# ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**ΑΛΓΟΡΙΘΜΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ-ΤΑΞΙΝΟΜΗΣΗΣ**

*Μουζάκη Δήμητρα Α.Μ. 9011*

*Κοσμάς Νικόλαος*

**ΕΦΑΡΜΟΓΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ**

**ΔΙΟΙΚΗΣΗ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ**

# ΠΕΡΙΕΧΟΜΕΝΑ

## 1. ΑΝΕΥΡΕΣΗ ΓΝΩΣΗΣ



## 2.ΤΑΞΙΝΟΜΗΣΗ



## 3 ΟΜΑΔΟΠΟΙΗΣΗ

# 1.ΑΝΕΥΡΕΣΗ ΓΝΩΣΗΣ

## 1.1 Γενική Αναφορά στην ανεύρεση γνώσης

Η ανεύρεση γνώσης(knowledge discovery), και η άμεσα συνεπακόλουθη τεχνική εξόρυξη από δεδομένα(data mining), αποτελεί την πλέον σύγχρονη θεμελιώδη τεχνική των Συστημάτων Υποστήριξης Αποφάσεων. Αφορά στην αυτόματη εξαγωγή γνώσης, κυρίως σε μορφή κανόνων/συμβουλών, από δεδομένα. Επιβλήθηκε από τη μεγάλη ανάπτυξη των εμπορικών και επιστημονικών βάσεων δεδομένων και την ανάγκη ανάλυσης του τεράστιου όγκου δεδομένων που αποθηκεύουν . Έχει υπολογιστεί ότι από το 1990 μέχρι το 1998 έχει εξαπλασιαστεί ο όγκος των αποθηκευμένων παγκοσμίως δεδομένων στα μεγάλα συστήματα Η/Υ(mainframes). Είναι χαρακτηριστική η ρήση του συγγραφέα και μελλοντολόγου John Naisbitt:<<Πνιγόμαστε στις πληροφορίες αλλά διψάμε για γνώση>> (<<we are drowning in information, but starving for knowledge>>).Η ανεύρεση γνώσης βασίζεται στη δυνατότητα ανάλυσης μεγάλου πλήθους δεδομένων, η οποία ανάλυση, αν δεν αυτοματοποιηθεί με την χρήση Η/Υ, είναι πρακτικά αδύνατη . Τα δεδομένα προς ανάλυση μπορεί να είναι από τα οικονομικά δεδομένα μίας επιχείρησης μέχρι εικόνες από δορυφόρους και μουσικές παρτιτούρες, ανάλογα με την εφαρμογή.

## 1.2 Ποιά τα προβλήματα που έχει εφαρμοστεί επιτυχημένα η ανεύρεση γνώσης

Η ανεύρεση γνώσης έχει εφαρμοστεί επιτυχημένα σε πλήθος προβλημάτων, τα πιο συνήθη από τα οποία είναι τα εξής :

Προώθηση προιόντων: π.χ. διαχείρηση και έρευνα της πελατειακής βάσης μίας επιχείρησης .

Λιανεμπόριο :π.χ. αναγνώριση των προτημήσεων των πελατών .

Οικονομικά :π.χ. αναγνώριση της επικινδυνότητας (risk analysis).

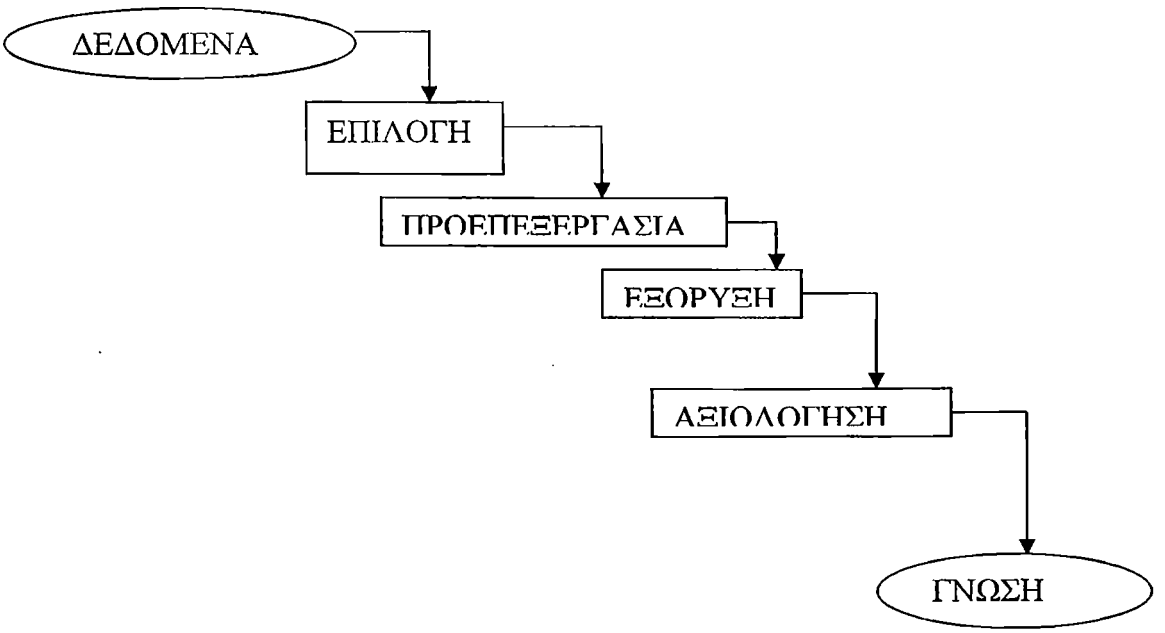
Βιομηχανία : π.χ. επεξεργασία των προδιαγραφών των ανταλλακτικών .

Υγεία:π.χ. διάγνωση

Ενέργεια : π.χ. πρόβλεψη των απαιτήσεων για κατανάλωση ηλεκτρικής ενέργειας .

## 1.3 Τι είναι η ανεύρεση γνώσης και ποιά τα βηματά που την χαρακτηρίζουν

Η ανεύρεση γνώσης είναι μία ακολουθία από βήματα , πολλές φορές επαναλαμβανόμενη , όπου σε κάθε βήμα ο χρήστης πρέπει να καθορίσει παραμέτρους, ανάλογα με την εφαρμογή . Τα βήματα που την συνιστούν παρουσιάζονται στην Εικόνα 1.




### 1.3.1 Πρώτο Στάδιο:Επιλογή

Στο στάδιο αυτό επιλέγονται από το σύνολο των δεδομένων εκείνα τα χαρακτηριστικά που ενδιαφέρουν. Αν θεωρήσουμε την περίπτωση κατά την οποία τα δεδομένα έχουν οργανωθεί σε πίνακες(tables) μίας σχεσιακής βάσης δεδομένων (relational database) στο πρώτο στάδιο επιλέγονται οι πίνακες αλλά και τα πεδία (fields) προς ανάλυση που ενδιαφέρουν. Πολύ συχνά απαιτείται να σχεδιασθεί ένα αρχιτεκτόνημα δεδομένων για την αποθήκευση των επιλεγμένων δεδομένων . Ένα βασικό προτέρημα της ανεύρεσης γνώσης (σε σύγκριση με τις παραδοσιακές στατιστικές μεθόδους) είναι ότι δεν απαιτεί κάποια υπόθεση εργασίας την οποία θα μπορούσε να επιβεβαιώσει ή όχι, αλλά παρέχει αυτόματα, υπό μορφή κανόνων, ένα σύνολο επιβεβαιωμένων τέτοιων υποθέσεων εργασίας. Απ'αυτήν την άποψη, όσα περισσότερα χαρακτηριστικά επιλεγούν στο πρώτο στάδιο τόσο περισσότερο εξαντλιτικοί είναι οι κανόνες που εξάγονται, δηλαδή τόσο πιο πλούσια είναι η

γνώση που ανευρίσκεται. Από την άλλη μεριά όμως, η αύξηση αυτή έχει άμεση επίδραση στην αύξηση της πολυπλοκότητας του χρόνου της τεχνικής. Ανάλογα με τον αλγόριθμο εξόρυξης από δεδομένα που χρησιμοποιείται, η αύξηση της πολυπλοκότητας χρόνου μπορεί να είναι και απαγορευτική για την εφαρμογή της τεχνικής. Συνεπώς, ανάλογα με την περίπτωση εφαρμογής, ο χρήστης πρέπει να κάνει τους κατάλληλους συμβιβασμούς.

### 1.3.2 Δεύτερο Στάδιο: Προεπεξεργασία

Αφού επιλεγούν τα δεδομένα προς ανάλυση, πολύ συχνά απαιτείται μία προεπεξεργασία τους ώστε να ικανοποιούν τις απαιτήσεις για την εφαρμογή της τεχνικής και του αλγόριθμου εξόρυξης από δεδομένα που θα χρησιμοποιηθεί . Συχνά, αυτή η προεπεξεργασία, που γίνεται στο δεύτερο στάδιο, αφορά στην απομάκρυνση πιθανού θορύβου από τα δεδομένα και τη διαχείρηση των κενών τιμών στα πεδία . Η αντιμετώπιση του προβλήματος των κενών τιμών γίνεται με διάφορους τρόπους, από την απλή διαγραφή των εγγραφών(records) που περιέχουν πεδία με κενές τιμές μέχρι την αντικατάσταση των κενών τιμών με ειδική τιμή (π.χ. ΑΓΝΩΣΤΟ) ή με την πιο συχνά εμφανιζόμενη τιμή ή ακόμα με τιμές που προκύπτουν από μία πιθανοτική κατανομή, κ.α.Επίσης στο δεύτερο στάδιο γίνεται αν απαιτείται, ένας μετασχηματισμός των δεδομένων ώστε να έχουν την κατάλληλη μορφή για την εφαρμογή του αλγόριθμου εξόρυξης από δεδομένα, που θα χρησιμοποιηθεί. Ο μετασχηματισμός μπορεί να γίνεται για πολλούς διαφορετικούς λόγους. Για παράδειγμα μπορούν να χρησιμοποιηθούν μέθοδοι μείωσης των χαρακτηριστικών (πεδίων) που έχουν επιλεγεί (ελάττωση διαστάσεων), ώστε να μειωθεί η πολυπλοκότητα χρόνου, όπως έχει αναφερθεί παραπάνω. Μπορεί, βέβαια, να γίνουν και απλοί μετασχηματισμοί των τιμών κάποιων πεδίων, όπως για παράδειγμα η μετατροπή της ημερομηνίας γέννησης του υπαλλήλου σε ήλικία ή η αντικατάσταση των τιμών εισοδήματος με διαστήματα του υπαλλήλου σε ηλικία ή η αντικατάσταση των τιμών εισοδήματος με διαστήματα τιμών(χαμηλό, μέσο, υψηλό)ή δημιουργία ενός νέου πεδίου σαν συνάρτηση κάποιων άλλων, κ.α.Από τους πλέον συνήθεις μετασχηματισμούς είναι η μετατροπή του εύρους των αριθμητικών τιμών των πεδίων και η μετατροπή κατηγορικών δεδομένων σε αριθμητικά.

