείο μπορούν να θεωρηθούν σαν μία υποομάδα.

Μόλις ο αριθμός των καθορισμένων υποομάδων γίνει ίσος με τον απαιτούμενο αριθμό υποομάδων τότε η φάση ολοκληρώνεται.

Η δεύτερη φάση του αλγορίθμου αφορά ουσιαστικά μία εξειδικευμένη πρώτη φάση. Κατά τη διάρκειά της ο C²P χρησιμοποιεί μία διαφορετική αναπαράσταση ομάδας ώστε να παραχθεί το τελικό σχήμα ομαδοποίησης συγχωνεύοντας δυο ομάδες σε κάθε βήμα του ώστε να μπορέσει να γίνει έλεγχος της ομαδοποίησης.

Πιο συγκεκριμένα η φάση δύο παίρνει τα κέντρα των υποομάδων που έχουμε βρει στη φάση ένα και έτσι σε κάθε επανάληψη της δεύτερης με τη μέθοδο CPQ βρίσκει το πιο κοντινό ζευγάρι ομάδων μεταξύ των αντιπροσωπευτικών σημείων.

Σε κάθε βήμα συγχώνευσης δύο ομάδων τα σημεία r μεταξύ όλων των σημείων των συγχωνευμένων ομάδων που βρίσκονται πιο κοντά στο κέντρο επιλέγονται ως αντιπρόσωποι της νέας ομάδας.

Χρησιμοποιώντας περισσότερα σημεία-αντιπροσώπους αντί του κέντρου των ομάδων, ο C²P μπορεί να συλλάβει το σχήμα και το μέγεθος των ομάδων. Επιπλέον καθορίζεται και το κοντινότερο ζευγάρι με τη μέθοδο CPQ.

Η φάση τελειώνει όταν επιτευχθεί ο απαιτούμενος αριθμός ομάδων που πρέπει να δημιουργηθούν. Συμπεραίνουμε ότι η δεύτερη φάση λειτουργεί ως ένας συσσωρευτικός αλγόριθμος ομαδοποίησης (**http://www.vldb.org/com,** Βαζιργιάννης, 2003).

2.2.4 ***Ομαδοποίηση βασισμένη στην πυκνότητα:***

Ο τρόπος αυτός της ομαδοποίησης στηρίζεται στην οργάνωση ομάδων των γειτονικών στοιχείων ενός συνόλου δεδομένων με βάση κάποια κριτήρια πυκνότητας.

Οι αλγόριθμοι που βασίζονται σε αυτή τη μέθοδο θεωρούν τις ομάδες ως πυκνές περιοχές στοιχείων στο χώρο όπου εξετάζονται τα δεδομένα μας, οι οποίες περιοχές χωρίζονται από τις υπόλοιπες που περιέχουν χαμηλή πυκνότητα.

Αλγόριθμοι βασισμένοι στην πυκνότητα:

- *DBSCAN*

Η βασική ιδέα του αλγορίθμου είναι ότι έχουμε ένα σύνολο από στοιχεία (D). Η περιοχή που εκτείνεται σε καθορισμένη πάντα ακτίνα (Eps) γύρω από τη «γειτονιά» κάθε στοιχείου κάθε ομάδας θα πρέπει να περιέχει έναν ελάχιστο αριθμό (minPts) στοιχείων.

Βάση των παραπάνω ισχύει ότι:

Ένα στοιχείο p είναι *άμεσα πυκνά προσεγγίσιμο* εάν:

a) Το στοιχείο ανήκει στο υποσύνολο των στοιχείων που βρίσκονται στη γειτονιά του q και,

b) Ο αριθμός των στοιχείων που βρίσκονται στη γειτονιά του q είναι μεγαλύτερος από ένα όριο του minPts (ελάχιστου αριθμού στοιχείων).

b*) *Τα στοιχεία αυτά ονομάζονται στοιχεία πυρήνα (core objects). Τα υπόλοιπα στοιχεία τα οποία δεν ανήκουν σε αυτή την κατηγορία λέγονται στοιχεία-όχι-πυρήνα (non-core-objects).*

Ένα στοιχείο p *είναι πυκνά προσεγγίσιμο* από ένα στοιχείο q, όπου p>Dq, εάν υπάρχει μία ακολουθία από p1,p2,p3,….pn στοιχεία όπου p1=q και pn=p, τέτοια ώστε το pi+1 να είναι πυκνά προσεγγίσιμο από το pi.

Ένα στοιχείο p *είναι πυκνά συνδεδεμένο* με ένα στοιχείο q εάν υπάρχει ένα στοιχείο o, τέτοιο ώστε τόσο το p όσο και το q να είναι πυκνά προσεγγίσιμα από το o.

Μία ομάδα C στο σύνολο των στοιχείων D είναι ένα υποσύνολο του D το οποίο ικανοποιεί τις ακόλουθες συνθήκες:

Για κάθε p,q Є D: εάν p Є c και q>Dp, τότε q Є C.

Για κάθε p,q Є C: το p είναι πυκνά συνδεδεμένο με το q

Έστω ότι C1,C2,….Cn ορίζουμε τις ομάδες του συνόλου D. Το σύνολο των στοιχείων του συνόλου τα οποία δεν ανήκουν σε καμία ομάδα Cn, το ονομάζουμε θόρυβο.

Βασική προϋπόθεση για να αρχίσει η εφαρμογή του αλγορίθμου είναι ο χρήστης να ορίσει την παράμετρο Eps της ακτίνας, στην οποία θα εκτείνεται η γειτονία κάθε στοιχείου του συνόλου των δεδομένων, και του ελάχιστου αριθμού σημείων minPts που ενδέχεται να υπάρχει στη γειτονιά.

Ο DBSCAN ξεκινά παίρνοντας ένα τυχαίο σημείο p του συνόλου και ανακτά τα στοιχεία που είναι πυκνά προσεγγίσιμα από το p.

Αν το p είναι στοιχείο πυρήνα, τότε ο αλγόριθμος ορίζει μία ομάδα. Αν δε συμβαίνει αυτό και το p είναι ένα ακραίο στοιχείο, τότε σημαίνει πως δεν υπάρχει κάποιο στοιχείο πυκνά προσεγγίσιμο από το p και έτσι το p συμπεριλαμβάνεται στο θόρυβο και ο DBSCAN συνεχίζει με την επεξεργασία των υπόλοιπων στοιχείων της βάσης δεδομένων που εξετάζουμε και η διαδικασία συνεχίζεται διαδοχικά.

Εντούτοις, ο αλγόριθμος αυτός σε περίπτωση που ανακαλύψει ομάδες με ακανόνιστα σχήματα αντιμετωπίζει προβλήματα διότι επηρεάζεται από τις τιμές των Eps και minPts οι οποίες δύσκολα προσδιορίζονται.

Επίσης, στην περίπτωση που υπάρχει μία πυκνή σειρά σημείων που συνδέει δύο ομάδες ο DBSCAN μπορεί να τελειώσει με την συγχώνευση των ομάδων αυτών καθώς υστερεί στην ευρωστία σημείων.

Λόγω του ότι εφαρμόζεται απευθείας στο σύνολο των δεδομένων προς ανάλυση καθίσταται ασύμφορος για επεξεργασία των στοιχείων μεγάλων βάσεων δεδομένων εξ' αιτίας του μεγάλου κόστους.

Τέλος, εάν πρόκειται για μεγάλο δείγμα ενδέχεται να υπάρχουν μεγάλες διακυμάνσεις στην πυκνότητα των σημείων μέσα σε κάθε ομάδα. Ωστόσο, ο αλγόριθμος αυτός δεν εφαρμόζει την τεχνική της τυχαίας δειγματοληψίας και για αυτό το λόγο δεν μπορεί να περιοριστεί το μέγεθος της εισόδου των στοιχείων.

Δομή DBSCAN

Algorithm DBSCAN (D,Eps,minPts)

// Προϋπόθεση: Όλα τα στοιχεία στο σύνολο δεδομένων D δεν έχουν τοποθετηθεί σε ομάδες.

FOR all objects o in D DO:

IF o δεν έχει ταξινομηθεί

Κάλεσε τη συνάρτηση expand_cluster προκειμένου να κατασκευαστεί μία ομάδα με ακτίνα Eps και ελάχιστο αριθμό στοιχείων MinPts το οποίο θα περιέχει το ο.

Function expand_cluster(o,D,Eps,minPts):

Ανάκτηση της Eps-γειτονιάς Neps(o) του ο;

If ׀ nEps(o) ׀<minPts// δηλαδή ο δεν είναι στοιχείο πυρήνα

THEN σημείωσέ το ο σαν θόρυβο, RETURN;

Else //το ο είναι στοιχείο πυρήνα

SELECT ένα νέο cluster_id και σημείωσε όλα τα στοιχεία στο Neps(o)

Με το τρέχον cluster_id;

Ώθησε όλα τα στοιχεία από το Neps(o)-{o} στη στοίβα seeds;

WHILE not seeds.empty() do

currentObject:=seed.top();

ανάκτηση της Eps-γειτονιάς του τρέχοντος στοιχείου;

If ׀ nEps(currentObject) ׀ >= minPts

SELECT όλα τα στοιχεία Neps(currentObject) που δεν έχουν ταξινομηθεί ακόμα ή έχουν σημειωθεί σαν θόρυβος,

Τοποθέτησε τα μη- κατηγοριοποιημένα στοιχεία στη στοίβα seeds

AND σημείωσε όλα αυτά τα αντικείμενα με το τρέχον cluster_id;

Seeds.pop(); RETURN

Γενικότερα έχει αποδειχθεί πως κατά την εισαγωγή ή τη διαγραφή ενός στοιχείου p, τα στοιχεία που επηρεάζονται είναι αυτά που ανήκουν στη γειτονιά του p καθώς επίσης και όλα

τα στοιχεία που είναι πυκνά προσεγγίσιμα από κάθε στοιχείο του συνόλου D στο οποίο περιλαμβάνεται το p.

Τα υπόλοιπα στοιχεία των ομάδων τα οποία δεν ανήκουν στο σύνολο των στοιχείων που επηρεάζονται κατά την εισαγωγή ή τη διαγραφή ενός στοιχείου p, δεν μεταβάλλονται καθόλου.

Κατά συνέπεια με βάση τον αλγόριθμο DBSCAN είναι δυνατό να σχεδιαστούν άλλοι αποδοτικοί αλγόριθμοι ώστε να ανταπεξέλθουν στις εισαγωγές ή διαγραφές των στοιχείων κατά τη διάρκεια της ομαδοποίησης ([21]).

§ *DENCLUE*

Ο αλγόριθμος που θα εξετάσουμε ανήκει και αυτός στην κατηγορία της ομαδοποίησης με βάση την πυκνότητα ο οποίος δίνει μία καινούρια προσέγγιση στο θέμα της ομαδοποίησης σε βάσεις δεδομένων.

Βασική ιδέα του Denclue είναι ότι θεωρεί τη συνολική πυκνότητα του συνόλου των στοιχείων σαν ένα άθροισμα συναρτήσεων επιρροής (influence functions).

Με τον όρο «συναρτήσεις επιρροής» εννοούμε τις συναρτήσεις οι οποίες περιγράφουν τις επιδράσεις κάθε σημείου από το σύνολο δεδομένων στη γειτονιά του.

Οι ομάδες προσδιορίζονται βάση των τοπικών μεγίστων της συνολικής συνάρτησης πυκνότητας (destiny attractors).

Ακόμη και οι ομάδες εκείνες οι οποίες έχουν ακανόνιστο σχήμα περιγράφονται εύκολα από μία απλή εξίσωση που και αυτή βασίζεται στη συνολική συνάρτηση πυκνότητας.

Ο Denclue για να χειριστεί καλά τα σύνολα δεδομένων που περιέχουν θόρυβο και για να ανακαλύψει τις ομάδες με ακανόνιστο σχήμα βασίζεται στις εξής παραμέτρους:

στη παράμετρο σ η οποία δείχνει τον τρόπο με τον οποίο ένα στοιχείο από το σύνολο δεδομένων επιδρά στην γειτονιά του

στη παράμετρο λ που περιγράφει εάν τα τοπικά μέγιστα της συνολικής συνάρτησης είναι σημαντικά και εάν ενδέχεται πιθανή μείωσή τους ώστε να βελτιωθεί η αποδοτικότητα ([21]).