### 1.3.3 Τρίτο Στάδιο: Εξόρυξη από δεδομένα

Στο τρίτο στάδιο γίνεται αρχικά η επιλογή της κατάλληλης τεχνικής εξόρυξης από δεδομένα και έπειτα η επιλογή του κατάλληλου αλγόριθμου που την υλοποιεί .Επιπλέον, καθορίζονται οι παράμετροι των αλγορίθμων αυτών, ανάλογα με την περίπτωση εφαρμογής .

### 1.3.4 Τεταρτο Στάδιο:Αξιολόγηση

Το τελευταίο αυτό στάδιο είναι βασικό για την επιτυχία της μεθοδολογίας και αφορά αφενός στην αξιολόγησή της σημαντικότητας των εξαγομένων γνώσεων και αφετέρου στην παρουσιασή τους με όσο το δυνατό πιο κατανοητό και φιλικό τρόπο , στον τελικό χρήστη .Η σπουδαιότητα των εξαγομένων γνώσεων, που αναφέρεται με τον όρο σημασιολογική σπουδαιότητα, μπορεί να οριστεί με βάση συγκεκριμένες παραμέτρους του αλγόριθμου εξόρυξης από δεδομένα που χρησιμοποιήθηκε . Όσον αφορά στην παρουσίαση των εξαγομένων γνώσεων , θα πρέπει να σημειωθεί ότι αποτελεί καθοριστικό παράγοντα αποδοχής της μεθοδολογίας . Η απλή παράθεση των εξορυχθέντων κανόνων οι οποίοι μπορεί να είναι μέχρι και μερικές εκατοντάδες , απαιτεί από τον χρήστη επιπλέον προσπάθεια προσπέλασης αυτών που ενδιαφέρουν. Χρειάζονται περισσότερο περίτεχνες μέθοδοι παρουσίασης .Για παράδειγμα, έχει προταθεί η χρήση Έμπειρων Συστημάτων για την προσπέλαση των εξορυχθέντων κανόνων.

Είναι φανερό ότι το βασικότερο στάδιο της εξόρυξης από δεδομένα , αν και έχει υπολογιστεί ότι το 80% του χρόνου ολοκλήρωσης της ακολουθίας των παραπάνω βημάτων καταναλώνεται στα υπόλοιπα, μάλλον υποβοηθητικά, στάδια .
Στα επόμενα θα αναφερθούμε αποκλειστικά σε αυτό το στάδιο, περιγράφοντας τις διάφορες τεχνικές και αλγόριθμους που το υλοποιούν .

1.3.4.1 Οι τύποι γνώσης που μπορούν να εξαχθούν αυτόματα με την μορφή κανόνων

Υπάρχουν τρεις κύριοι τύποι γνώσης που μπορούν να εξαχθούν αυτόματα, με τη μορφή κανόνων, από τα δεδομένα:

1. Οι κανόνες ταξινόμησης
2. Οι κανόνες ομαδοποίησης
3. Οι κανόνες συσχέτισης

Οι μέθοδοι εξαγωγής τέτοιων κανόνων επίσης ποικίλουν. Έχουν χρησιμοποιηθεί μέθοδοι εξόρυξης από τα επιστημονικά πεδία της Μηχανικής Μάθησης (Machine Learning) ,της στατιστικής ανάλυσης (statistical analysis), κ.λ.π.Όμως ,στα επόμενα θα επικεντρωθούμε στις μεθόδους που χρησιμοποιούν αλγόριθμους Μηχανικής Μάθησης (Machine Learning) ,της στατιστικής ανάλυσης (statistical analysis) , της αριθμητικής ταξινόμησης (numerical taxonomy) της παλινδρόμησης(regression analysis) της ανάλυσης χρονοσειρών (time series analysis), κ.λ.π.Όμως, στα επόμενα θα επικεντρωθούμε στις μεθόδους που χρησιμοποιούν αλγόριθμους Μηχανικής Μάθησης (Machine Learning), θεωρώντας ότι αυτές έχουν το πλεονέκτημα έναντι των υπολοίπων να μπορούν να βασίζονται σε ήδη υπάρχουσα γνώση (background knowledge) για να εξάγουν νέα γνώση σ΄ ένα εννοιολογικό επίπεδο (conceptual level) και ταυτόχρονα να παρέχουν αποδεικτικά πειστήρια για την ορθότητα αυτής της νέας γνώσης πάλι σε ένα εννοιολογικό επίπεδο (conceptual level) .

# 2. ΤΑΞΙΝΟΜΗΣΗ

## 2.1 Τι είναι η ταξινόμηση

Η ταξινόμηση (classification) είναι μία από τις πιο δημοφιλείς και αποτελεσματικές τεχνικές εξόρυξης από δεδομένα. Έχει ερευνηθεί ιδιαίτερα τα τελευταία χρόνια και έχει προσφέρει πολύ σημαντικές εφαρμογές.Οι αλγόριθμοι ταξινόμησης (classification algorithms) [ π.χ. Boutsinas01, Clark, Quinlan, Rumelhart ] εφαρμόζονται σε δεδομένα, προταξινομημένα σε συγκεκριμένες κλάσεις, με στόχο την εξαγωγή κανόνων, τους οποίους στη συνέχεια μπορούμε να χρησιμοποιήσουμε για να ταξινομήσουμε νέα δεδομένα σ'αυτές τις συγκεκριμένες κλάσεις. Ένα σύνολο εξαγόμενων κανόνων ονομάζεται και ταξινομητής (Classifer) .

## 2.2 Η λειτουργικότητα ταξινομητή

Συγκεκριμένα, αν ένα σύνολο από δεδομένα δοθούν σαν είσοδος σ' έναν αλγόριθμο ταξινόμησης, ο αλγόριθμος <<μαθαίνει>>, αντιστοιχεί στη δημιουργία ενός συνόλου

από κανόνες. Στη συνέχεια ο αλγόριθμος μπορεί, βασιζόμενος σε αυτούς τους κανόνες να ταξινομήσει νέα δεδομένα.

## 2.3 Εφαρμογές από τους κανόνες ταξινόμησης

Οι κανόνες ταξινόμησης είναι οι πλέον διαδεδομένοι και χρησιμοποιούμενοι Μερικές συνήθεις εφαρμογές τους είναι: η αναγνώριση των κατάλληλων πελατών για αποστολή διαφημιστικών φυλλαδίων, η αναγνώριση της επικινδυνότητας των υποψηφίων για δάνειο από μια τράπεζα ,η διάγνωση μιας ασθένειας από τα συμπτώματα, κ.α.

### 2.3.1 Παράδειγμα: Δημιουργία Περιγραφών Πελατών

Ας θεωρήσουμε την εφαρμογή της δημιουργίας περιγραφών πελατών (customer profiles).Δηλαδή, ας υποθέσουμε, για παράδειγμα, ότι έχουμε έναν πίνακα μιας σχεσιακής βάσης δεδομένων σαν τον παρακάτω Πίνακα1, που χρησιμοποιείται για να <<μάθει >>ένας αλγόριθμος ταξινόμησης :

| ΗΛΙΚΙΑ | ΕΠΑΓΓΕΛΜΑ | ΠΕΡΙΟΧΗ | ΚΛΑΣΗ |
|---|---|---|---|
| 37 | ΠΑΙΔΙΑΤΡΟΣ | ΠΛΑΚΑ | ΚΑΛΟΣ |
| 33 | ΔΙΚΗΓΟΡΟΣ | ΠΕΙΡΑΙΑΣ | ΚΑΚΟΣ |
| 45 | ΟΔΗΓΟΣ | ΝΙΚΑΙΑ | ΜΕΤΡΙΟΣ |
| 37 | ΧΕΙΡΟΥΡΓΟΣ | ΚΗΦΗΣΙΑ | ΚΑΛΟΣ |
| ... | ... | ... | ... |

Πίνακας 1. Πίνακας Πελατών .

Το σύστημα θα <<μάθει >> κανόνες για το ποιός είναι ο τύπος του <<ΚΑΛΟΥ>>, του <<ΚΑΚΟΥ>>, και του <<ΜΕΤΡΙΟΥ>> πελάτη.Συνεπώς, ένας νέος πελάτης ανάλογα με την ηλικία , το επάγγελμα και περιοχη του ,μπορεί να ταξινομηθεί κατάλληλα ,με όλες τις συνέπειες που μ πορεί να έχει αυτό στη συμπεριφορά της επιχείρησης προς αυτόν .

## 2.4 Βήματα Αλγορίθμου Ταξινόμησης

Τα βήματα εφαρμογής ενός αλγορίθμου ταξινόμησης είναι απλά.Ας υποθέσουμε ότι έχουμε έναν πίνακα μίας σχεσιακής βάσης δεδομένων, δηλαδή ένα σύνολο από εγγραφές που η κάθε μια έχει ήδη ταξινομηθεί σε συγκεκριμένη κλάση της κάθε εγγραφής . Το σύνολο αυτό καλείται σύνολο εκπαίδευσης (training set) .Τα υπόλοιπα πεδία των εγγραφών, ή κάποια από αυτά που εμείς θα επιλέξουμε , είναι αυτά με τα οποία θα γίνει η εκπαίδευση .Μετά την εκπαίδευση μπορούμε να χρησιμοποιήσουμε τον ταξινομητή που προκύπτει σε άλλες εγγραφές για τις οποίες δεν ξέρουμε σε ποιά κλάση ανήκουν .Μπορεί να ελεγχθεί η ακρίβεια ταξινόμησης του ταξινομητή που εξήχθη με βάση ένα δεύτερο προταξινομημένο σύνολο εγγραφών , το σύνολο ελέγχου (test set). Ταξινομούμε το σύνολο ελέγχου εκ νέου χρησιμοποιώντας τον ταξινομητή και μετράμε το ποσοστό των λανθασμέ νων ταξινομήσεων (error rate).

## 2.5 Τύποι των Αλγορίθμων Ταξινόμησης

Υπάρχουν δύο βασικοί τύποι αλγόριθμων ταξινόμησης ανάλογα με την δομή του ταξινομητή που παράγουν .Οι αλγόριθμοι που παράγουν δέντρα αποφάσεων (decision trees) και αυτοί που παράγουν λίστες αποφάσεων (decision lists).

### 2.5.1 Παράδειγμα ταξινομητή για δέντρα αποφάσεων και λίστες αποφάσεων

| ΠΟΙΝΙΚΟ ΜΗΤΡΩΟ | ΕΙΣΟΔΗΜΑ | ΈΓΚΡΙΣΗ ΔΑΝΕΊΟΥ |
|---|---|---|
| Όχι | Χαμηλό | Όχι |
| Ναι | Χαμηλό | Όχι |
| Όχι | Υψηλό | Ναι |
| Ναι | Υψηλό | Όχι |

Το συγκεκριμένο σύνολο εκπαίδευσης έχει δύο πεδία, τα 'Ποινικό μητρώο' και 'εισόδημα' και ένα πεδίο κλάσης το 'έγκριση δανείου'.

Αντίστοιχα, ένας αλγόριθμος ταξινόμησης που παράγει λίστες αποφάσεων, εξάγει την παρακάτω λίστα κανόνων :

Αν ποινικό μητρώο:όχι και Εισόδημα :χαμηλό τότε Έγκριση Δανείου :όχι

Αν ποινικό μητρώο: όχι και Εισόδημα :υψηλό τότε Έγκριση Δανείου :Ναι

Αν ποινικό μητρώο: ναι τότε Έγκριση Δανείου :όχι

## 2.6 Οι αλγόριθμοι Ταξινόμησης συμπερασματικά

Οι ταξινομητές που παράγονται από τους αλγόριθμους ταξινόμησης δεν είναι μοναδικοί.Εξαρτώνται από τους παραμέτρους που καθορίζει ο χρήστης και σχετίζονται με την επιλογή του κριτηρίου με το οποίο επιλέγονται είτε κόμβοι ενός δέντρου αποφάσεων είτε πεδία σε μια λίστα αποφάσεων .Το κριτήριο αυτό συνήθως βασίζεται σε εκτιμήσεις του πόσο καλά χωρίζεται κάθε φορά το σύνολο εκπαίδευσης , ώστε να καταλήξουμε στο τέλος σε υποσύνολα που ανήκουν σε μια κλάση .Τέλος ,σημαντικό ζήτημα για τέτοιους αλγόριθμους είναι η αντεμετώπιση του θορύβου αλλά και η διαχείρηση των κενών τιμών στα πεδία .

Η εξαγωγή δέντρων αποφάσεων από σύνολα δεδομένων και η χρήση τους για ταξινόμηση είναι ίσως η παλαιότερη έκφραση της εξόρυξης από δεδομένα. Τα δέντρα αποφάσεων έχουν στη ρίζα τους και στους ενδιάμεσους κόμβους τιμές των διαφόρων πεδίων και στα φύλλα τους τιμές του πεδίου κλάσης . Ο κάθε κόμβος διακλαδώνεται προς τα κάτω , έχοντας ένα κλαδί για κάθε διακριτή τιμή του πεδίου. Σε περίπτωση που το πεδίο είναι συνεχές(continuous) αριθμητικό χωρίζεται το εύρος του πεδίου σε διαστήματα (value ranges) και ο κόμβος διακλαδώνεται με βάση αυτά.

Ένας αλγόριθμος που παράγει δέντρα αποφάσεων ακολουθεί συνήθως αναλυτική προσέγγιση δημιουργεί δηλαδή το δέντρο από τη ριζά και προχωράει προς τα κάτω. Με βάση όλο το σύνολο εκπαίδευσης , επιλέγει το πεδίο που θα τοποθετηθεί στη ρίζα του δέντρου . Για κάθε διακριτή τιμή του πεδίου αυτού , ορίζεται ένα υποσύνολο εγγραφών , οι εγγραφές του οποίου έχουν στο συγκεκριμένο πεδίο τη συγκεκριμένη διακριτή τιμή. Ο αλγόριθμος αναδρομικά, έχοντας κάνει την πρώτη διακλάδωση,

προσπαθεί να βρεί για το καθένα από τα υποσύνολα το δικό του υποδέντρο αποφάσεων (decision subtree).Μόλις συναντήσει υποσύνολο, το οποίο ανήκει όλο σε μία μόνο κλάση, σταματά τη διακλάδωση προς τα κάτω και τοποθετεί στο σημείο εκείνο του δέντρου ένα φύλλο με την κλάση στην οποία ανήκει το υποσύνολο .
Η διαφορά των αλγορίθμων ταξινόμησης που παράγουν δέντρα αποφάσεων επικεντρώνεται βασικά στον τρόπο με τον οποίο επιλέγεται κάθε φορά το πεδίο με βάση το οποίο θα γίνει η διακλάδωση . Συνήθως χρησιμοποιούνται στατιστικά κριτήρια που εφαρμόζονται πάνω στο εξεταζόμενο σύνολο εγγραφών . Αν τα πεδία έχουν επαρκή πληροφορία , είναι πάντα δυνατόν να βρεθεί ένα δέντρο αποφάσεων που ταξινομεί σωστά όλες τις εγγραφές του συνόλου εκπαίδευσης . Συνήθως μάλιστα, υπάρχουν περισσότερα του ενός τέτοια δέντρα αποφάσεων .Βέβαια, το ουσιώδες θέμα είναι το δέντρο αποφάσεων που θα προκύψει να είναι χρήσιμο για ταξινόμηση και νέων εγγραφών, εκτός του συνόλου εκπαίδευσης . Η αρχή που ακολουθείται για να επιλεγεί το καλύτερο δέντρο αποφάσεων είναι να προτιμηθεί το απλούστερο. Μάλιστα, ορισμένοι αλγόριθμοι ταξινόμησης , έχουν ένα επιπλέον βήμα μετά την παραγωγή του δέντρου αποφάσεων κατά το οποίο γίνεται η απλούστεσή του.

## 2.7 Ο ID3 (Induction of Decision Trees)-Αλγόριθμος ταξινόμησης

Ο αλγόριθμος ID3 παρουσιάστηκε ολοκληρωμένα από τον J.R. Quinlan στο περιοδικό Machine Learning το 1986.Η αρχική του έκδοση είχε δημοσιευτεί από τον ίδιο το 1979.Το όνομά του προκύπτει από τα αρχικά Induction of Decision Trees(3→three).Ο αλγόριθμος αυτός θεωρήθηκε πρωτοποριακός την εποχή εκείνη εξαιτίας της πληρότητας της μελέτης αλλά και της αποτελεσματικότητάς του ,ακόμα και σε ειδικές ανεπιθύμητες καταστάσεις συνόλου εκπαίδευσης , όπως ο θόρυβος (noise) ή οι κενές τιμές (missing values) .Η έρευνα που ακολούθησε ανέδειξε πολλές βελτιωμένες εκδόσεις του (π.χ. οι αλγόριθμοι C4.5 ,ID5, κ.α.), ενώ τα περισσότερα εμπορικά συστήματα εξόρυξης από δεδομένα βασίζονται σε αλγόριθμους που αποτελούν εκδόσεις του ID3.

Ο ID3 ανήκει στην οικογένεια των συστημάτων μάθησης TDIDT (Top-Down Induction of Decision Trees), ακολουθώντας την αναλυτική προσέγγιση (top - down).

Δέχεται σαν είσοδο ένα σύνολο εκπαίδευσης οι εγγραφές του οποίου έχουν προταξινομηθεί σε κλάσεις .Ο αλγόριθμος ,στην αρχική του μορφή ,θεωρεί δύο διακριτές τιμές κλάσης οι οποίες συμβολίζονται στη βιβλιογραφία ως P(Positive) και N(Negative).Βέβαια , ο αλγόριθμος , εύκολα μπορεί να επεκταθεί και για περισσότερες των δυο τιμές κλάσης. Αντί να εξάγεται το δέντρο αποφάσεων από ολόκληρο το σύνολο εκπαίδευσης , χρησιμοποιείται ένα <<παράθυρο>> , δηλαδή ένα υποσύνολο των εγγραφών. Με το δέντρο που προκύπτει ταξινομείται όλο το σύνολο εκπαίδευσης και ελέγχεται η ακρίβεια της ταξινόμησης. Αν όλες οι εγγραφές έχουν ταξινομηθεί σωστά, τότε το υπάρχον δέντρο γίνεται αποδεκτό και ο αλγόριθμος τερματίζει . Διαφορετικά , προστίθεται κι άλλες εγγραφές στο <<παράθυρο>> και η διαδικασία επαναλαμβάνεται .Αυτό συνεχίζεται μέχρι όλες οι εγγραφές να ταξινομούνται σωστά από το δέντρο. Εμπειρικά έχει βρεθεί ότι ο αλγόριθμος τερματίζει έτσι γρηγορότερα από το αν γινόταν χρήση ολοκλήρου του συνόλου εκπαίδευσης . Πάντως , για να εξασφαλιστεί η επιτυχής κατάληξη του αλγορίθμου, πρέπει να υπάρχει δυνατότητα το <<παράθυρο >> να μεγαλώσει τόσο ώστε να περιέχει το σύνολο των εγγραφών.