2.2.5 ***Ομαδοποίηση βασισμένη σε πλέγμα:***

Οι αλγόριθμοι που ανήκουν σε αυτή την κατηγορία χρησιμοποιούνται στην ανάλυση χωρικών δεδομένων.

Κατά την εφαρμογή τους διαιρούν το χώρο σε ένα πεπερασμένο αριθμό κελιών, τα πλέγματα, υπολογίζοντας έτσι τον μέσο, τη διακύμανση, το ελάχιστο και μέγιστο άκρο καθώς και τον τύπο κατανομής που οι τιμές ακολουθούν διαμορφώνοντας έτσι μία ιεραρχική δομή στα προς εξέταση δεδομένα (Βαζιργιάννης, 2003).

Αλγόριθμοι βασισμένοι σε πλέγμα:

- ***STING***

Η μέθοδος που εφαρμόζει αυτός ο αλγόριθμος είναι η διαίρεση των χωρικών περιοχών σε ορθογώνια κελία βάση μιας ιεραρχικής δομής. Η μέθοδος αυτή μέσω των ομάδων που έχουν διαμορφωθεί υπολογίζει τις στατιστικές παραμέτρους των αριθμητικών στοιχείων που βρίσκονται μέσα σε κάθε κελί-πλέγμα.

Στη συνέχεια παράγεται μία ιεραρχική δομή των κελιών πλέγματος όπου κάθε κόμβος της δομής αυτής συνοψίζει τις πληροφορίες για τα στοιχεία που την αποτελούν οι οποίες μπορεί να βρίσκονται σε διαφορετικά επίπεδα.

Έτσι βασισμένος σε αυτήν την ιεραρχική δομή ο STING επιτρέπει τη χρήση των πληροφοριών ομαδοποίησης στην αναζήτηση των ερωτήσεων ή της αποδοτικής ανάθεσης ενός νέου στοιχείου σε ομάδες προς επεξεργασία (Τασουλής, 2007)

- ***Wave Cluster***

Ο αλγόριθμος αυτός ανακαλύφθηκε σχετικά πρόσφατα και βασίζεται και αυτός στην τεχνική της ομαδοποίησης βάση πλέγματος. Το όνομά του προέρχεται από τον μετασχηματισμό επεξεργασίας σημάτων διαφορετικής συχνότητας (wavelet transformation), ο οποίος εφαρμόζεται στα χωρικά δεδομένα του πεδίου συχνοτήτων.

Αρχικά τα δεδομένα ορίζοντας μια πολυδιάστατη δομή πλέγματος επάνω το διάστημά τους. Κάθε κελί πλέγματος περιέχει τις πληροφορίες για το σύνολο των σημείων που βρίσκονται

μέσα στο κελί. Μετατρέποντας τα χωρικά δεδομένα στο πεδίο συχνοτήτων μετασχηματίζεται το αρχικό διάστημα των δεδομένων αυτών.

Κατά τη διάρκεια του μετασχηματισμού η συνέλιξη των στοιχείων του κελιού με μια κατάλληλη συνάρτηση οδηγεί σε ένα μετασχηματισμένο διάστημα όπου προσδιορίζονται οι φυσικές ομάδες στα δεδομένα.

Ο μετασχηματισμός wavelet είναι ιδιαίτερα χρήσιμος επειδή χρησιμοποιεί φίλτρα για να δώσει έμφαση σε περιοχές που τα δεδομένα σχηματίζουν ομάδες ενώ ταυτόχρονα περιορίζουν την ασθενέστερη πληροφορία στα όρια των ομάδων αυτών. Η όλη διαδικασία απομακρύνει τα outliers αναλύοντας τα πολλές φορές, κάτι που συμβάλλει αποδοτικά στη μείωση του κόστους.

Σε τελευταίο στάδιο με την εφαρμογή του Wavelet Cluster προσδιορίζονται οι ομάδες με την εύρεση των πυκνών περιοχών στο μετασχηματισμένο χώρο (Βαζιργιάννης,2003).

2.2.6 ***Ομαδοποίηση υποχώρων:***

Η ομαδοποίηση υποχώρων εφαρμόζεται σε προβλήματα που προκύπτουν από τα δεδομένα υψηλών διαστάσεων. Αυτό συμβαίνει διότι λόγω της ύπαρξης των πολλών διαστάσεων, μαζί και αυτών που αντιστοιχούν σε θόρυβο, σχεδόν κάθε υποσύνολο παρουσιάζει χαμηλή πυκνότητα σημείων.

Οι αλγόριθμοι αυτής της κατηγορίας προσπαθούν να ανακαλύψουν ποια υποσύνολα ενός αρχικού συνόλου δεδομένων παρουσιάζουν καλύτερα αποτελέσματα ομαδοποίησης.

Με βάση αυτά αναλύουμε παρακάτω τους εξής αλγόριθμους που αντιστοιχούν σε αυτή την κατηγορία (Βαζιργιάννης, 2003).

- *CLIQUE*

Η μέθοδος που εφαρμόζει ο αλγόριθμος αυτός έχει σαν στόχο, ακολουθώντας μία προσέγγιση βασισμένη στην πυκνότητα ώστε να προσδιοριστούν οι ομάδες, να ορίζει αυτόματα

διάφορους υποχώρους του αρχικού συνόλου δεδομένων οι οποίοι θα επιτρέπουν στα στοιχεία του συνόλου που εξετάζουμε να ομαδοποιούνται καλύτερα.

Η διαδικασία που ακολουθεί ο CLIQUE είναι να προχωρά από τους χαμηλότερους σε διάσταση υποχώρους προς εκείνους με την κάθε φορά υψηλότερη ανακαλύπτοντας σε κάθε περίπτωση υποχώρου τις πυκνές περιοχές.

Για να προσδιοριστεί σωστά η πυκνότητα των σημείων ο Clique αρχικά τμηματοποιεί το χώρο των δεδομένων σε κελιά που έχουν ίσα διανύσματα μεταξύ τους.

Για ένα δεδομένο σύνολο διαστάσεων ο συνδυασμός των αντίστοιχων διαστημάτων στον αντίστοιχο υποχώρο (όπου κάθε διάστημα ορίζεται από το χρήστη του αλγορίθμου και αντιστοιχεί σε μία διάσταση του συνόλου των δεδομένων) λέγεται *μονάδα* (unit). Μία μονάδα καλείται πυκνή ένα ο αριθμός των σημείων που περιέχει είναι μεγαλύτερος από ένα δεδομένο όριο *t*, το οποίο όριο καθορίζεται και αυτό από το χρήστη.

Γενικότερα ο αλγόριθμος που περιγράφουμε ανακαλύπτει όλα τα πυκνά στοιχεία σε ένα χώρο με k διαστάσεις δημιουργώντας πυκνούς υποχώρους (k-1) διαστάσεων και έπειτα τα συνδέει μεταξύ τους ώστε να περιγραφούν οι ομάδες.

§ *PROCLUS*

Στο προηγούμενο κομμάτι αναφερθήκαμε στον clique αλγόριθμο ο οποίος όπως είπαμε εξετάζει διαφορετικούς υποχώρους για διαφορετικές ομάδες οι οποίες είναι συνδεδεμένες μεταξύ τους.

Ωστόσο ο παραπάνω αλγόριθμος υστερεί στον υπολογισμό μίας τμηματοποίησης των στοιχείων των οποίων οι ομάδες δε συνδέονται.

Έτσι ανακαλύφθηκε ο PROCLUS αλγόριθμος ο οποίος αναζητά σύνολα διαστάσεων τέτοια ώστε τα σημεία των δεδομένων που εμπεριέχονται να είναι πολύ πυκνά ομαδοποιημένα στους αντίστοιχους υποχώρους.

Εδώ ο χρήστης καθορίζει πάλι τον αριθμό των ομάδων που δημιουργούνται καθώς επίσης και το μέσο αριθμό διαστάσεων ανά ομάδα.

Αρχικά ο αλγόριθμός μας επιλέγει ένα τυχαίο σύνολο σημείων και σταδιακά βελτιώνει την ποιότητά τους εκτελώντας μία διαδικασία hill climbing η οποία είναι επαναληπτική και απορρίπτει έτσι τα σημεία από το σύνολο εκείνων που δεν ταιριάζουν στα υπό ανάλυση δεδομένα.

Στη συνέχεια, ο PROCLUS επιλέγει τις διαστάσεις κατά μήκος των οποίων τα σημεία έχουν τη μικρότερη μέση απόσταση μεταξύ τους προκειμένου να βρει το σύνολο των διαστάσεων που επηρεάζουν περισσότερο την κάθε ομάδα (Βαζιργιάννης,2003)

ΕΝΟΤΗΤΑ Γ’

3.1 ΘΕΩΡΗΤΙΚΗ ΣΥΓΚΡΙΣΗ ΜΕΤΑΞΥ ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ

Η μέθοδος της ομαδοποίησης είναι γεγονός πως συνδυάζει έννοιες διαφορετικών επιστημονικών πεδίων όπως η μηχανική μάθηση, η αναγνώριση προτύπων, στοιχεία στατιστικής και όλα αυτά πάνω σε βάσεις δεδομένων.

Για το λόγο αυτό και δημιουργήθηκαν οι αλγόριθμοι που εξετάσαμε παραπάνω, οι οποίοι ναι μεν χρησιμοποιούνται όλοι στη μέθοδο εξόρυξης γνώσης από δεδομένα, αλλά καθένας από αυτούς εξετάζει διαφορετικά το κάθε φορά σύνολο δεδομένων προς ανάλυση.

Ταξινομήσαμε λοιπόν τους αλγόριθμους σε τέσσερις διαφορετικές κατηγορίες με βάση τη μέθοδο ομαδοποίησης που ακολουθούν. Οι κατηγορίες των αλγορίθμων λοιπόν είναι διαιρετικοί, ιεραρχικοί, βασισμένοι στην πυκνότητα και βασισμένοι σε γράφους αλγόριθμοι.

Οι παραπάνω κατηγορίες αλγορίθμων διαφέρουν ως προς τα εξής:

τον τύπο δεδομένων που υποστηρίζουν, αν δηλαδή τα δεδομένα μας είναι αριθμητικά ή κατηγορικά

τη μορφή των ομάδων που δημιουργούνται

τη δυνατότητα να εντοπιστεί ο θόρυβος και τα outliers

το κριτήριο ομαδοποίησης που χρησιμοποιείται

την πολυπλοκότητα του κάθε αλγορίθμου της κάθε κατηγορίας

Παρακάτω βρίσκονται συγκεντρωμένοι συνοπτικά σε πίνακες οι κατηγορίες με τους αλγορίθμους που αντιστοιχούν στην κάθε μία.

3.1.1 *Διαιρετικοί Αλγόριθμοι:*

Όνομα	Τύπος δεδομένων	Outliers	Αποτελέσματα	Κριτήριο Ομαδοποίησης
K-means	Αριθμητικά	Όχι	Κέντρα ομάδων	$\text{Min}_{v1,v2..vk}(E_k)$ $E_k = \Sigma_k \Sigma_v d^2(x_k, v_i)$,για κάθε i=1
K-modes	Κατηγορικά	Όχι	Modes ομάδων	$\text{Min}_{\alpha1,\alpha2..\alpha k}(E_k)$ $E_k = \Sigma_k \Sigma_v d(X_i, Q_i)$,για κάθε i=1
PAM	Αριθμητικά	Όχι	Medoids ομάδων	$\text{Min}(TC_{ih}) = \Sigma_j C_{jih}$
CLARA	Αριθμητικά	Όχι	Medoids ομάδων	$\text{Min}(TC_{ih}) = \Sigma_j C_{jih}$
CLARANS	Αριθμητικά	Όχι	Medoids ομάδων	$\text{Min}(TC_{ih}) = \Sigma_j C_{jih}$
Fuzzy c-means	Αριθμητικά	Όχι	Κέντρα ομάδων	$\text{Min}_{u,v1,v2,..,vk}(J_m(U,V))$

3.1.2 *Ιεραρχικοί Αλγόριθμοι*

Όνομα	Τύπος	Outliers	Αποτελέσματα	Κριτήριο

	δεδομένων			Ομαδοποίησης
BIRCH	Αριθμητικά	Ναι	CF (πρόκειται για τον αριθμό των στοιχείων σε μία ομάδα N, το άθροισμα των στοιχείων στην ομάδα LS , το άθροισμα των τετραγώνων των N στοιχείων SS)	Ένα στοιχείο ανατίθεται στην ομάδα σύμφωνα με το επιλεγμένο μέτρο απόστασης. Επίσης ο ορισμός των ομάδων βασίζεται στις απαιτήσεις ότι ο αριθμός των σημείων σε κάθε ομάδα πρέπει να ικανοποιεί κάποιο όριο.
CURE	Αριθμητικά	Ναι	Ανάθεση δεδομένων σε ομάδες	Οι ομάδες με τα κοντινότερα ζεύγη αντιπροσώπων συγχωνεύονται σε κάθε εκτέλεση.
ROCK	Κατηγορικά	Ναι	Ανάθεση δεδομένων σε ομάδες	Max Ei

3.1.3 *Ιεραρχικοί και βασισμένοι σε γράφους Αλγόριθμοι:*

Όνομα	Τύπος	Αφηρημένα σχήματα	Outliers	Αποτελέσματα	Κριτήριο

	δεδομένων	ομάδων			ομαδοποίησης
C^2P	Αριθμητικά	Ναι	Όχι	Ανάθεση δεδομένων σε ομάδες	<u>Φάση1:</u> Εύρεση m-υποομάδων στο γράφημα που αναπαριστά το σύνολο δεδομένων. <u>Φάση2:</u> Με βάση τον αλγόριθμο Self-CPQ βρίσκουμε το πλησιέστερο ζευγάρι ομάδων. Έπειτα οι ομάδες συγχωνεύονται και η διαδικασία προχωράει επαναληπτικά μέχρι να οριστεί ο απαιτούμενος αριθμός

					ομάδων.

3.1.4 ***Βασισμένοι στην πυκνότητα Αλγόριθμοι:***

Όνομα	Τύπος δεδομένων	Αφηρημένα σχήματα ομάδων	Outliers	Αποτελέσματα	Κριτήριο ομαδοποίησης
DBSCAN	Αριθμητικά	Ναι	Ναι	Ανάθεση δεδομένων σε ομάδες	Ενώνει σημεία που είναι πυκνά προσεγγίσιμα από κάποια ομάδα
DENCLUE	Αριθμητικά	Ναι	Ναι	Ανάθεση δεδομένων σε ομάδες	Βασίζεται σε δύο παραμέτρους, στην παράμετρο σ η οποία καθορίζει την επίδραση ενός στοιχείου του συνόλου των δεδομένων μας στην γειτονία του και στην ξ παράμετρο, η οποία

					περιγράφει εάν ένας density-attractor είναι σημαντικός, επιτρέποντας μία μείωση του αριθμού των density-attractors και βοηθά στην βελτίωση της αποδοτικότητας.

3.1.5 ***Βασισμένοι σε πλέγμα Αλγόριθμοι:***

Όνομα	Τύπος δεδομένων	Αφηρημένα σχήματα ομάδων	Outliers	Αποτελέσματα	Κριτήριο ομαδοποίησης
WaveCluster	Χωρικά	Ναι	Ναι	Κατηγοριοποιημένα αντικείμενα	Αποσύνθεση του χώρου των χαρακτηριστικών με εφαρμογή μετασχηματισμού σε ομάδες θέτοντας κάποια όρια

					ομάδων.
STING	Χωρικά	Ναι	Ναι	Κατηγοριοποιημένα αντικείμενα	Διαίρεση του χώρου σε τετράγωνα κελιά και εφαρμογή μιας ιεραρχικής δομής. Κάθε κελί στο υψηλό επίπεδο τμηματοποιείται σε έναν αριθμό από μικρότερα κελιά στο επόμενο χαμηλότερο επίπεδο.

3.1.6 ***Χωρικοί Αλγόριθμοι:***

Όνομα	Τύπος δεδομένων	Αφηρημένα σχήματα	Outliers	Αποτελέσματα	Κριτήριο ομαδοποίησης
CLIQUE	Αριθμητικά	Ναι	Ναι	Περιγραφή πυκνών περιοχών	Εύρεση πυκνών περιοχών σε

					υποχώρους
PROCLUS	Αριθμητικά	Όχι	Όχι	Διάσταση ομάδων	Εύρεση όμοιων στοιχείων σε υποχώρους

(Βαζιργιάννης,2003)

3.2 ΠΑΡΑΤΗΡΗΣΕΙΣ:

§ ΔΙΑΙΡΕΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ

Στον πρώτο πίνακα αναφερόμαστε στα χαρακτηριστικά των διαιρετικών αλγορίθμων. Οι διαιρετικοί αλγόριθμοι έχουν σαν χαρακτηριστικό ότι δε μπορούν να χειριστούν τον θόρυβο και τα outliers και δε μπορούν να εφαρμοστούν σε ακανόνιστου σχήματος ομάδες καθώς βασίζονται σε μια συγκεκριμένη υπόθεση για να χωρίσουν το σύνολο των ομάδων, σύμφωνα με την οποία προσδιορίζεται εκ των προτέρων ο αριθμός των ομάδων. Επίσης παρατηρούμε πως οι αλγόριθμοι αυτοί εφαρμόζονται κυρίως πάνω σε αριθμητικά δεδομένα εκτός από τον k-Modes ο οποίος εφαρμόζεται σε κατηγορικά.

Ωστόσο έχουμε δει πως ο αλγόριθμος αυτός βασίζεται στη μέθοδο k-Means, η οποία αφορά αριθμητικά δεδομένα, ώστε να ανακαλύπτει ομάδες μέσα σε ένα σύνολο δεδομένων χρησιμοποιώντας παράλληλα νέες έννοιες για να μπορεί να εφαρμόζεται σε κατηγορικά δεδομένα.

Έτσι τα κέντρα των ομάδων κατά τον k-modes λέγονται modes ενώ χρησιμοποιείται ένα διαφορετικό μέτρο ανομοιότητας για να εξεταστούν τα νέα δεδομένα.

Όλοι λοιπόν οι διαιρετικοί αλγόριθμοι καθορίζουν από πριν τον αριθμό των ομάδων εκτός από τον CLARANS ο οποίος για να ξεκινήσει την εφαρμογή του πρέπει να είναι γνωστός ο μέγιστος αριθμός των γειτόνων μιας ομάδας και τα τοπικά ελάχιστα προκειμένου να

καθοριστεί η τμηματοποίηση ενός συνόλου δεδομένων πάνω στο οποίο θα εφαρμοστεί ο αλγόριθμος αυτός.

Το αποτέλεσμα της διαδικασίας της τμηματοποίησης είναι η εύρεση ενός συνόλου των κέντρων των ομάδων που έχουν προσδιοριστεί. Συμβαίνει όμως να μην μπορεί να βρεθεί απόλυτα το κέντρο μιας ομάδας. Έτσι κάποιοι αλγόριθμοι, όπως για παράδειγμα ο PAM χρησιμοποιούν τα πιο κεντρικά τοποθετημένα στοιχεία μιας ομάδας τα οποία αυτά στοιχεία ονομάζονται medoids.

Ο στόχος λοιπόν των διαιρετικών αλγορίθμων είναι να ελαχιστοποιηθεί η απόσταση των αντικειμένων μέσα σε μία ομάδα από το κέντρο της. Το κριτήριο που εφαρμόζει ο k-Means στοχεύει σε αυτόν το στόχο απόλυτα. Υπάρχουν και κάποιοι αλγόριθμοι που στοχεύουν στην ελαχιστοποίηση της απόστασης των αντικειμένων μιας ομάδας από το medoid της όπως ο PAM.

Οι αλγόριθμοι CLARA και CLARANS βασίζονται στο κριτήριο ομαδοποίησης που ο PAM χρησιμοποιεί. Εφαρμόζονται σε δείγματα του αρχικού συνόλου δεδομένων στο οποίο πραγματοποιείται η διαδικασία της ομαδοποίησης και μπορούν να εξετάσουν μεγάλα σύνολα δεδομένων, σε αντίθεση με τον PAM ο οποίος εφαρμόζεται σε μικρά σύνολα.

Συγκεκριμένα ο CLARA απεικονίζει πολλαπλάσια δείγματα του συνόλου των δεδομένων και εφαρμόζει το κριτήριο ομαδοποίησης του PAM σε κάθε δείγμα βρίσκοντας έτσι την καλύτερη ομαδοποίηση στα δεδομένα προς εξέταση.

Τα μειονεκτήματα όμως της μεθόδου αυτής είναι ότι η αποδοτικότητά της εξαρτάται από τον μέγεθος των δειγμάτων και ότι τα αποτελέσματα που παράγονται από την ομαδοποίηση βασίζονται μόνο στα δείγματα ενός συνόλου δεδομένων. Έτσι εάν το δείγμα δεν είναι απόλυτα αντιπροσωπευτικό μια ομαδοποίηση η οποία βασίζεται στα δείγματα δεν θα είναι αντιπροσωπευτική για ολόκληρο το σύνολο των δεδομένων.

Ο CLARANS από την άλλη είναι ένας συνδυασμός των PAM και CLARA που όμως είναι αποδοτικότερος και περισσότερο επεκτάσιμος από αυτούς.

Για παράδειγμα, ο CLARANS ψάχνει σε ένα υποσύνολο του συνόλου των δεδομένων προκειμένου να καθορίσει τις ομάδες ενώ ο PAM όχι.

Τα υποσύνολα υπολογίζονται τυχαία σε κάθε βήμα ης αναζήτησης αντίθετα με τον CLARA ο οποίος έχει ένα σταθερό βήμα κάθε φορά.

Τέλος, ο Fuzzy k-means αποτελεί έναν αντιπροσωπευτικό αλγόριθμο της διαδικασίας της ασαφούς ομαδοποίησης, σύμφωνα με την οποία οι αλγόριθμοι που ανήκουν σε αυτήν την κατηγορία θεωρούν ότι ένα αντικείμενο μπορεί να ανήκει σε μία μόνο ομάδα, παρ' όλο που τα όρια μιας ομάδας μπορούν δύσκολα να καθοριστούν.

Ο Fuzzy k-means βασίζεται στις έννοιες του k-Means για να χωρίσει ένα σύνολο δεδομένων σε ομάδες ενώ ταυτόχρονα εισάγει τη έννοια της αβεβαιότητας και αναθέτει τα στοιχεία στις ομάδες με διαφορετικό βαθμό πίστης.

§ ΙΕΡΑΡΧΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ

Όσον αφορά τον πίνακα με τους ιεραρχικούς αλγορίθμους ομαδοποίησης παρατηρούμε πως κάποιοι από τους αλγορίθμους αυτούς είναι πιο αποδοτικοί στη διαχείριση θορύβου από ότι είναι οι διαιρετικοί. Παρ ' όλα ταύτα υστερούν από άποψη αποδοτικότητας όταν ο αριθμός των δεδομένων είναι πολύ υψηλός.

Ωστόσο ο BIRCH αντιμετωπίζει αυτό το πρόβλημα χρησιμοποιώντας μία ιεραρχική δομή δεδομένων που την ονομάζει CF-tree για την ομαδοποίηση πολλών φάσεων. Έτσι μία μόνο σάρωση των δεδομένων μπορεί να δώσει μία καλή ομαδοποίηση ενώ περισσότερες από μία σαρώσεις βελτιώνουν ολοένα και περισσότερο την ποιότητα της ομαδοποίησης.

Όμως εφαρμόζεται μόνο σε αριθμητικά δεδομένα και επηρεάζεται από τη σειρά με την οποία αυτά εισάγονται. Παράγει δηλαδή διαφορετικές ομάδες ανάλογα με τη σειρά εισόδου των δεδομένων.

Επίσης δεν είναι αποδοτικός όταν οι ομάδες παρουσιάζουν ανομοιομορφία καθώς χρησιμοποιεί ως αντιπροσωπευτικό στοιχείο το κέντρο κάθε ομάδας.

Από την άλλη ο CURE συνδυάζει κατά την εφαρμογή του την τυχαία δειγματοληψία και ομαδοποίηση για να διαχειριστεί μεγάλες βάσεις δεδομένων αναπαριστώντας κάθε ομάδα με πολλαπλά σημεία όταν αναγνωρίζει ομάδες μη-σφαιρικού σχήματος και μεγάλων αποκλίσεων.