## 2.7.1 Παράμετροι του αλγορίθμου ID3

i. Μια παράμετρος , επομένως αυτού του αλγορίθμου ,είναι τι ποσοστό των εγγραφών θα περιέχει το <<παράθυρο>> , και με τι ρυθμό θα μεγαλώσει εφόσον δεν είναι επαρκές.

ii. Επόμενη και σημαντικότερη παράμετρος είναι το κριτήριο επιλογής του πεδίου για κάθε κόμβο , με το οποίο θα γίνει η διακλάδωση .Χρησιμοποιείται σαν κριτήριο επιλογής ένα μέγεθος δανεισμένο από την Θεωρία της Πληροφορίας , η . Εντροπία η οποία δίδεται από τον τύπο $-\Sigma p\{a\} \Sigma p\{c|a\}\log p\{c|a\}$, για το χαρακτηριστικό (πεδίο) α, όπου $P\{c|a\}$ είναι η δεσμευμένη πιθανότητα να ισχύει η τιμή κλάσης c όταν ισχύει η τιμή a του a. Το μέγεθος Εντροπία, ως αντίθετο στο

μέγεθος πληροφορία, δίνει μια εκτίμηση του πόσο λανθασμένα χωρίζεται καλύτερα το σύνολο εκπαίδευσης .

## 2.7.2 Τα βήματα του αλγορίθμου ID3

Τα βήματα του αλγορίθμου ID3 έχουν ως εξής:

1. Διάλεξε ένα πεδίο για ρίζα του δέντρου απόφασης και σχημάτισε διακλάδωση με ένα φύλλο για κάθε διαφορετική τιμή (ή διάστημα) αυτού του πεδίου
2. Το δέντρο απόφασης που έχει μέχρι στιγμής κατασκευασθεί χρησιμοποιείται για να ταξινομήσει το σύνολο εκπαίδευσης . Αν όλες οι εγγραφές που ταξινομούνται σε ένα συγκεκριμένο φύλλο ανήκουν στην ίδια κλάση , ονόμασε το φύλλο μ΄ αυτή την κλάση . Αν όλα τα φύλλα έχουν ονομασθεί με κάποια κλάση ο αλγόριθμος τελειώνει
3. Αλλιώς, για κάθε φύλλο που δεν έχει ονομασθεί με κάποια κλάση, διάλεξε ένα πεδίο που δεν έχει προηγουμένως επιλεγεί στο μονοπάτι από το φύλλο εώς τη ρίζα , ονόμασε το φύλλο (κόμβος πλέον) μ ΄ αυτό το πεδίο και σχημάτισε διακλάδωση μ΄ένα φύλλο για κάθε διαφορετική τιμή (ή διάστημα ) αυτού του πεδίου. Συνέχισε στο βήμα 2

## 2.7.3 Παράδειγμα εφαρμογής αλγορίθμου ID3 –Ασθενής οφθαλμιατρικής κλινικής

### 2.7.3.1 Χαρακτηριστικά (πεδία) και οι τιμές που μπορούν να πάρουν αυτά τα χαρακτηριστικά

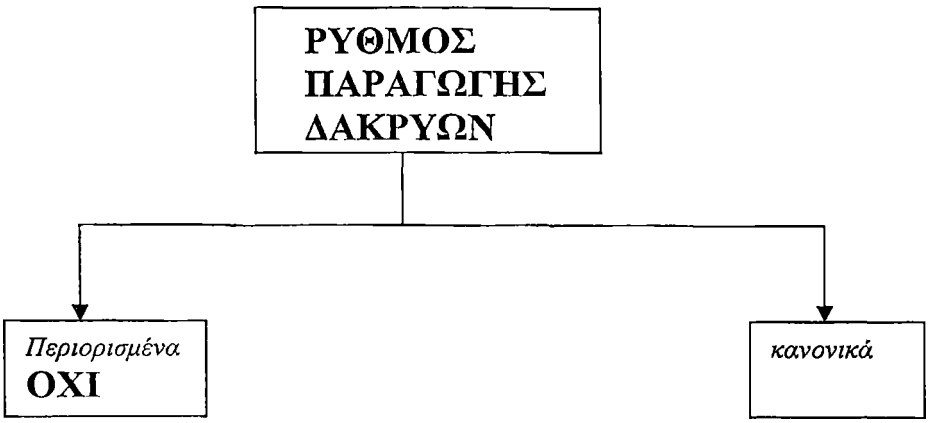
Έστω ότι το σύνολο εκπαίδευσης αποτελείται από 24 καταγεγραμμένες περιπτώσεις ασθενών μιας οφθαλμιατρικής κλινικής που χρειάστηκε να φορέσουν φακούς επαφής.Τα χαρακτηριστικά (πεδία) που καταγράφηκαν για κάθε περίπτωση καθώς και οι τιμές που μπορούν να πάρουν αυτά τα χαρακτηριστικά είναι:

1. ηλικία (age):που μπορεί να πάρει αντίστοιχα τις κωδικοποιημένες τιμές (1) νεαρή (young) ,(2) προ-πρεσβυωπική (pre-presbyopic), (3) πρεσβυωπική (presbyopic)
2. συνταγή για ματογυάλια (spectacle prescription) που μπορεί να πάρει τις κωδικοποιημένες τιμές (1)μυωπίας(myope) , (2) υπερμετρωπίας (hypermetrope)

3. αστιγματισμός (astigmatic) που μπορεί να πάρει αντίστοιχα τις
Κωδικοποιημένες τιμές (1)όχι (no) ,(2) ναι (yes)
4. ρυθμός παραγωγής δακρύων (tear production rate): που μπορεί να πάρει
Αντίστοιχα τις κωδικοποιημένες τιμές (1) περιορισμένα(reduced), (2) κανονικά
(normal)
5. φακοί επαφής (contact lenses) που μπορεί να πάρει αντίστοιχα τις
Κωδικοποιημένες τιμές (1)σκληροί (hard),(2)μαλακοί(soft),(3)όχι(no)
Το τελευταίο χαρακτηριστικό αντιστοιχεί στην κλάση.Το σύνολο εκπαίδευσης
φαίνεται στον Παρακάτω πίνακα:

| ΑΣΘΕΝΗΣ | ΗΛΙΚΙΑ | ΣΥΝΤ.ΓΙΑ ΜΑΤΟΓΥΑΛ. | ΑΣΤΙΓΜΑΤ. | ΡΥΘΜ.ΠΑΡΑΓ.ΔΑΚΡ. | ΦΑΚΟΙ ΕΠΑΦΗΣ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 3 |
| 2 | 1 | 1 | 1 | 2 | 2 |
| 3 | 1 | 1 | 2 | 1 | 3 |
| 4 | 1 | 1 | 2 | 2 | 1 |
| 5 | 1 | 2 | 1 | 1 | 3 |
| 6 | 1 | 2 | 1 | 2 | 2 |
| 7 | 1 | 2 | 2 | 1 | 3 |
| 8 | 1 | 2 | 2 | 2 | 1 |
| 9 | 2 | 1 | 1 | 1 | 3 |
| 10 | 2 | 1 | 1 | 2 | 2 |
| 11 | 2 | 1 | 2 | 1 | 3 |
| 12 | 2 | 1 | 2 | 2 | 1 |
| 13 | 2 | 2 | 1 | 1 | 3 |
| 14 | 2 | 2 | 1 | 2 | 2 |
| 15 | 2 | 2 | 2 | 1 | 3 |
| 16 | 2 | 2 | 2 | 2 | 3 |
| 17 | 3 | 1 | 1 | 1 | 3 |
| 18 | 3 | 1 | 1 | 2 | 3 |
| 19 | 3 | 1 | 2 | 1 | 3 |
| 20 | 3 | 1 | 2 | 2 | 1 |
| 21 | 3 | 2 | 1 | 1 | 3 |
| 22 | 3 | 2 | 1 | 2 | 2 |
| 23 | 3 | 2 | 2 | 1 | 3 |
| 24 | 3 | 2 | 2 | 2 | 3 |

Συνεπώς, σύμφωνα με το πρώτο βήμα, στη ρίζα του δέντρου επιλέγεται το πεδίο <<ρυθμός παραγωγής δακρύων>> και σχηματίζεται διακλάδωση για τις τιμές του <<περιορισμένα>> και <<κανονικά>> (βλέπε εικόνα 3).Παρατηρούμε ότι όλες οι εγγραφές που ταξινομούνται στο φύλλο <<περιορισμένα>> ανήκουν στην ίδια κλάση <<όχι>>, οπότε, σύμφωνα με το δεύτερο βήμα, ονομάζουμε το φύλλο μ' αυτή την κλάση.




Για το άλλο φύλλο πρέπει, σύμφωνα με το τρίτο βήμα, να διαλέξουμε ένα πεδίο που δεν έχει προηγουμένως επιλεγεί στο μονοπάτι από το φύλλο έως τη ρίζα. Η εντροπία για κάθε πιθανό πεδίο είναι:

1. Entropy(ηλικία)=1,3333
2. Entropy(συνταγή για ματογυάλια) =1,4591
3. Entropy(αστιγματισμός)=0,7842

Οπότε επιλέγουμε <<αστιγματισμός>> και ονομάζουμε τον κόμβο μ' αυτό το πεδίο . Στη συνέχεια σχηματίζουμε διακλάδωση μ' ένα φύλλο για κάθε διαφορετική τιμή(<<ναι>>, <<όχι>>) αυτού του πεδίου, όπως φαίνεται στην Εικόνα 4.


Η διαδικασία συνεχίζεται μέχρις ότου ονομασθούν όλα τα φύλλα. Το τελικό δέντρο απόφασης φαίνεται στην Εικόνα 5.