Καθορίζει τα αντιπροσωπευτικά σημεία για μία ομάδα επιλέγοντας τα πιο απομακρυσμένα σημεία από μία ομάδα και συρρικνώνοντας τα προς το κέντρο της κάθε ομάδας κατά ένα συγκεκριμένο ποσοστό.

Όμως, ο αλγόριθμος αυτός είναι ευαίσθητος στον αριθμό των αντιπροσωπευτικών σημείων, στον παράγοντα συρρίκνωσης και στον αριθμό των τμημάτων που δημιουργεί. Η ποιότητα των αποτελεσμάτων της ομαδοποίησης εξαρτάται από αυτούς τους τρεις παράγοντες.

Ο αλγόριθμος ROCK είναι ένας αντιπροσωπευτικός αλγόριθμος της κατηγορίας των ιεραρχικών αλγορίθμων ομαδοποίησης ο οποίος χρησιμοποιείται για κατηγορικά δεδομένα. Έτσι εισάγει μία διαφορετική έννοια προκειμένου να μετρήσει την εγγύτητα ανάμεσα στα διάφορα ζεύγη σημείων που δημιουργούνται στις ομάδες, οποία έννοια καλείται link.

Δεν χρησιμοποιεί την τεχνική των τυχαίων δειγμάτων και χειρίζεται με επιτυχία τα δεδομένα που παρουσιάζουν σημαντικές διαφορές στο μέγεθος των ομάδων.

§ Ιεραρχικοί και βασισμένοι σε γράφους Αλγόριθμοι

Η επόμενη κατηγορία αλγορίθμων είναι εκείνη των ιεραρχικών και βασισμένων σε γράφους οι οποίοι συνδυάζουν ομαδοποίηση βασισμένη σε γράφους με τις μεθόδους των ιεραρχικών αλγορίθμων.

Ας επισημάνουμε πως η διαδικασία της ομαδοποίησης λαμβάνει υπ' όψη της τόσο το καταπώς συνδέονται δύο ομάδες όσο και την εγγύτητά τους πριν αυτές ενωθούν. Η διαδικασία της ένωσης τους βασίζεται σε ένα δυναμικό μοντέλο που προκύπτει μέσω της διαδικασίας που ακολουθεί ο εκάστοτε αλγόριθμος και το οποίο βοηθά στην ανακάλυψη ομοιογενών ομάδων. Αυτή η διαδικασία που περιγράφουμε εφαρμόζεται σε όλους τους τύπους δεδομένων αρκεί να καθοριστεί πρώτα η κατάλληλη συνάρτηση ομοιότητας.

Ο CHAMELEON μετρά την ανομοιότητα ανάμεσα στις ομάδες σε ένα δυναμικό μοντέλο αντίθετα με τους ιεραρχικούς αλγόριθμους. Παρόλο όμως που δουλεύει αποτελεσματικά για την εύρεση ομάδων με περίεργα σχήματα δεν μπορεί να εφαρμοστεί σε μεγάλες βάσεις δεδομένων.

Ο C^2P από την άλλη πλευρά συνδυάζει τα πλεονεκτήματα των ιεραρχικών και βασισμένων σε γράφους αλγορίθμων επιτυγχάνοντας καλή ποιότητα ομαδοποίησης ενώ ταυτόχρονα προσαρμόζεται στη διαχείριση μεγάλου όγκου δεδομένων.

Χρησιμοποιεί χωρικές μεθόδους προσπέλασης καθορίζοντας έτσι τα κοντινότερα μεταξύ τους σημεία.

Τέλος εκτελείται σε δύο φάσεις, όπου κατά την πρώτη ορίζει τον αριθμό των υποομάδων και κατά τη δεύτερη τον αριθμό των αντιπροσωπευτικών σημείων.

§ Βασισμένοι στην πυκνότητα Αλγόριθμοι

Οι αλγόριθμοι αυτής της κατηγορίας μπορούν να χειρίζονται αποδοτικά τον θόρυβο καθώς επίσης και διάφορα σύνολα δεδομένων των οποίων οι ομάδες έχουν ακανόνιστα σχήματα όπως σπιράλ, ελλειψοειδή, κυλινδρικά.

Ο DBSCAN αλγόριθμος είναι ευαίσθητος στις παραμέτρους Eps και MinPts. Με τον όρο Epts αναφερόμαστε στην ακτίνα που εκτείνεται η γειτονιά ενός αντικειμένου και με τον όρο MinPts, στον ελάχιστο αριθμό στοιχείων που βρίσκονται στη γειτονιά ενός αντικειμένου. Είναι φανερό λοιπόν πως αυτές οι παράμετροι είναι δύσκολο να καθοριστούν.

Ο DENCLUE ομοίως πρέπει να επιλέξει προσεχτικά τις τιμές των παραμέτρων εισόδου διότι οι παράμετροι αυτοί επηρεάζουν σημαντικά την ποιότητα των αποτελεσμάτων της ομαδοποίησης.

Ωστόσο υπερτερεί σχετικά με τους υπόλοιπους αλγορίθμους ομαδοποίησης διότι ο DENCLUE βασίζεται σε ένα καλά θεμελιωμένο μαθηματικό ορισμό ενώ ταυτόχρονα γενικεύει και άλλες μεθόδους ομαδοποίησης όπως είναι η διαιρετική και η ιεραρχική.

Επίσης, μπορεί να χειριστεί αποδοτικά σύνολα δεδομένων με μεγάλο όγκο θορύβου ενώ επιτρέπει τη μαθηματική περιγραφή ομάδων ακανόνιστου σχήματος σε πολυδιάστατα σύνολα δεδομένων.

Τέλος, χρησιμοποιεί τεχνικές ομαδοποίησης βασισμένης σε πλέγμα και διατηρεί πληροφορία μόνο για τα κελιά τα οποία περιέχουν σημεία, όπου χειρίζεται αυτά τα κελιά

χρησιμοποιώντας δομές βασισμένες σε δένδρα επιτυγχάνοντας έτσι να είναι γρηγορότερος από το DBSCAN.

Ωστόσο οι αλγόριθμοι που ανήκουν σε αυτή την κατηγορία δεν χρησιμοποιούν κάποια μορφή δειγματοληψίας με αποτέλεσμα να μη μπορούν να μειώσουν το κόστος εισόδου-εξόδου και αποτυγχάνουν σε περίπτωση που προσπαθούν να χρησιμοποιήσουν τεχνικές σαν την δειγματοληψία με σκοπό να μειώσουν το μέγεθος του συνόλου δεδομένων εισόδου.

Αυτό συμβαίνει γιατί μπορεί να υπάρχει μεγάλη διαφορά ανάμεσα στην πυκνότητα στις ομάδες του δείγματος σε σχέση με τις ομάδες του συνόλου των δεδομένων.

Βασισμένοι σε πλέγμα Αλγόριθμοι

Οι αλγόριθμοι που ανήκουν σε αυτή την κατηγορία έχουν σαν κύριο χαρακτηριστικό να ορίζουν ένα πλέγμα για τον χώρο των δεδομένων, μέσα στον οποίο καθορίζουν τις ομάδες.

Γενικότερα οι αλγόριθμοι οι οποίοι βασίζονται σε πλέγμα μπορούν να χειριστούν δεδομένα με ακανόνιστα σχήματα, να αναγνωρίσουν τον θόρυβο και να χειρίζονται μεγάλες βάσεις δεδομένων.

Ο STING είναι ένας από τους πιο γνωστούς αλγορίθμους που μπορούν να βασιστούν σε πλέγμα. Βασισμένος στη μέθοδο της δημιουργίας πλέγματος, χωρίζει το χώρο σε τετράγωνα κελιά, κατά τη διάρκεια που τον διασχίζει, αποθηκεύοντας παράλληλα τις στατιστικές παραμέτρους των τιμών των στοιχείων μέσα στα κελιά. Χάρη στη δομή του πλέγματος μπορεί και επεξεργάζεται παράλληλα τα δεδομένα των ομάδων που έχει δημιουργήσει καθώς επίσης και να τα ενημερώνει αυξητικά.

Μειονέκτημα του αλγορίθμου STING είναι πως η ποιότητα των αποτελεσμάτων που παράγει, εξαρτάται από την διαβάθμιση του χαμηλότερου επιπέδου του πλέγματος. Αυτό συμβαίνει διότι χρησιμοποιεί μία προσέγγιση πολλαπλής ανάλυσης για να αναλύσει τις ομάδες που έχουν δημιουργηθεί.

Επίσης ο αλγόριθμος αυτός δεν υπολογίζει τις χωρικές συσχετίσεις ανάμεσα στα «παιδία» και τα γειτονικά κελία ώστε να μπορέσει να κατασκευάσει το κελί «πατέρα» και έτσι όλα τα όρια

των ομάδων είναι είτε οριζόντια είτε κάθετα ,γι' αυτό και δεν είναι απόλυτα έγκυρη η ποιότητα των αποτελεσμάτων τα οποία παράγονται.

Ο WAVECLUSTER από την άλλη πλευρά, είναι σε θέση να ανακαλύψει ομάδες με περίεργα γεωμετρικά σχήματα καθώς στηρίζεται σε γνωστές τεχνικές επεξεργασίας σήματος.

Δεν απαιτεί να γνωρίζει από πριν τον αριθμό των ομάδων που υπάρχουν ούτε είναι απαραίτητο να γνωρίζει την ακτίνα της γειτονιάς.

Με λίγα λόγια δεν χρειάζεται να λάβει υπ' όψη τους τις παραπάνω παραμέτρους μολονότι η εκ των προτέρων εκτίμηση του αναμενόμενου αριθμού των ομάδων βοηθά στην επιλογή της σωστής ανάλυσης των ομάδων.

Χάρη σε μελέτες που έχουν διεξαχθεί αναφέρεται ότι ο WaveCluster αλγόριθμος είναι πιο αποδοτικός και η ποιότητα των αποτελεσμάτων του είναι πιο έγκυρη σε αντίθεση με άλλους όπως οι BIRCH, CLARANS, και DBSCAN.

Παρ' όλα αυτά ο αλγόριθμος δεν είναι τόσο αποδοτικός και δε προτιμάται όταν έχουμε να μελετήσουμε πολυδιάστατους χώρους.

§ ΧΩΡΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ

Οι αλγόριθμοι που ανήκουν σε αυτή την κατηγορία μπορούν να αναπαραστήσουν και να επεξεργαστούν μία πολύπλοκή χωρική δομή ενός συγκεκριμένου μεγέθους σχήματος. Οι τεχνικές που χρησιμοποιούν βασίζονται σε τετραδικό δένδρο, r-δένδρο, k-d δένδρο.

Κύριο χαρακτηριστικών των αλγορίθμων αυτών είναι ότι εντοπίζουν ομάδες με διαφορετικά γεωμετρικά σχήματα σε αντίθεση με τους άλλους αλγορίθμους οι οποίοι χρησιμοποιώντας κέντρα βάρους και απλές μετρήσεις δεν μπορούν να χειριστούν ομάδες με ακανόνιστα σχήματα.

ΕΝΟΤΗΤΑ Δ’

4.1 ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΩΝ DBSCAN, k-MODES ΚΑΙ SIMPLE K-MEANS

Στην ενότητα αυτή θα ασχοληθούμε με την εφαρμογή των αλγορίθμων ομαδοποίησης δεδομένων και συγκεκριμένα με τον DBSCAN, αλγόριθμο βασισμένο στην πυκνότητα και με τον Simple k-Means ο οποίος είναι ένας διαιρετικός αλγόριθμος καθώς επίσης και με τον k-Modes ο οποίος είναι και αυτός ένας διαιρετικός αλγόριθμος, το λογισμικό του οποίου δημιουργήθηκε από τον εισηγητή μας Νικόλαο Μαστρογιάννη.

Θα εφαρμόσουμε λοιπόν τους αλγορίθμους DBSCAN και k-Means πάνω σε ορισμένες βάσεις δεδομένων που συλλέξαμε από την ηλεκτρονική σελίδα της εφαρμογής WEKA.

Η πλατφόρμα WEKA είναι ένα ανοιχτού κώδικα εργαλείο που υλοποιεί μία πληθώρα από αλγορίθμους εξόρυξης δεδομένων. Πρόκειται ουσιαστικά για μία συλλογή από αλγορίθμους μηχανικής μάθησης για διαδικασίες εξόρυξης γνώσης.

Οι αλγόριθμοι μπορούν είτε να εφαρμοστούν απευθείας σε ένα σύνολο δεδομένων ή να χρησιμοποιηθούν από οποιαδήποτε εφαρμογή Java.

Το WEKA περιέχει εργαλεία που αφορούν την προ-επεξεργασία δεδομένων και τις μεθόδους εξόρυξης γνώσης από δεδομένα όπως είναι η ταξινόμηση, η ομαδοποίηση, η παλινδρόμηση και η εύρεση κανόνων συσχέτισης.

Επίσης περιλαμβάνει εργαλεία γραφικής αναπαράστασης των αποτελεσμάτων. Είναι σχεδιασμένο με τέτοιο τρόπο ώστε να είναι εύκολο να ενσωματωθούν καινούριοι αλγόριθμοι μηχανικής μάθησης.

Εμείς πρόκειται να εφαρμόσουμε τους αλγορίθμους DBSCAN στα αριθμητικά δεδομένα και τους Simple k-Means και k-Modes στα κατηγορικά δεδομένα, βασιζόμενοι σε δεδομένα τα οποία πήραμε από την βάση δεδομένων UCI Machine Learning Repository και τα μετατρέψαμε σε μορφή κατάλληλη ώστε να εφαρμοστούν στην πλατφόρμα WEKA.

Παρακάτω παρατίθενται λεπτομερώς τα στοιχεία κάθε βάσης δεδομένων που πρόκειται να εξετάσουμε.

4.1.1 **Αριθμητικά Δεδομένα:**

§ Breast Cancer Wisconsin(original):

Πρόκειται για μία βάση δεδομένων η οποία δόθηκε από τον Olvi Mangasarian και περιλαμβάνει δεδομένα σχετικά με τον καρκίνο του μαστού από τις αναφορές του Dr. WIlliam H. Wolberg από το ιατρικό πανεπιστήμιο του Wisconsin των Η.Π.Α.

Το σετ δεδομένων αποτελείται από 286 παρατηρήσεις οι οποίες εξετάζονται κάθε φορά ανάλογα με 5 μεταβλητές:

§ Την ηλικία κατά την οποία εμφανίζεται η ασθένεια

§ Αν η ασθένεια εμφανίζεται πριν ή μετά την εμμηνόπαυση

§ Το που βρίσκεται ο όγκος (αριστερά ή δεξιά)

§ Το σημείο του μαστού που είναι ο όγκος

§ Το αν δέχεται η ασθενής ακτινοβολία

§ Το αν επανεμφανίζεται η ασθένεια μετά τη θεραπεία

Από τα αποτελέσματα προέκυψαν τα εξής:

Όσον αφορά την ηλικία των ασθενών έχουμε τον εξής πίνακα:

Κλάση	Πληθυσμός
10-19	0
20-29	1

30-39	36
40-49	90
50-59	96
60-69	57
70-79	6
80-89	0
90-99	0

Παρατηρούμε λοιπόν πως η ασθένεια εμφανίζεται περισσότερο μεταξύ των ηλικιών 30-69, ενώ στις ηλικίες 0-19 και 80 και πάνω δεν παρατηρούνται περιστατικά.

Όσον αφορά την εμμηνόπαυση παρατηρήθηκε πως μόλις 7 γυναίκες οι οποίες είχαν περάσει το στάδιο της εμμηνόπαυσης ήταν ασθενείς, ενώ 129 γυναίκες έπασχαν από τη νόσο κατά τη διάρκεια της εμμηνόπαυσης. Οι υπόλοιπες, δηλαδή οι 150 έπασχαν από την νόσο πριν από την εμμηνόπαυση.

Ηλικία	Πληθυσμός
40+	7
40	129
Προ-εμμηνόπαυση	150

Έπειτα αναφερόμαστε στο μέγεθος του όγκου:

Μέγεθος	Πληθυσμός
0-4	8
5-9	4
10-14	28
15-19	30
20-24	50
25-29	54
30-34	60
35-39	19
40-44	22
45-49	3
50-54	8
55-59	0

Παρατηρούμε πως το μέγεθος του όγκου στις περισσότερες γυναίκες κυμαίνεται στα 10-40 χιλιοστά.

Έπειτα παρατηρούμε πως οι 152 γυναίκες εμφανίζουν τη νόσο στον αριστερό μαστό ενώ οι υπόλοιπες 134 στο δεξί. Δηλαδή η αναλογία είναι περίπου ίση.

Εδώ κάνουμε στατιστικές μετρήσεις σχετικά με το μέρος στο οποίο βρίσκεται ο όγκος του μαστού. Έχουμε λοιπόν τα εξής:

Περιοχή όγκου	Πληθυσμός
Αριστερά και ψηλά	97
Αριστερά και χαμηλά	110
Δεξιά και ψηλά	33
Δεξιά και χαμηλά	24
Στο κέντρο	21

Παρατηρούμε λοιπόν, ότι η ασθένεια εμφανίζεται κυρίως στο αριστερό μέρος του μαστού, και τις πιο πολλές φορές στην χαμηλή αριστερή περιοχή του μαστού.

Μετά οι 286 ασθενείς ρωτήθηκαν αν υπόκεινται σε όποια είδους ακτινοβολία και οι 218 απάντησαν αρνητικά. Μόλις 68 από αυτές κάνουν χημειοθεραπεία.

Τέλος, όσον αφορά το κατά πόσο επανεμφανίζεται η ασθένεια μετά τη θεραπεία στις ασθενείς, οι 201 απάντησαν αρνητικά, ενώ στις 85 περιπτώσεις υπάρχουν μεταστάσεις.

Εφαρμογή DBSCAN:

Λόγω του ότι ο αλγόριθμος αυτός εφαρμόζεται απευθείας στο σύνολο των δεδομένων προς ανάλυση καθίσταται ασύμφορος για επεξεργασία των στοιχείων μεγάλων βάσεων δεδομένων. Έτσι λοιπόν δεν είναι ο κατάλληλος για να τρέξει τα 286 δεδομένα μας.

Επομένως προχωράμε σε διαχωρισμό του ποσοστού των παρατηρήσεων μας κατά 66% και εκτελώντας τον αλγόριθμο προκύπτει πως η ακτίνα γύρω από τη γειτονιά κάθε στοιχείου κάθε ομάδας έχει μέγεθος 0,9 και περιέχει 6 στοιχεία (minPoints).

Όμως πάλι δεν είναι αποδοτικός διότι δεν παρουσιάζει τις ομάδες των στοιχείων.

Οπότε αφαιρούμε αυτή τη φορά 2 από τις 6 μεταβλητές ,εκείνες που αφορούν το που εμφανίζεται ο όγκος και εκείνες που αφορούν την περιοχή του μαστού όπου παρουσιάζεται η νόσος.

Προκύπτει λοιπόν ότι ομαδοποιούνται οι 80 από τις 286 παρατηρήσεις σε 11 ομάδες:

0	6
1	6
2	12
3	7
4	6
5	7
6	7
7	9
8	6
9	7
10	7

Τώρα θα αφαιρέσουμε την μεταβλητή που αφορά το αν δέχονται ακτινοβολία οι ασθενείς. Προκύπτουν 16 ομάδες οι οποίες περιέχουν 127 παρατηρήσεις:

0	8

1	8
2	14
3	9
4	8
5	6
6	8
7	6
8	6
9	7
10	10
11	6
12	10
13	6
14	7
15	8

Αν αφαιρέσουμε και την περιοδικότητα της νόσου, την τελευταία μεταβλητή μας προκύπτουν 18 ομάδες όπου περιλαμβάνουν συνολικά 182 παρατηρήσεις:

0	13
1	177
2	14
3	8
4	12
5	9
6	7
7	7
8	18
9	7
10	9
11	8
12	16
13	6
14	7
15	6

16	10
17	8

Έπειτα αφαιρούμε το μέγεθος του όγκου της νόσου και παρουσιάζονται 7 ομάδες οι οποίες στο σύνολό τους περιλαμβάνουν 278 παρατηρήσεις:

0	81
1	59
2	33
3	55
4	35
5	9
6	6

Τέλος θα αφαιρέσουμε τη μεταβλητή της εμμηνόπαυσης και θα αφήσουμε να εκτελέσει ο αλγόριθμος μόνο τις παρατηρήσεις που αναφέρονται στην ηλικία. Από αυτή την εφαρμογή προκύπτουν 5 ομάδες οι οποίες συνολικά περιλαμβάνουν 285 στοιχεία.

0	90
1	96
2	57
3	36

4	6

Εφαρμογή k-Means:

Εφαρμόζοντας τώρα τον k-Means στα στοιχεία που έχουμε συλλέξει παρατηρούμε πως σε σχέση με τον DBSCAN αλγόριθμο, ο k-Means ομαδοποιεί κατ' ευθείαν τα δεδομένα μας από την πρώτη κιόλας εκτέλεσή του.

Πιο συγκεκριμένα:

Κατά την πρώτη εκτέλεση του αλγορίθμου σε όλο το δείγμα ως έχει τα αποτελέσματα που παίρνουμε είναι η δημιουργία δύο ομάδων, η πρώτη εκ των οποίων περιέχει 226 παρατηρήσεις και η δεύτερη 60.

Κατά τη δεύτερη εκτέλεση του αλγορίθμου αγνοούμε τη μεταβλητή του μαστού και παίρνουμε τα ίδια αποτελέσματα. Όπως επίσης το ίδιο συμβαίνει και αν αφαιρέσουμε και τις μεταβλητές που αφορούν την περιοχή που εμφανίζεται η νόσος και το αν δέχεται το άτομο ακτινοβολία.

Στην περίπτωση όμως που αφαιρούμε τη μεταβλητή της περιοδικότητας της εμφάνισης της νόσου παρατηρούμε πως αυτή τη φορά πάλι δημιουργούνται δύο ομάδες, εκ των οποίων η πρώτη περιέχει 170 παρατηρήσεις ενώ η δεύτερη 116.

Έπειτα αφαιρώντας τη μεταβλητή του μεγέθους που έχει ο όγκος οι δυο ομάδες περιέχουν 169 και 117 παρατηρήσεις αντίστοιχα.

Τέλος, αφήνοντας μόνο τη μεταβλητή που αφορά την ηλικία να εξεταστεί παρατηρούμε πως στην πρώτη ομάδα που δημιουργείται οι παρατηρήσεις είναι 196 ενώ στη δεύτερη ομάδα 90.

Ας εξετάσουμε τώρα την περίπτωση όπου ο αλγόριθμος εφαρμόζεται στο 66% των παρατηρήσεων, δηλαδή οι 188. Οι ομάδες που δημιουργούνται περιέχουν 98 στοιχεία και είναι δύο, η μία περιλαμβάνει τα 58 στοιχεία ενώ η άλλη τα υπόλοιπα 40.

Εάν ο αλγόριθμος εφαρμοστεί στο 50% των συνολικών παρατηρήσεων μας παρατηρούμε πως δημιουργούνται δύο ομάδες, εκ των οποίων η πρώτη περιέχει 84 στοιχεία και η δεύτερη 59.

- Ionosphere

Το σετ δεδομένων που θα εξετάσουμε αυτή τη φορά αποτελείται από σήματα που δέχεται το ραντάρ τα οποία σήματα αφορούν τα ηλεκτρόδια που βρίσκονται στην ιονόσφαιρα.

Αποτελείται από 351 στοιχεία και 35 μεταβλητές, εκ των οποίων οι 34 είναι συνεχείς ενώ η $35^{η}$ παίρνει δυαδική μορφή καθώς αφορά τα καλά σήματα που λαμβάνουν τα ραντάρ και τα οποία εμφανίζουν αποδεικτικά στοιχεία για το είδος της δομής της ιονόσφαιρας και, τα κακά των οποίων τα σήματα ξεπερνούν την ιονόσφαιρα και άρα δεν εμφανίζουν δείγματα της δομής της.

Η βάση δεδομένων πάνω στην οποία θα εκτελεστούν οι αλγόριθμοί μας δόθηκε από το πανεπιστήμιο ερευνών John Hopkins από μία ομάδα φυσικών ερευνητών.

Εφαρμογή DBSCAN:

Κατά την πρώτη εκτέλεση του αλγορίθμου παρατηρούμε πως ομαδοποιούνται οι 240 παρατηρήσεις σε δύο ομάδες, μία των 223 και άλλη μία των υπόλοιπων 17. Απομένουν 111 ακόμη.