## 2.7.4 Παράδειγμα αλγορίθμουID3 –Αυτόματη προσγείωση ενός διαστημόπλοιου

Εφαρμογή του αλγορίθμου βάση του οποίου μπορούμε να εξάγουμε συμπεράσματα για τις συνθήκες κάτω από τις οποίες ενδείκνυται η αυτόματη προσγείωση (auto landing) ενός διαστημοπλοίου σε σχέση με τον χειροκίνητο έλεγχο.

| A / A | Stability | Error | Sign | Wind | Magnitude | Visibility | Class |
|---|---|---|---|---|---|---|---|
| 1 | stab | MM | PP | tail | Low | No | auto |
| 2 | xstab | MM | PP | tail | Low | Yes | no auto |
| 3 | stab | LX | PP | tail | Low | Yes | no auto |
| 4 | stab | XL | PP | tail | Low | Yes | no auto |
| 5 | stab | MM | nn | tail | Low | Yes | no auto |
| 6 | stab | MM | PP | tail | Out of range | Yes | no auto |
| 7 | stab | SS | PP | tail | Low | Yes | auto |
| 8 | stab | SS | PP | tail | Medium | Yes | auto |
| 9 | stab | SS | PP | tail | Strong | Yes | auto |
| 10 | stab | MM | PP | head | Low | Yes | auto |
| 11 | stab | MM | PP | head | Medium | Yes | auto |
| 12 | stab | MM | PP | tail | Low | Yes | auto |
| 13 | stab | MM | PP | tail | Medium | Yes | auto |
| 1 | stab | MM | PP | tail | Strong | Yes | no |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4 | | | | | | | auto |
| 15 | stab | MM | PP | tail | Strong | Yes | auto |

Πίνακας 1

Θα υπολογίσουμε για κάθε ένα από τα χαρακτηριστικά την εντροπία.

- Από τον πίνακα για το χαρακτηριστικό «stability» έχουμε:

  P(no auto/stab) = 5/14

  P(auto/stab) = 9/14

  P(stab) = 14/15

  P(no auto/xstab) = 1/1 = 1

  P(auto/xstab) = 0

  P(xstab) = 1/15

  Άρα $E_{stability}$ = - { 14/15· [5/14 · log(5/14) + 9/14 · log(9/14)] +

  1/15 · (1 · log(1) + 0 · log(0) ) } =

  = - { 14/15 · [ 0.357 · (-0,447) + 0.642 · (-0.191) ] } =

  = - { 0.933 · [ -0.159 -0.122 ] } = 0.933 · 0.281 = 0.262

- Για το χαρακτηριστικό «error» έχουμε:

  P(MM) = 10/15

  P(auto/MM) = 6/10

  P(no auto/MM) = 4/10

  P(LX) = 1/15

  P(auto/LX) = 0

  P(no auto/LX) = 1/1 = 1

  P(XL) = 1/15

  P(auto/XL) = 0

  P(no auto/XL) = 1

  P(SS) = 3/15

  P(auto/SS) = 3/3 = 1

  P(no auto/SS) = 0

  Άρα $E_{error}$ = - {10/15 · [6/10 · log(6/10) + 4/10 · log(4/10)] +

  1/15 · [0 · log(0) + 1 · log(1)] + 1/15 · [0 · log(0) + 1 · log(1)]

$$+3/15 \cdot [1 \cdot \log(1) + 0 \cdot \log(0)\ ]\ \} =$$

$$= -\ 10/15 \cdot [0.6 \cdot (-0.221) + 0.4 \cdot (-0.397)\ ] =$$

$$= -\ 10/15 \cdot (\ -0.133 - 0.159\ ) = 0.29 \cdot 10/15 = \underline{0.194}$$

- Για το χαρακτηριστικό «sign» έχουμε:

  P(PP) = 14/15

  P(auto/PP) = 9/14

  P(no auto) = 5/14

  P(nn) = 1/15

  P(auto/nn) = 0

  P(no auto/nn) = 1

  Άρα $E_{sign} = -\ \{\ 14/15 \cdot [9/14 \cdot \log(9/14) + 5/14 \cdot \log(5/14)] +$

  $1/15 \cdot [0 \cdot \log(0) + 1 \cdot \log(1)\ ]\ \} =$

  $= -\ \{\ 0.933 \cdot (\ -0.642 \cdot 0.19 - 0.357 \cdot 0.447\ )\ \} =$

  $= -\ \{\ 0.933 \cdot (\ -0.121 - 0.159)\ \} = \underline{0.261}$

- Για το χαρακτηριστικό «wind» έχουμε:

  P(tail) = 12/15

  P(auto/tail) = 7/12

  P(no auto/tail) = 5/12

  P(head) = 3/15

  P(auto/head) = 2/3

  P(no auto/head) = 1/3

  Άρα $E_{wind} = -\ \{12/15 \cdot [\ 7/12 \cdot \log(7/12) + 5/12 \cdot \log(5/12)\ ] +$

  $13/15 \cdot [2/3 \cdot \log(2/3) + 1/3 \cdot \log(1/3)\ ]\ \} =$

  $= -\ \{12/15 \cdot (-0.583 \cdot 0.234 - 0.416 \cdot 0.38) +$

  $13/15 \cdot (-0.666 \cdot 0.176 - 0.333 \cdot 0.477)\ \} =$

  $= -\ \{0.8 \cdot (-0.136\ -0.158) + 0.2 \cdot (-0.117 - 0.158)\ \} =$

  $= 0.2352 + 3.6972 = \underline{0.238}$

- Για το χαρακτηριστικό «magnitude» έχουμε:

  P(Low) = 8/15

  P(auto/Low) = 4/8 = 1/2

P(no auto/Low) = 4/8 = 1/2

P(out of range) = 1/15

P(auto/out of range) = 0

P(no auto/out of range) = 1

P(medium) = 3/15

P(auto/medium) = 3/3 = 1

P(no auto/medium) = 0

P(strong) = 3/15

P(auto/strong) = 2/3

P(no auto/strong) = 1/3

Άρα $E_{magnitude}$ = - {8/15 · [4/8 · log(4/8) + 4/8 · log(4/8)] +

1/15 · [0 · log(0) + 1 · log(1)] +

3/15 · [1 · log(1) + 0 · log(0)] +

3/15 · [2/3 · log(2/3) + 1/3 · log(1/3)] } =

= 0.53 · (2 · 0.5 · 0.301) + 0.2 · (0.666 · 0.176 +

0.333 · 0.477) =

= 0.159 + 0.2 (0.117 + 0.158) = 0.214

- Για το χαρακτηριστικό «visibility» έχουμε:

P(No) = 1/15

P(auto/No) = 1/1 = 1

P(no auto/No) = 0

P(Yes) = 14/15

P(auto/Yes) = 8/14

P(no auto/Yes) = 6/14

Άρα $E_{visibility}$ = - {1/15 · [1 · log(1) + 0 log(0)] +

14/15 · [8/14 · log(8/14) + 6/14 · log(6/14)]} =

= 0.9333 · (0.571 · 0.243 + 0.428 · 0.367) =

= 0.9333 · (0.134 + 0.157) = 0.275

Παρατηρούμε ότι το χαρακτηριστικό «error» έχει μικρότερη εντροπία. Οπότε ο διαχωρισμός θα γίνει με βάση το χαρακτηριστικό αυτό:

- Όταν error = MM

| Stability | Error | Sign | Wind | Magnitude | Visibility | Class |
|---|---|---|---|---|---|---|
| stab | MM | PP | tail | Low | No | auto |
| xstab | MM | PP | tail | Low | Yes | no auto |
| stab | MM | nn | tail | Low | Yes | no auto |
| stab | MM | PP | tail | Out of range | Yes | no auto |
| stab | MM | PP | head | Low | Yes | auto |
| stab | MM | PP | head | Medium | Yes | auto |
| stab | MM | PP | tail | Low | Yes | auto |
| stab | MM | PP | Tail | Medium | Yes | auto |
| stab | MM | PP | head | Strong | Yes | no auto |
| stab | MM | PP | tail | Strong | Yes | auto |

Πίνακας 2

- Όταν «error» = LX τότε το «class» = no auto[1]
- Όταν «error» = XL τότε το «class» = no auto[2]
- Όταν «error» = SS, τότε

| Stability | Error | Sign | Wind | Magnitude | Visibility | Class |
|---|---|---|---|---|---|---|
| stab | SS | PP | tail | Low | Yes | auto |
| stab | SS | PP | tail | Medium | Yes | auto |
| stab | SS | PP | tail | Strong | Yes | auto |

Παρατηρούμε πως όταν «error» = SS τότε «class» = auto, άρα δεν χρειάζεται να υπολογίσουμε τις εντροπίες

- Όταν «error» = MM για το χαρακτηριστικό «stability» έχουμε:

P(stab) = 9/10

P(auto/stab) = 6/9 = 2/3

[1] Δεν χρειάζεται να υπολογίσουμε τις εντροπίες
[2] Δεν χρειάζεται να υπολογίσουμε τις εντροπίες

P(no auto/stab) = 3/9 = 1/3

P(xstab) = 1/10

P(auto/xstab) = 0

P(no auto/xstab) = 1

Άρα $E_{stability}$ = - {9/10 · [6/9 · log(6/9) + 3/9 · log(3/9)] +

1/10 · [0 · log(0) + 1 · log(1)] } =

= 0.9 · (0.11738 + 0.1593) = <u>0.24877</u>

- Όταν «error» = ΜΜ για το χαρακτηριστικό «sign» έχουμε:

P(PP) = 9/10

P(auto/PP) = 6/9 = 2/3

P(no auto/PP) = 3/9 = 1/3

P(nn) = 1/10

P(auto/nn) = 0

P(no auto/nn) = 1

Άρα $E_{sign}$ = - {9/10 · [2/3 · log(2/3) + 1/3 · log(1/3)] +

1/10 · [0 · log(0) + 1 · log(1)] } =

= 0.9 · (0.1173 + 0.15983) = <u>0.24877</u>

- Όταν «error» = ΜΜ για το χαρακτηριστικό «wind» έχουμε:

P(tail) = 7/10

P(auto/tail) = 4/7

P(no auto/tail) = 3/7

P(head) = 3/10

P(auto/head) = 2/3

P(no auto/head) = 1/3

Άρα $E_{wind}$ = - {7/10 · [4/7 · log(4/7) + 3/7 · log(3/7)] +

3/10 · [2/3 · log(2/3) + 1/3 · log(1/3)]} =

= 0.7 · (0.571 · 0.243 + 0.428 · 0.367) + 0.3 · 0.275 =

= <u>0.289</u>

- Όταν «error» = ΜΜ για το χαρακτηριστικό «magnitude» έχουμε:

P(Low) = 5/10 = 1/2

P(auto/Low) = 3/5

P(no auto/Low) = 2/5

P(out of range) = 1/10

P(auto/out of range) = 0

P(no auto/out of range) = 1

P(medium) = 2/10 = 1/5

P(auto/medium) = 2/2 = 1

P(no auto/medium) = 0

P(strong) = 2/10 = 1/5

P(auto/strong) = 1/2

P(no auto/strong) = 1/2

Άρα $E_{magnitude}$ = - {1/2 · [3/5 · log(3/5) + 2/5 · log(2/5)] +

1/10 · [0 · log(0) + 1 · log(1)] +

1/5 · [1 · log(1) + 0 · log(0)] +

1/5 · [1/2 · log(1/2) + 1/2 · log(1/2)]} =

= 1/2 · (0.6 · 0.221 + 0.4 · 0.397) + 1/5 · (2 · 1/2 · 0.301) =

= <u>0.2962</u>

- Όταν «error» = ΜΜ για το χαρακτηριστικό «visibility» έχουμε:

P(No) = 1/10

P(auto/No) = 1

P(no auto/No) = 0

P(Yes) = 9/10

P(auto/Yes) = 5/9

P(no auto/Yes) = 4/9

Άρα $E_{visibilitye}$ = - {1/10 · [ 1 · log(1) + 0 · log(0)] +

9/10 · [5/9 · log(5/9) + 4/9 · log(4/9)] } =

= 0.9 · (0.555 · 0.255 + 0.444 · 0.352) = <u>0.2673</u>

Παρατήρηση: όταν το «error» = ΜΜ το χαρακτηριστικό «stability» και το χαρακτηριστικό «sign» έχουν ίσες εντροπίες.

Διακρίνω δυο περιπτώσεις

1. Επιλέγω το χαρακτηριστικό «stability», άρα ο διαχωρισμός θα γίνει με βάση αυτό.

- Όταν «error» = MM και «stability» = xstab τότε class = «no auto»
- Όταν «error» = MM και «stability» = stab τότε προκύπτει ο πίνακας 3,

| Stability | Error | Sign | Wind | Magnitude | Visibility | Class |
|---|---|---|---|---|---|---|
| stab | MM | PP | tail | Low | No | auto |
| stab | MM | nn | tail | Low | Yes | no auto |
| stab | MM | PP | tail | Out of range | Yes | no auto |
| stab | MM | PP | head | Low | Yes | auto |
| stab | MM | PP | head | Medium | Yes | auto |
| stab | MM | PP | tail | Low | Yes | auto |
| stab | MM | PP | tail | Medium | Yes | auto |
| stab | MM | PP | head | Strong | Yes | no auto |
| stab | MM | PP | tail | Strong | Yes | auto |

Πίνακας 3

- Όταν «error» = MM και «stability» = stab για το χαρακτηριστικό «sign» έχουμε:

P(PP) = 8/9

P(auto/PP) = 36/8 = 3/4

P(no auto/PP) = 2/8 = 1/4

P(nn) = 1/9

P(auto/nn) = 0

P(no auto/nn) = 1

Άρα $E_{sign}$ = - {8/9 · [3/4 · log(3/4) + 1/4 · log(1/4)] +

1/9 · [0 · log(0) + 1 · log(1)] } =

= 8/9 · (0.0983 + 0.1005) = <u>0.2164</u>

- Όταν «error» = MM και «stability» = stab για το χαρακτηριστικό «wind» έχουμε:

  P(tail) = 6/9 = 2/3

  P(auto/tail) = 4/6 = 2/3

  P(no auto/tail) =2/6 = 1/3

  P(head) = 3/9 = 1/3

  P(auto/head) = 2/3

  P(no auto/head) = 1/3

  Άρα $E_{wind}$ = - {2/3 · [2/3 · log(2/3) + 1/3 · log(1/3)] +

  1/3 · [2/3 · log(2/3) + 1/3 · log(1/3)] } =

  = 0.2762

- Όταν «error» = MM και «stability» = stab για το χαρακτηριστικό «magnitude» έχουμε:

  P(low) = 4/9

  P(auto/low) = 3/4

  P(no auto/low) = 1/4

  P(out of range) = 1/9

  P(auto/out of range) = 0

  P(no auto/out of range) = 1

  P(medium) = 2/9

  P(auto/medium) = 2/2 = 1

  P(no auto/medium) = 0

  P(strong) = 2/9

  P(auto/strong) = 1/2

  P(no auto/strong) = 1/2

  Άρα $E_{magnitude}$ = - {4/9 · [3/4 · log(3/4) + 1/4 · log(1/4)] +

  1/9 · [0 · log(0) + 1 · log(1)] +

  2/9 · [1 · log(1) + 0 · log(0)] +

  2/9 · [1/2 · log(1/2) + 1/2· log(1/2)] } =

  = 0.1742

- Όταν «error» = MM και «stability» = stab για το χαρακτηριστικό «visibility» έχουμε:

  P(no) = 1/9

  P(auto/no) = 1

  P(no auto/no) = 0

  P(Yes) = 8/9

  P(auto/Yes) = 8/9

  P(no auto/Yes) = 3/8

  Άρα $E_{vosibility}$ = - {1/2 · [1 · log(1) + 0 · log(0)] +

  8/9 · [5/8 · log(5/8) + 3/8 · log(3/8)] } =

  = 0.2549

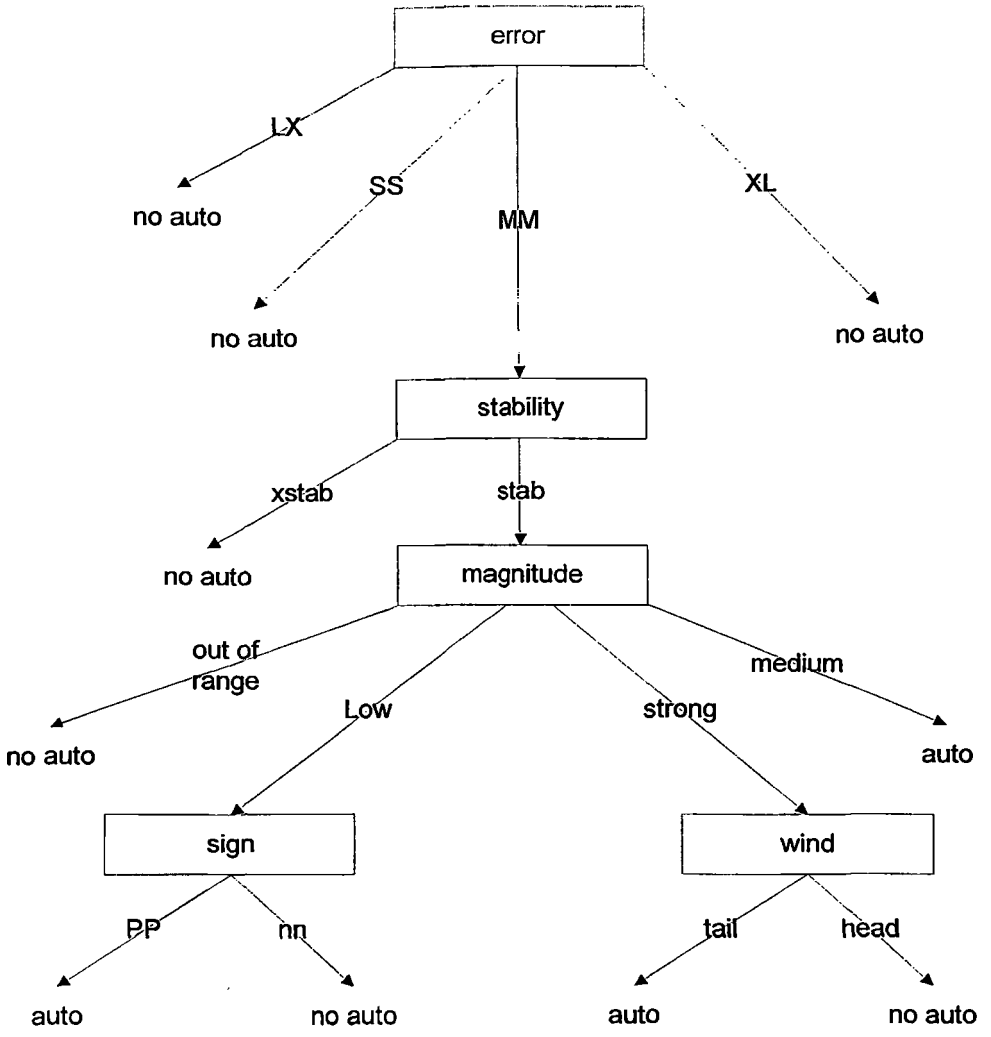
Παρατήρηση: το χαρακτηριστικό «magnitude» έχει την μικρότερη εντροπία. Από τον πίνακα 3 έχουμε:

- Όταν «error» = MM και «stability» = stab και «magnitude» = out of range, τότε «class» = no auto
- Όταν «error» = MM και «stability» = stab και «magnitude» = medium, τότε «class» = auto
- Όταν «error» = MM και «stability» = stab και «magnitude» = Low, τότε προκύπτει ο πίνακας που ακολουθεί

| Stability | Error | Sign | Wind | Magnitude | Visibility | Class |
|---|---|---|---|---|---|---|
| stab | MM | PP | tail | Low | No | auto |
| stab | MM | nn | tail | Low | Yes | no auto |
| stab | MM | PP | head | Low | Yes | auto |
| stab | MM | PP | tail | Low | Yes | auto |

- Όταν «error» = MM και «stability» = stab και «magnitude» = Low, υπολογίζουμε πάλι τις εντροπίες:

- Όταν «error» = MM και «stability» = stab και «magnitude» = Low για το χαρακτηριστικό «sign» έχουμε:




# 3.ΟΜΑΔΟΠΟΙΗΣΗ

## 3.1 Διαφοροποίηση κανόνων ομαδοποίηση-Ταξινόμηση

Οι κανόνες ομαδοποίησης είναι επίσης πολύ διαδεδομένοι και εφαρμόζονται ευρέως. Διαφέρουν από τους κανόνες ταξινόμησης στο ότι τα δεδομένα που χρησιμοποιούνται για μάθηση δεν είναι προταξινομημένα. Για παράδειγμα , ο παρακάτω πίνακας είναι όμοιος με τον Πίνακα 1 , αλλά λείπει το πεδίο <<κλάση>> , δηλαδή ο χαρακτηρισμός του κάθε πελάτη. Το ζητούμενο εδώ είναι να ομαδοποιηθούν τα δεδομένα σε (k) ομάδες , των οποίων τον αριθμό (k) δίνει ο

χρήστης . Άρα το αποτέλεσμα είναι (k) κανόνες που περιγράφουν ο καθένας μια από (k) ομάδες .Για παράδειγμα, μια ομάδα που προκύπτει από τα δεδομένα του παρακάτω πίνακα, θα μπορούσε να είναι <<ΝΕΟΙ, ΙΑΤΡΟΙ, ΚΑΤΟΙΚΟΙ ΑΚΡΙΒΩΝ ΠΕΡΙΟΧΩΝ>>.

| ΗΛΙΚΙΑ | ΕΠΑΓΓΕΛΜΑ | ΠΕΡΙΟΧΗ |
|---|---|---|
| 35 | ΙΑΤΡΟΣ | ΠΛΑΚΑ |
| 47 | ΔΙΚΗΓΟΡΟΣ | ΠΕΙΡΑΙΑΣ |
| 55 | ΟΔΗΓΟΣ | ΝΙΚΑΙΑ |
| 33 | ΙΑΤΡΟΣ | ΚΗΦΙΣΙΑ |
| ..... | ..... | ..... |

Ο διαχωρισμός ενός μεγάλου συνόλου αντικειμένων σε ομογενείς ομάδες (clustering) είναι μια από τις πλέον θεμελιώδεις τεχνικές εξόρυξης από δεδομένα .Η τεχνική ομαδοποίησης χωρίζει ουσιαστικά ένα σύνολο εγγραφών σε ομάδες έτσι ώστε οι εγγραφές που βρίσκονται στην ίδια ομάδα να έχουν περισσότερες ομοιότητες μεταξύ τους ,με βάση ορισμένα προκαθορισμένα κριτήρια, απ' ότι με εγγραφές άλλων ομάδων.Τα τελευταία χρόνια η τεχνική ομαδοποίησης εφαρμόζεται πολύ συχνά στην ανάλυση μεγάλων συνόλων δεδομένων για μεγάλη ποικιλία εφαρμογών. Έχουν αποδειχθεί ιδιαίτερα χρήσιμες για παράδειγμα στην βιολογία για εξαγωγή ιεραρχιών των οργανισμών του φυτικού βασιλείου, στη ψυχολογία για κατάταξη των ατόμων σε κατηγορίες προσωπικοτήτων, στην αστρονομία για την κατηγοριοποίηση γαλαξιών και γενικότερα των ουράνιων σωμάτων. Σημαντική είναι, βέβαια, η χρησιμότητά τους στη διοίκηση επιχειρήσεων. Σήμερα είναι ιδιαίτερα σημαντικό για τις επιχειρήσεις να μπορούν να αξιολογήσουν ένα νέο πελάτη με βάση την ομάδα στην οποία κατατάσσεται ή ακόμα να προσδιορίσουν τα χαρακτηριστικά των πελατών που αποφέρουν μεγάλα κέρδη στην εταιρεία .Με βάση αυτό το διαχωρισμό των πελατών που αποφέρουν μεγάλα κέρδη στην εταιρεία. Με βάση αυτό το διαχωρισμό των πελατών μπορούν να προσανατολίσουν τη στρατηγική της εταιρείας στην εξειδικευμένη εξυπηρέτηση ορισμένων ομάδων . Για παράδειγμα, από την ανάλυση ενός πολύ μεγάλου συνόλου πελατών, μπορεί να μειωθεί το υψηλό κόστος μίας διαφημιστικής εκστρατείας που βασίζεται στην αποστολή ενημερωτικών φυλλαδίων. Αυτό γίνεται περιορίζοντας το πλήθος των πελατών στους οποίους απευθύνεται, επιλέγοντας αυτούς με μεγάλη πιθανότητα να αντιδράσουν θετικά.

## 3.2 Εφαρμογές των κανόνων ομαδοποίησης

Μερικές συνήθεις εφαρμογές των κανόνων ομαδοποίησης είναι: η κατανομή της συνδρομητικής ή πελατειακής βάσης μίας επιχείρησης σε ομάδες ,ώστε να ακολουθηθεί κατάλληλη πολιτική σε κάθε μία, η επεξεργασία ερωτηματολογίων, η ομαδοποίηση συνδρομητών χρηστών που προσπελαύνουν μία ιστοσελίδα, ο εντοπισμός ταυτόσημων δεδομένων (για παράδειγμα, τραπεζικοί λογαριασμοί που αφορούν συγγενείς ιδιοκτήτες ), η αναγνώριση αντικειμένων σε φωτογραφίες. Πολλοί αλγόριθμοι ομαδοποίησης έχουν προταθεί κατά καιρούς [π.χ. Huang, Jain, Sander, Vrahatis]. Η αποδοτικότητα τους σχετίζεται άμεσα με το είδος των δεδομένων που θα διαχειριστούν .

## 3.2.1 Προυπόθεση για επιλογή κατάλληλου αλγόριθμου

Για να μπορέσει να γίνει η επιλογή του κατάλληλου αλγορίθμου απαραίτητη προϋπόθεση είναι η μελέτη των δεδομένων που θα χρησιμοποιηθούν για τον προσδιορισμό κυρίως του κριτηρίου ομοιότητας των εγγραφών μίας ομάδας .Έχουν προταθεί διάφορα μέτρα (αν)ομοιότητας, ανάλογα με τη φύση των δεδομένων . Τα πλέον γνωστά είναι: η Ευκλείδεια απόσταση, ο υπολογισμός του αριθμού των διαφορετικών τιμών σε συγκεκριμένα πεδία, η συχνότητα εμφάνισης των διαφόρων τιμών στα πεδία αλλά και συνδυασμοί τους .Γενικά , η τεχνική της ομαδοποίησης μπορεί να είναι: η στατιστική ή αριθμητική (statistical/numerical clustering) και εννοιολογική(conceptual clustering).Στην πρώτη περίπτωση χρησιμοποιούνται διάφορα αριθμητικά κριτήρια ομοιότητας ενώ στη δεύτερη ο προσδιορισμός των ομάδων βασίζεται στο νόημα και στις έννοιες που αυτά αντιπροσωπεύουν .Έτσι στην αριθμητική ομαδοποίηση οι ομάδες που προκύπτουν περιγράφονται από αριθμητικές τιμές ενώ στη εννοιολογική ομαδοποίηση οι ομάδες που προκύπτουν περιγράφονται από αριθμητικές τιμές ενώ στη εννοιολογική ομαδοποίηση από κατηγορικές . Πολλοί από τους αλγορίθμους ομαδοποίησης απαιτούν το σύνολο εκπαίδευσης που επεξεργάζονται να είναι είτε αριθμητικό (π.χ k-means[Jain]) είτε κατηγορικό (π.χ. k-modes [Huang]).Σ'αυτές τις περιπτώσεις μπορεί να γίνει μετασχηματισμός όπως

αναφέρθηκε προηγούμενα. Υπάρχουν βέβαια και αλγόριθμοι ομαδοποίησης που επιτρέπουν μικτό σύνολο εκπαίδευσης (π.χ. k-prototypes [Huang]).

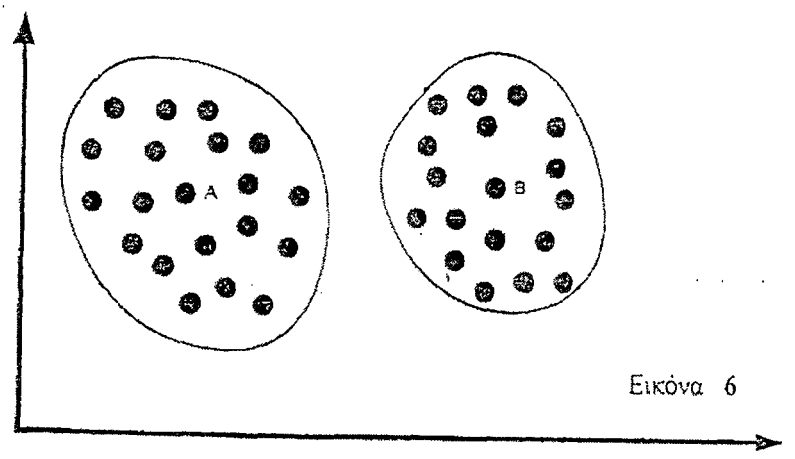
## 3.3 Ο K-Means αλγόριθμος ομαδοποίησης –Σύντομη αναφορά

Στη συνέχεια θα περιγραφεί αναλυτικά ένας από τους βασικότερους αλγόριθμους ομαδοποίησης, ο k-means. Ο αλγόριθμος k-means παρουσιάστηκε αρχικά το 1967. Από τότε ένα μεγάλο πλήθος από διάφορες εκδόσεις του έχει προταθεί στη βιβλιογραφία. Η συντριπτική πλειοψηφία των αλγορίθμων ομαδοποίησης βασίζονται στην ιδέα του k-means ενώ τον χρησιμοποιούν τα περισσότερα εμπορικά συστήματα εξόρυξης από δεδομένα αλλά και τα πιο γνωστά συστήματα στατιστικής ανάλυσης.