Κατά τη δεύτερη εκτέλεση του αλγορίθμου αφαιρούμε την πρώτη μεταβλητή από τις 35 και βλέπουμε ότι σχηματίζονται πάλι δύο ομάδες, οι πρώτη αποτελείται από 223 στοιχεία που παίρνουν την τιμή 0, και η δεύτερη από 22 οι οποίες παίρνουν την τιμή 1. Απομένουν 106 στοιχεία χωρίς να αντιστοιχίζονται σε κάποια ομάδα.

Αφαιρώντας και την επόμενη μεταβλητή παραμένουν τα ίδια αποτελέσματα. Όσο όμως προχωράμε στην αφαίρεση των υπόλοιπων μεταβλητών παρατηρούμε πως σε κάθε αφαίρεση η δεύτερη ομάδα της οποίας τα στοιχεία παίρνουν την τιμή 1 αυξάνεται κατά ένα στοιχείο κάθε φορά, με αποτέλεσμα αφού αφαιρεθούν και οι 34 η πρώτη ομάδα να περιλαμβάνει 223 στοιχεία και η δεύτερη 126.

Αν τώρα αφαιρέσουμε μόνο τη μεταβλητή που παίρνει δυαδική τιμή, βλέπουμε ότι σχηματίζεται μόνο μία ομάδα με 247 στοιχεία τα οποία παίρνουν την τιμή 0, και υπολείπονται 104 στοιχεία που δεν ανήκουν σε καμία ομάδα.

Σε περίπτωση που εκτελέσουμε τον αλγόριθμο στο 66% των παρατηρήσεων μας, παρατηρούμε πως σχηματίζονται δύο ομάδες εκ των οποίων η πρώτη περιέχει 6 στοιχεία που παίρνουν μηδενική τιμή και 74 στοιχεία που παίρνουν την τιμή 1. 40 στοιχεία μένουν χωρίς να ομαδοποιηθούν. Δηλαδή από τις 231 παρατηρήσεις μόνο οι 120 επεξεργάζονται από τον αλγόριθμο. Οι υπόλοιπες προκαλούν θόρυβο και είναι δύσκολο να εξηγηθούν.

Εάν εφαρμόσουμε τον αλγόριθμο στο 50% των παρατηρήσεων μας δημιουργούνται 3 ομάδες. Η πρώτη περιέχει 7 παρατηρήσεις οι οποίες λαμβάνουν την τιμή 0, η δεύτερη περιλαμβάνει 100 στοιχεία που παίρνουν την τιμή 1 και η τρίτη 6 στοιχεία που παίρνουν την τιμή 2. Οι 63 μένουν χωρίς να ομαδοποιηθούν.

Εφαρμογή k-Means:

Κατά την πρώτη εφαρμογή του αλγορίθμου σε ποσοστό 50%, ομαδοποιούνται οι 176 παρατηρήσεις σε 2 ομάδες. Η πρώτη περιέχει 97 στοιχεία με μηδενική τιμή και η δεύτερη 79 τα οποία λαμβάνουν την τιμή 1.

Στη δεύτερη εφαρμογή σε όλο το σετ δεδομένων ο αλγόριθμος πάλι τις ομαδοποιεί όλες σε δύο ομάδες, η πρώτη περιέχει τις 194 οι οποίες παίρνουν την τιμή 0 και η δεύτερη τις 157 με τιμή 1.

Έπειτα παρατηρούμε πως αφαιρώντας μεταβλητές τα στοιχεία συνεχίζουν να μοιράζονται σε δύο ομάδες εξίσου αρμονικά όπως στην προηγούμενη εκτέλεση χωρίς να δημιουργούν θόρυβο ή να αφήνουν στοιχεία τα οποία δεν ανήκουν σε κάποια ομάδα.

- Iris

Αυτή είναι ίσως η πιο διαδεδομένη βάση δεδομένων ώστε να εφαρμοστούν οι αλγόριθμοι εξόρυξης δεδομένων που εφαρμόζουμε.

Δόθηκε από τον R.A. Fisher και το σετ δεδομένων περιέχει 3 τάξεις των 50 στοιχείων, όπου κάθε κλάση αναφέρεται στον τύπο του φυτού Iris:

Iris Setosa, Iris Versicolour, Iris Virginica.

Οι μεταβλητές είναι 5, συμπεριλαμβανομένου και της μεταβλητής με τις τάξεις. Η πρώτη αναφέρεται στο μήκος του κορμού του φυτού, η δεύτερη στο πλάτος του. Η Τρίτη και η τέταρτη μεταβλητή αναφέρονται στο μήκος και το πλάτος των φύλλων αντίστοιχα.

Εφαρμογή DBSCAN:

Κατά την πρώτη εκτέλεση του αλγορίθμου παρατηρούμε πως δημιουργούνται 3 ομάδες η κάθε μία από αυτές περιέχει 50 στοιχεία και παίρνουν τιμές: 0,1,2 αντίστοιχα.

Αφαιρώντας κάθε φορά μία μεταβλητή παρατηρούμε πως όσες και να αφαιρεθούν το αποτέλεσμα παραμένει το ίσιο με προηγουμένως.

Άρα συμπεραίνουμε πως ο αλγόριθμος ομαδοποιεί άριστα τις παρατηρήσεις μας.

Εφαρμογή k-Means:

Κατά την πρώτη εφαρμογή του k-Means σε όλο το σετ δεδομένων βλέπουμε πως δημιουργούνται 2 ομάδες, εκ των οποίων η πρώτη περιέχει τα 100 στοιχεία και παίρνει μηδενική τιμή και η δεύτερη τα υπόλοιπα 50 με τιμή 1.

Πάλι συμβαίνει όσες μεταβλητές και αν αφαιρέσουμε να προκύπτουν τα ίδια αποτελέσματα.

- Wine

Στο παρόν δείγμα αναφερόμαστε σε μία έρευνα χημικής ανάλυσης κρασιού που πραγματοποιήθηκε σε μία περιοχή της Ιταλίας.

Τα δείγμα αποτελείται από δεδομένα τα οποία είναι αποτελέσματα μίας χημικής ανάλυσης κρασιών των οποίων τα αμπέλια καλλιεργήθηκαν στην ίδια περιοχή της Ιταλίας και τα οποία αποτελούνται από τρεις διαφορετικές ποικιλίες αμπελιού.

Η χημική ανάλυση προσδιόρισε 13 συστατικά που εμπεριέχονται σε κάθε τύπο ποικιλίας των 178 φιαλών κρασιού που εξετάσαμε.

Οι 13 συνεχείς μεταβλητές είναι:

- Κλάση
- Αλκοόλη
- Μηλικό Οξύ
- Τέφρα
- Τέφρα Αλκοόλης
- Μαγνήσιο
- Συνολική Φαινόλη
- Flavanoids
- Non-flavanoid φαινόλες
- Προ ανθοκυανόλη
- Απόχρωση
- Έντονο χρώμα
- Αραίωμα οίνου
- Προλίνη

Το δείγμα μας λοιπόν αποτελείται από 3 τάξεις όπως αυτές φαίνονται παρακάτω:

Κλάση	Πληθυσμός
1	59

2	71
3	48

Τα επίπεδα της αλκοόλης καθορίζονται ως εξής:

Στατιστικές	Τιμή
Minimum	11,03
Maximum	14,83
Μέσος	13,001
Τυπική Απόκλιση	0,812

Τα επίπεδα του μηλικού οξέος είναι:

Στατιστικές	Τιμή
Minimum	0,74
Maximum	5,8
Μέσος	2,336
Τυπική Απόκλιση	1,117

Τα επίπεδα τέφρας όπως φαίνονται παρακάτω:

Στατιστικές	Τιμή

Minimum	1,36
Maximum	3,23
Μέσος	2,367
Τυπική Απόκλιση	0,274

Τα επίπεδα της τέφρας της αλκοόλης:

Στατιστικές	Τιμή
Minimum	10,6
Maximum	30
Μέσος	19,495
Τυπική Απόκλιση	3,34

Τα επίπεδα του μαγνησίου είναι τα εξής:

Στατιστικές	Τιμή
Minimum	70
Maximum	162
Μέσος	99,742
Τυπική Απόκλιση	14,282

Τα επίπεδα της συνολικής φαινόλης που περιέχονται σε κάθε μπουκάλι είναι:

Στατιστικές	Τιμή
Minimum	0,98
Maximum	3,88
Μέσος	2,295
Τυπική Απόκλιση	0,626

Τα επίπεδα flavanoids είναι τα εξής:

Στατιστικές	Τιμή
Minimum	0,34
Maximum	5,08
Μέσος	2,029
Τυπική Απόκλιση	0,999

Τα επίπεδα non-flavanoids είναι:

Στατιστικές	Τιμή
Minimum	0,13
Maximum	0,66
Μέσος	0,362

Τυπική Απόκλιση	0,124

Τα επίπεδα προανθοκυανόλης είναι τα εξής:

Στατιστικές	Τιμή
Minimum	0,41
Maximum	3,58
Μέσος	1,591
Τυπική Απόκλιση	0,572

Τα επίπεδα της έντασης χρώματος του οίνου είναι τα εξής:

Στατιστικές	Τιμή
Minimum	1,28
Maximum	13
Μέσος	5,058
Τυπική Απόκλιση	2,318

Τα επίπεδα απόχρωσης του κάθε οίνου είναι:

Στατιστικές	Τιμή
Minimum	0,48

Maximum	1,71
Μέσος	0,957
Τυπική Απόκλιση	0,229

Τα επίπεδα αραίωσης κάθε φιάλης οίνου είναι:

Στατιστικές	Τιμή
Minimum	1,27
Maximum	4
Μέσος	2,612
Τυπική Απόκλιση	0,71

Τα επίπεδα προλίνης κάθε φιάλης:

Στατιστικές	Τιμή
Minimum	278
Maximum	1680
Μέσος	746,893
Τυπική Απόκλιση	314,907

Εφαρμογή DBSCAN:

Κατά την πρώτη εκτέλεση του αλγορίθμου για ομαδοποίηση των στοιχείων του σετ δεδομένων παρατηρούμε πως δημιουργούνται 3 ομάδες ενώ μόνο ένα στοιχείο παραμένει εκτός κάποιας ομάδας:

Ομάδες	Τιμή Ομάδας	Πληθυσμός
1	0	59
2	1	70
3	2	48

Ενώ αν κατά τις επόμενες εφαρμογές του αλγορίθμου επιδιώξουμε να αφαιρέσουμε μία ή περισσότερες μεταβλητές, το αποτέλεσμα που παίρνουμε είναι παντού το ίδιο: όλα μας τα στοιχεία μένουν ανομαδοποίητα.

Αν τώρα εφαρμόσουμε τον αλγόριθμο σε ποσοστό 66% των στοιχείων παρατηρούμε πως ομαδοποιούνται τα 61 από τα 117 στις εξής ομάδες:

Ομάδες	Πληθυσμός
0	13
1	24
2	24

Αν όμως εφαρμόσουμε τον DBSCAN στο 50% των στοιχείων μας, δηλαδή στα 89, ομαδοποιούνται όλα εκτός από ένα στοιχείο που μένει εκτός ομάδας:

Ομάδες	Πληθυσμός

0	19
1	36
2	33

Εφαρμογή k-Means:

Εφαρμόζοντας τον αλγόριθμο αυτόν παρατηρούμε ότι κατά την πρώτη εκτέλεση του παίρνουμε τα εξής αποτελέσματα:

Ομάδες	**Τιμή Ομάδων**	**Πληθυσμός**
1	0	59
2	1	119

Κατά τη δεύτερη εφαρμογή του αφαιρούμε την πρώτη μεταβλητή που αφορά την τάξη των οίνων και παρατηρούμε ότι συμβαίνουν τα εξής:

0	108
1	70

Όσο προχωράμε στην αφαίρεση μεταβλητών παρατηρούμε πως εξακολουθούν να υπάρχουν οι ίδιες ομάδες με τη διαφορά πως αφαιρώντας ανά δύο τις μεταβλητές αυξάνεται κατά 2 ο αριθμός της πρώτης ομάδας ενώ ταυτόχρονα μειώνεται κατά 2 ο αριθμός των στοιχείων της δεύτερης.

Γενικότερα παρατηρούμε πως και οι 2 αλγόριθμοι ομαδοποιούν τα δεδομένα μας ορθά.

4.1.2 **Κατηγορικά Δεδομένα**

- **§** Balance-Scale

Στο παρόν σετ δεδομένων αναφερόμαστε σε 625 δεδομένα τα οποία είναι αποτελέσματα μοντελοποίησης ενός ψυχολογικού πειράματος.

Κάθε παράδειγμα κατηγοριοποιείται σαν να πρόκειται για μία ζυγαριά η οποία ή γέρνει στα δεξιά ή στα αριστερά είτε ισορροπεί.

Οι μεταβλητές είναι αριστερό βάρος, αριστερή απόσταση, δεξί βάρος, δεξιά απόσταση.

Ο σωστός τρόπος ώστε να βρεθεί η κατάλληλη ομάδα είναι η εξής εξίσωση:

Αριστερό βάρος * αριστερή απόσταση

Δεξί βάρος * δεξιά απόσταση

Σε περίπτωση που αυτά είναι ίσα τότε υπάρχει ισορροπία.

Οι πληροφορίες για τις μεταβλητές μας είναι οι εξής:

Όνομα τάξης: 3 (L,B,R)

Αριστερό βάρος: 5 (1,2,3,4,5)

Αριστερή απόσταση: 5 (1,2,3,4,5)

Δεξί βάρος: 5 (1,2,3,4,5)

Δεξιά απόσταση: 5 (1,2,3,4,5)

Οι παραπάνω πληροφορίες προέκυψαν από τον Siegler, R. S. (1976).

Το σε δεδομένων μας απεικονίζεται στην πλατφόρμα weka ως εξής:

Κλάση	Πληθυσμός
Left	288
Balance	49
Right	288

Για το πώς κατανέμονται οι παρατηρήσεις σύμφωνα με κάθε μέτρηση ισχύουν τα παρακάτω για όλες:

Στατιστικές	Τιμή
Minimum	1
Maximum	5
Μέσος	3
Τυπική Απόκλιση	1415

Εφαρμογή k-modes

Κατά την πρώτη ομαδοποίηση, η οποία βασίστηκε σε τυχαία επιλογή των αρχικών κέντρων. Δημιουργούνται 3 ομάδες στις οποίες αναθέτονται δεδομένα ως έχει:

Ομάδες	Πληθυσμός
1	352

2	177
3	96

Στη συνέχεια ωστόσο, και μετά από δύο επαναλήψεις ο αλγόριθμος κατέληξε στην ίδια ομαδοποίηση:

Ομάδες	Πληθυσμός
1	352
2	177
3	96

Έπειτα από 5 συνεχόμενες επαναλήψεις της αρχικής ομαδοποίησης, πάντα με τυχαία επιλογή των αρχικών κέντρων, ο αλγόριθμος καταλήγει πάλι στα ίδια αποτελέσματα:

Ομάδες	Πληθυσμός
1	352
2	177
3	96

- Car evaluation

Σε αυτό το σετ δεδομένων αναφερόμαστε σε μία βάση δεδομένων που περιγράφει χαρακτηριστικά διαφόρων αυτοκινήτων ώστε να συμβάλλει στη λήψη αποφάσεων των πελατών που προτίθενται να τα αγοράσουν.

Το δείγμα μας αποτελείται από 1728 δεδομένα, καθένα από τα οποία περιλαμβάνει:

- Αυτοκίνητο: αποδοχή αυτοκινήτου
- Τιμή: ολόκληρη τιμή, αγοράς, συντήρησης
- Τεχνολογία: τεχνολογικά χαρακτηριστικά
- Άνεση: άνεση, αριθμός θυρών, χωρητικότητα ατόμων, μέγεθος χωρητικότητας αποσκευών
- Ασφάλεια: ασφάλεια αυτοκινήτου

Οι μεταβλητές μας λοιπόν είναι οι εξής:

- Αγορά (πολύ υψηλή, υψηλή, μέτρια, χαμηλή)
- Τεχνολογία (πολύ υψηλή, υψηλή, μέτρια, χαμηλή)
- Θύρες (2,3,4,5 και περισσότερες)
- Άτομα (2,4 και περισσότερα)
- Χωρητικότητα αποσκευών (μικρή, μέτρια, μεγάλη)
- Ασφάλεια (χαμηλή, μέτρια, υψηλή)

Και οι κλάσεις είναι:

- μη-αποδεκτό αυτοκίνητο
- αποδεκτό

- καλό
- πολύ καλό

Οπότε σύμφωνα με τα παραπάνω, όσον αφορά την μεταβλητή της αγοράς και της τεχνολογίας αντίστοιχα, τα στοιχεία κατανέμονται ως εξής:

Κλάση	Πληθυσμός
Πολύ υψηλή	432
Υψηλή	432
Μέτρια	432
Χαμηλή	432

Όσον αφορά τη μεταβλητή των θυρών του οχήματος:

Κλάση	Πληθυσμός
2	432
3	432
4	432
5 και πάνω	432

Όσον αφορά τη χωρητικότητα ατόμων:

Κλάση	Πληθυσμός

2	576
4	576
περισσότερα	576

Όσον αφορά τη χωρητικότητα αποσκευών:

Κλάση	Πληθυσμός
Μικρές	576
Μέτριες	576
Μεγάλες	576

Εφαρμογή k-Modes:

Κατά την πρώτη ομαδοποίηση, η οποία βασίστηκε σε τυχαία επιλογή αρχικών κέντρων, ανατέθηκαν οι παρακάτω αριθμοί αντικειμένων ως εξής:

Ομάδες	Πληθυσμός
1	445
2	447
3	409

4	427

Ωστόσο, μετά από 8 επαναλήψεις, τα αποτελέσματα στα οποία καταλήγει ο αλγόριθμος είναι:

Ομάδες	Πληθυσμός
1	431
2	354
3	473
4	470

Κατά τη δεύτερη ομαδοποίηση που εκτέλεσε ο αλγόριθμός μας, πάλι με τυχαία επιλογή κέντρων, ανατέθηκαν τα παρακάτω αποτελέσματα:

Ομάδες	Πληθυσμός
1	480
2	469
3	343
4	436

Μετά από 10 επαναλήψεις καταλήγουμε στα εξής αποτελέσματα:

Ομάδες	Πληθυσμός

1	364
2	568
3	382
4	414

Κατά την τρίτη ομαδοποίηση του k-modes ανατέθηκαν τα παρακάτω αντικείμενα:

Ομάδες	Πληθυσμός
1	407
2	488
3	398
4	435

Έπειτα από 23 επαναλήψεις τα αποτελέσματα έχουν ως εξής:

1	387
2	374
3	379
4	588

Κατά την τέταρτη ομαδοποίηση του αλγορίθμου ανατέθηκαν τα παρακάτω νούμερα αντικειμένων:

Ομάδες	Πληθυσμός
1	312
2	354
3	666
4	1581

Ύστερα από 17 επαναλήψεις αυτή τη φορά τα τελικά νούμερα αντικειμένων που ανατίθενται στις ομάδες είναι:

Ομάδες	Πληθυσμός
1	423
2	385
3	496
4	424

Τέλος, κατά την πέμπτη ομαδοποίηση όπου πάντα βάση τυχαία επιλεγμένων κέντρων, αναθέτονται στις ομάδες τα αντικείμενα ως εξής:

Ομάδες	Πληθυσμός
1	555
2	433

3	340
4	400

Μετά από 23 επαναλήψεις έχουμε τα εξής νούμερα αντικειμένων στις ομάδες:

1	455
2	456
3	401
4	416

§ Tic Tac Toe

Η συγκεκριμένη βάση δεδομένων αφορά τα παιχνίδια ης μορφής tic tac toe όπου κωδικοποιούνται τα σύνολα των πιθανών συνθέσεων στο τέλος κάθε παρτίδας παιχνιδιού. Η έννοια-στόχος είναι «νίκη για χ» όπου ο «χ» θα έχει παίξει πρώτος.

Ουσιαστικά πρόκειται για τα παιχνίδια της μορφής της τρίλιζας τα οποία παίζονται με δύο παίχτες, καθένας από τους οποίους επιλέγει σε ποιο «τετραγωνάκι» θα παίξει κάθε φορά.

Η συγκεκριμένη βάση δεδομένων αφορά 558 δεδομένα υπό εξέταση και κατασκευάστηκε από τον David W. Aha.

Οι μεταβλητές μας έχουν ως εξής:

i. Επάνω αριστερό τετράγωνο: {x,o,b}

ii. Επάνω μεσαίο τετράγωνο: {x,o,b}

iii. Επάνω δεξί τετράγωνο: {x,o,b}

iv. Μεσαίο αριστερό τετράγωνο: {x,o,b}

v. Μεσαίο κεντρικό τετράγωνο: {x,o,b}

vi. Μεσαίο δεξί τετράγωνο: {x,o,b}

vii. Κάτω αριστερό τετράγωνο: {x,o,b}

viii. Κάτω μεσαίο τετράγωνο; {x,o,b}

ix. Κάτω δεξί τετράγωνο: {x,o,b}

x. Τάξη: {θετική, αρνητική}

Εφαρμογή k-modes:

Στην πρώτη ομαδοποίηση, η οποία βασίστηκε σε τυχαία επιλογή αρχικών κέντρων, ανατέθηκαν οι παρακάτω αριθμοί αντικειμένων στις εξής ομάδες:

Ομάδες	Πληθυσμός
1	303
2	655

Στη συνέχεια ωστόσο και μετά από 9 επαναλήψεις καταλήγουμε στην παρακάτω ομαδοποίηση:

Ομάδες	Πληθυσμός
1	335
2	623

Έπειτα, κατά την πρώτη επανάληψη της εφαρμογής της αρχικής ομαδοποίησης παίρνουμε τα εξής αποτελέσματα:

Ομάδες	Πληθυσμός
1	303
2	655

Μετά από πάλι 9 επαναλήψεις καταλήγει ο αλγόριθμος στα ίδια αποτελέσματα όπως παραπάνω:

Ομάδες	Πληθυσμός
1	335
2	623

Κατά τη δεύτερη επανάληψη της εφαρμογής ο k-modes βασισμένος πάντα σε τυχαία επιλογή των αρχικών κέντρων μας δίνει τα εξής αποτελέσματα:

Ομάδες	Πληθυσμός
1	307
2	651

Μετά από 6 επαναλήψεις ο αλγόριθμος καταλήγει στα εξής αποτελέσματα:

Ομάδες	Πληθυσμός
1	330

2	628

Κατά το τρίτο τρέξιμο του k-modes ο αλγόριθμος καταλήγει στην ανάθεση των εξής αριθμών των αντικειμένων:

Ομάδες	Πληθυσμός
1	307
2	651

Έπειτα από 6 επαναλήψεις πάλι καταλήγουμε στα εξής:

Ομάδες	Πληθυσμός
1	330
2	628

Τέλος κατά την πέμπτη επανάληψη της αρχικής ομαδοποίησης παίρνουμε ακριβώς τα ίδια αποτελέσματα με παραπάνω:

Ομάδες	Πληθυσμός
1	307
2	651

Και πάλι μετά από 6 επαναλήψεις καταλήγουμε στα τελικά αποτελέσματα:

Ομάδες	Πληθυσμός

1	330
2	628

4.2 ΣΥΜΠΕΡΑΣΜΑΤΑ

Στο σημείο αυτό ολοκληρώνεται η έρευνά μας στο θέμα της εξόρυξης γνώσης από βάσεις δεδομένων με τη χρήση των αλγορίθμων ομαδοποίησης.

Σε όλη μας την εργασία εξετάσαμε λεπτομερώς τη διαδικασία ανακάλυψης γνώσης (KDD) όπως επίσης και τις διάφορες μεθόδους εξόρυξης δεδομένων.

Αναφερθήκαμε σε μία πληθώρα αλγορίθμων που αφορούν τη μέθοδο της ομαδοποίησης δεδομένων εφαρμόζοντας τους k-Modes, k-Means και DBSCAN αλγορίθμους σε συγκεκριμένες βάσεις δεδομένων ώστε να εξετάσουμε την εφαρμογή όλων όσων είχαμε αναφέρει περί του τρόπου εξόρυξης γνώσης από βάσεις δεδομένων.

Χρησιμοποιήσαμε βάσεις δεδομένων τις οποίες δανειστήκαμε από την ιστοσελίδα του πανεπιστημίου της Καλιφόρνια: www.ics.uci.edu/ προκειμένου να εφαρμόσουμε τους παραπάνω αλγορίθμους ώστε να είμαστε σε θέση να εξάγουμε χρήσιμα συμπεράσματα και διαπιστώσεις.

Όσον αφορά τους αλγορίθμους που εφαρμόσαμε στα αριθμητικά δεδομένα μας, δηλαδή τους DBSCAN και Simple k-Means, προβήκαμε στο εξής συμπέρασμα:

Ο DBSCAN σαν αλγόριθμος βασισμένος στην πυκνότητα εφαρμόζεται απευθείας στο σύνολο των δεδομένων προς ανάλυση και έτσι καθίσταται ασύμφορος για επεξεργασία των στοιχείων μεγάλων βάσεων δεδομένων.

Η βασική ιδέα του αλγορίθμου είναι ότι έχουμε ένα σύνολο από στοιχεία (D). Η περιοχή που εκτείνεται σε καθορισμένη πάντα ακτίνα γύρω από τη «γειτονιά» κάθε στοιχείου κάθε ομάδας θα πρέπει να περιέχει έναν ελάχιστο αριθμό στοιχείων.

Σε περίπτωση που ανακαλύψει ομάδες με ακανόνιστα σχήματα αντιμετωπίζει προβλήματα διότι επηρεάζεται από την τιμή της ακτίνας και του ελάχιστου αριθμού στοιχείων τα οποία δύσκολα προσδιορίζονται.

Όντας γνώστες ,πλέον, των παραπάνω καταλήξαμε στο ότι η εφαρμογή του αλγορίθμου αυτού ήταν περίπλοκη και χρονοβόρα σε αντίθεση με εκείνη του k-Means αλγόριθμου.

Ο Simple k-Means από την άλλη πλευρά ακολουθεί μία απλή επαναληπτική μέθοδο ώστε να χωρίσει ένα σύνολο δεδομένων σε ένα καθορισμένο αριθμό k ομάδων επιλέγοντας k σημεία από το d σύνολο διανυσμάτων. Τα σημεία αυτά ορίζονται ως αρχικοί αντιπρόσωποι των k ομάδων, προκειμένου να ελαχιστοποιηθεί η παρακάτω αντικειμενική συνάρτηση που ορίζεται ως η μέση τετραγωνική απόσταση των σημείων από τα πλησιέστερα κέντρα (k) των ομάδων.

Παρατηρώντας την εφαρμογή του K-Means στα δεδομένα μας καταλήξαμε στο ότι δεν είναι τόσο χρονοβόρος όσο ο DBSCAN και είναι περισσότερο απλοϊκός στην εκτέλεσή του αφού αποτελεί μία καθιερωμένη πλέον μέθοδο ομαδοποίησης με αποτελεσματικές τεχνικές για διάφορα πεδία ορισμού των , κάθε φορά , υπό ανάλυση δεδομένων.

Όσον αφορά τον k-Modes, τον οποίο εφαρμόσαμε με τη βοήθεια του εισηγητή μας Νικόλαου Μαστρογιάννη, συμπεράναμε πως η εφαρμογή αυτού του αλγορίθμου στα κατηγορικά μας δεδομένα ήταν η πιο απλή.

Ο k-Modes λαμβάνοντας υπ’ όψη μόνο δεδομένα που αφορούν λεκτικές τιμές έχει σαν βάση τον k-means με κάποιες τροποποιήσεις. Χρησιμοποιεί διαφορετικά μέτρα ανομοιότητας έτσι ώστε να μπορούν να εφαρμοστούν στις λεκτικές τιμές. Αυτός αντικαθιστά τα k-κέντρα με τα k-modes και χρησιμοποιεί μεθόδους που βασίζονται στη συχνότητα εμφάνισης των τιμών προκειμένου να ενημερώνονται τα κέντρα των ομάδων, δηλαδή τα k-modes.

Από την εφαρμογή του παρατηρήσαμε πως στην πρώτη ομαδοποίηση που εκτελεί βασίζεται σε τυχαία επιλογή αρχικών κέντρων, όπου αναθέτει συγκεκριμένους αριθμούς αντικειμένων στις ομάδες που δημιουργούνται και στη συνέχεια εφαρμόζει πολλές επαναλήψεις στην

αρχική αυτή εφαρμογή ώστε να καταλήξει στην πιο «ταιριαστή» ομαδοποίηση, εκείνη δηλαδή που ομαδοποιεί καλύτερα τα δεδομένα υπό εξέταση.

Συνοψίζοντας, καταλήγουμε στο ότι η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων και η εξόρυξη δεδομένων από αυτές είναι μία επιστήμη η οποία είναι αρκετά χρήσιμη σε πολλούς τομείς της καθημερινότητάς μας, διότι στον απαιτητικό κόσμο της πληροφορίας μόνο όταν εξετάζονται τα στοιχεία από πολλές διαφορετικές απόψεις μπορούν να γίνουν ενδιαφέρουσες ανακαλύψεις.

Ειδικότερα, μας δίνεται η δυνατότητα να ταξινομήσουμε, να ομαδοποιήσουμε και να συσχετίσουμε τον τεράστιο αυτό όγκο της γνώσης που υπάρχει στις βάσεις δεδομένων ώστε να είμαστε σε θέση να καταλήξουμε σε κάποια ενδεχομένως χρήσιμα συμπεράσματα στον τομέα που κάθε φορά ερευνούμε.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] S.M. Weiss(1997), Predictive Data Mining: a Practical Guide, Morgan Kaufmann Pub.

[2] Christofer Westphal. Teresa Blaxton(1998):Data mining methods and tools for solving real world problems, Bookstore].

[3] Sang Jung Lee, Keng Siau "A review of Data Mining techniques", Industrial Management and Data Systems (101/1[2001],pp.41-46).

[4] Frawley., Piatetsky-Shapiro,G.Smyth,P., and Matheus 1996, "Advances in Knowledge Discovery and Data Mining", AAAI Press/The MIT Press.

[5] Holshermier, Siebes 1997.

[6] Μ.Βαζιργιάννης, Μ.Χαλκιδη: «Εξορυξη Δεδομένων, 2003

[7] M.H Dunham, Data Mining: «Εισαγωγικά και προηγμένα θέματα εξόρυξης γνώσης από δεδομένα» (επιμέλεια ελληνικής έκδοσης: Β.Βερύκιος,Γ.Θεοδωρίδης Εκδόσεις Νέων Τεχνολογιών ,2004

[8] en.wikipedia.org/wiki/**Data_mining**.

[9] http://www.cs.uoi.gr/pitoura/courses/dm, Θεώνη Πιτουρά διδάσκων Πανεπιστημίου.

[10] W.H.Inmon, Building the data warehouse,1998.

[11] Σημειώσεις μαθήματος καθηγητή Β.Βουτσινά «Συστήματα Υποστήριξης Αποφάσεων 2, (κεφάλαιο Ανεύρεσης Γνώσης)

[12] tsirakis@ceid.upatras.gr

[13] http://gtziralis.googlepages.com

[14]LogisticsCourse LectureOnForecasting.Tziralis/

[15] http://courses.dbnet.ntua.gr

[16] Γεώργιος Παλιούρας, Εξόρυξη γνώσης από δεδομένα 1992.

[17] Fayyad,G.Piatetsky-Shapiro,P.Smyth and R.Uthurusamy: "Advances in Knowledge Discovery and Data Mining" Workshops on Knowledge in Databases, 1991-1994

[18]"Classification" (Tutorial) Department of Informatics, Athens University of Economics & Business

(www.dbnet.aueb.gr/courses/datamining/slides_class.zip.)

[19] Ι.Βλαχαβάς, Π.Κεφαλάς: «Τεχνητή Νοημοσύνη», εκδόσεις Θεσσαλονίκη

[20] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques"

[21] Melvin F. Janowitz, "A combinational Introduction to Cluster Analysis"

[22] Top ten algorithms in data mining, Springer-Verlag London Limited

[23] www.icsd.aegean.gr/lectures/kavalieratou.pdf

[24] David Heckerman, 1997

[25] Alireza Osarech, Majid Mirmehdi, Barry Tomas and Ritchard Markham "Automatic Recognition of Exudative Maculopathy using Fuzzy C-Means Clustering and Neural Networks", Department of Computer Science, University of Bristol

[26] http://mmlab.ceid.upatras.gr/

[27] Νέοι αλγόριθμοι υπολογιστικής νοημοσύνης και ομαδοποίησης για την εξόρυξη πληροφορίας, Τασουλής Δημήτρης,2007

[28] (http://www.ted.unipi.gr/Uploads/Files/Material/Courses/15_1105659031.pdf

Επιπρόσθετη Βιβλιογραφία που μελετήσαμε:

[29] Σημειώσεις μαθήματος καθηγητή Β.Βουτσινά «Συστήματα Υποστήριξης Αποφάσεων 2, (κεφάλαιο Ανεύρεσης Γνώσης)

[30]Melvin F. Janowitz, “A combinational Introduction to Cluster Analysis”

[31] T.Mitchell “Decision Tree Learning” in T.Mitchell, Machine Learning The Mc Graw-Hill Companies, Inc1997,pp.52-78

[32] Περιοδικό RAM «Το ένθετο πρόβλημα της ανάλυσης – E-Commerce»

[33]W.A.Wallace ,F.Ozden Gur Ali, “Bringing the gap between business objectives and parameters of data mining algorithms”, Decision Support Systems (1997)

[34] Βασίλης Μ.Παπαδάκης «Στρατηγική των επιχειρήσεων: Ελληνική και διεθνής Εμπειρία, Αθήνα 2002, εκδόσεις Μπένου

[35] Johannes Grabmeier, Andreas Rudolph “Techniques of Cluster algorithms in Data Mining, kdd,6,303-360,2002

[36]Frank Block ,Phd. “Data mining the Insightful Way” ,Whitepaper

[37] SAS White paper “Data Mining in the insurance industry: Solving business problems using SAS Enterprise Miner Software

[38] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York

[39] http://dtps.unipi.gr/files/notes/2007-2008/eksamino_5/apothhkes_kai_eksoryksh_dedomenwn/lecture7_clustering28229.pdf ΟΜΑΔΟΠΟΙΗΣΗ ΥΠΟΧΩΡΩΝ

[40] http://infolab.cs.unipi.gr/courses/dwdm/slides/9-spatial.pdf

[41] Shu-Hsien Liao “Knowledge management technologies and applications-literature review from 1995-2002”, Expert Systems with Applications 25 (2003)pp.155-164

[42] www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/**datamining**.htm - 13k

[43] http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/fayyad96.pdf

[44] Προχωρημένα θέματα βάσεων δεδομένων- Αποθήκες δεδομένων, http://courses.dbnet.ntua.gr

[45] http://csie.org/~dm/clustering.1.1107.ppt#280,41,Birch: A Hybrid Clustering Algorithm

[46] Δ.Μπουραντάς ,Ν.Παπαλεξανδρή «Εισαγωγή στη ιοίκηση Επιχειρήσεων» ενότητα 6,8 εκδόσεις Μπένου Αθήνα 1998

[47] Jing Luan “Data Mining Applications in higher education” (SPSS)

[48] 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD) 1995,19998

[49] Journal of Data Mining and Knowledge Discovery, 1997

[50] Γ. Νικηφορίδης «Στοιχεία ιατρικής Πληροφορικής»

[51] R. Brause, T. Langsdorf, M. Hepp, “Neural Data Mining for credit card fraud detection”

[52] Gialunca Bontempi, “Data Mining for prediction”

[53] IDC & Cap Gemini: “Four elements of customer relationship Management

[54] Kurt Thearling “An introduction to Data Mining”

www.thearling.com

[55] Jonatan Shapiro “Genetic Algorithms in Machine Learning”

[56] Διαδικτυακός τόπος: Department of computer science, university of Manchester

[57] Jiawei Han, “Data Mining: An overview from a database perspective”, PAKDD Conference 1998

[58] Ανδρέας Α.Κιντής «Στατιστικές και Οικονομετρικές Μέθοδοι» εκδόσεις Gutenberg ,Αθήνα 1999