### 3.3.1 Λειτουργία του αλγόριθμου K-means –που στηρίζεται ο αλγόριθμος

Ο αλγόριθμος k-means έχει αποδειχθεί ότι είναι αποδοτικός σε πάρα πολλές πρακτικές εφαρμογές. Διαχωρίζει τα δεδομένα του συνόλου εκπαίδευσης σε k ομάδες , όπου καθορίζεται από τον χρήστη . Η λειτουργία του βασίζεται σε διαδοχικές επαναλήψεις κατά τις οποίες τα δεδομένα κατατάσσονται σε κάποια ομάδα με βάση την ομοιότητα που παρουσιάζουν με το μέσο αυτής της ομάδας . Στη συνέχεια θα περιγραφούν οι βασικές αρχές και τα γενικά βήματα του k-means αλγορίθμου .
Ο αλγόριθμος στηρίζεται ουσιαστικά σε κάποια αντιπροσωπευτικά δείγματα(means) κάθε ομάδας . Κάθε μια από τις k ομάδες που θα δημιουργηθούν θα περιέχει ένα αντιπροσωπευτικό δείγμα το οποίο ουσιαστικά θα αντιπροσωπεύει την ομάδα καθώς θα αποτελεί μια τυπική (μέση) περιγραφή της ομάδας. Για να γίνει πιο κατανοητό ,ας θεωρήσουμε ότι συμβολίζουμε τις εγγραφές μιας ομάδας σαν σημεία σ' ένα πολυδιάστατο επίπεδο . Οι διαστάσεις αυτού του επιπέδου θα είναι όσες και τα χαρακτηριστικά (attributes) , (παράβαλε <<πεδία>>), των εγγραφών του συνόλου εκπαίδευσης . Το αντιπροσωπευτικό δείγμα θεωρούμε ότι θα είναι το κέντρο βάρους του πολυδιάστατου σχήματος που ορίζουν τα σημεία αυτά. Στο παρακάτω σχήμα της

Εικόνας 6 , θεωρούμε ότι οι εγγραφές χαρακτηρίζονται από δυο πεδία και επομένως μπορούν να αναπαρασταθούν σε δισδιάστατο επίπεδο . Όπως φαίνεται έχουν σχηματιστεί δυο ομάδες και τα αντιπροσωπευτικά δείγματα αυτών θα είναι τα σημεία Α και Β που είναι τα κέντρα βάρους των ομάδων .




Εικόνα 6

Ο αλγόριθμος προσπαθεί να κατατάξει τις εγγραφές στις διάφορες ομάδες έτσι ώστε μετά τον τερματισμό, κάθε εγγραφή ν' ανήκει σ' εκείνη την ομάδα από της οποίας το αντιπροσωπευτικό δείγμα απέχει λιγότερο σε σχέση μ'αυτά των υπολοίπων ομάδων .

## 3.3.2 Βήματα του αλγορίθμου K-Means

Πιο αναλυτικά τα βήματα του αλγορίθμου είναι τα εξής:

1. Προσδιόρισε το k
2. Πάρε τα αρχικά k αντιπροσωπευτικά δείγματα(π.χ. πάρε τις k πρώτες εγγραφές σαν αντιπροσωπευτικά δείγματα )
3. Επανέλαβε
    - Για κάθε εγγραφή βρες την απόστασή της από τα αντιπροσωπευτικά δείγματα και θεώρησε ότι ανήκει στην ομάδα του πιο κοντινότερου αντιπροσωπευτικού δείγματος .
    - Υπολόγισε τα νέα αντιπροσωπευτικά δείγματα (κέντρα βάρους ) των ομάδων μέχρι να μη γίνονται αλλαγές

Στην πρώτη φάση θα πρέπει να επιλεγούν τα αντιπροσωπευτικά δείγματα στα οποία θα βασιστεί η αρχική κατάταξη των εγγραφών στις ομάδες . Επιλέγονται έτσι, είτε τυχαία είτε με κάποιες ευρετικές μεθόδους, τα k αρχικά 'σημεία' στον πολυδιάστατο

χώρο. Μ' αυτόν τον τρόπο κάθε ομάδα αντιπροσωπεύεται από ένα από τα k αντιπροσωπευτικά δείγματα . Στη συνέχεια, για κάθε εγγραφή υπολογίζουμε την ομοιότητα που παρουσιάζει με το αντιπροσωπευτικό δείγμα κάθε ομάδας .Αν θεωρήσουμε το παράδειγμα με τον πολυδιάστατο χώρο και την Ευκλείδεια απόσταση σαν μέτρο ομοιότητας, ουσιαστικά στο βήμα αυτό υπολογίζεται η απόσταση που έχει το σημείο που αντιστοιχεί σε κάθε εγγραφή από κάθε σημείο που αντιστοιχεί σε κάθε εγγραφή από κάθε σημείο που αντιστοιχεί σ' ένα αντιπροσωπευτικό δείγμα . Αφού καταταχθούν όλες οι εγγραφές σε κάποια ομάδα, για κάθε μια από αυτές επαναπροσδιορίζεται το αντιπροσωπευτικό δείγμα μπορεί να μην τις αντιπροσωπεύει πλήρως . Μ' αυτόν τον τρόπο προσπαθούμε να ελαχιστοποιήσουμε την 'εσωτερική ' ανομοιομορφία των ομάδων. Η ελαχιστοποίηση αυτή αντιστοιχεί στην ελαχιστοποίηση μίας συνάρτησης κόστους(cost function) . Στις διάφορες εκδοχές του k-means η συνάρτηση κόστους, μπορεί να έχει διαφορετική μορφή, αλλά σε όλες εκφράζει ουσιαστικά το πόσο ικανοποιητική είναι η ομαδοποίηση που πραγματοποιήθηκε μετρώντας τις ανομοιότητες που παρουσιάζονται μέσα στην κάθε ομάδα. Έτσι αφού προσδιοριστεί το νέο αντιπροσωπευτικό δείγμα για την κάθε ομάδα υπολογίζεται η συνάρτηση κόστους . Τα παραπάνω βήματα επαναλαμβάνονται μέχρι να διαπιστωθεί ότι η σύσταση των ομάδων δεν έχει αλλάξει σημαντικά από την προηγούμενη επανάληψη ή ότι η συνάρτηση κόστους δεν παρουσιάζει σημαντική μείωση.

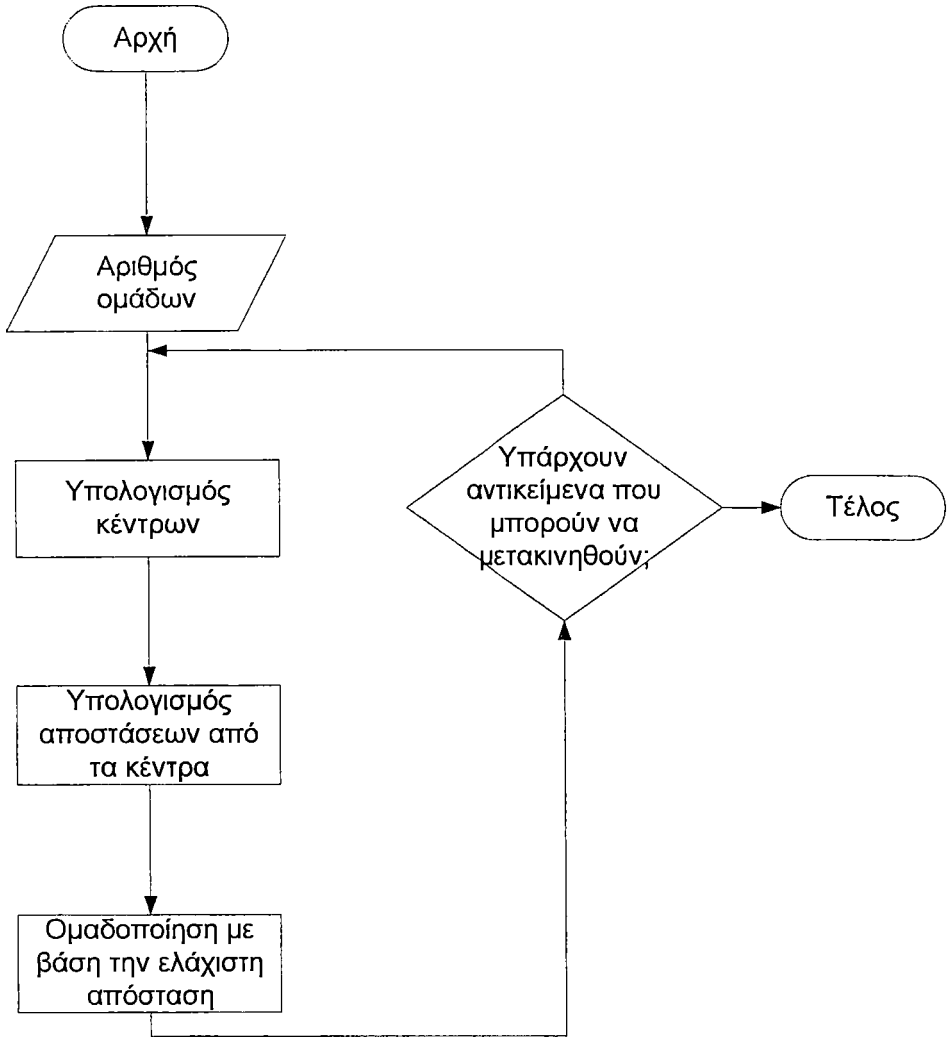
### 3.3.3 Τι είναι ο αλγόριθμος K-Means

Ο αλγόριθμος k-means είναι ένας αλγόριθμος κατηγοριοποίησης που ομαδοποιεί τα αντικείμενα σε k ομάδες με βάση τα χαρακτηριστικά τους. Το k είναι ένας θετικός ακέραιος. Η βασική ιδέα είναι να πραγματοποιηθεί κατηγοριοποίηση που επιτυγχάνει ελαχιστοποίηση των αποστάσεων μεταξύ των αντικειμένων και των κέντρων (centroids).

### 3.3.4 Ποιά είναι η σχηματική αναπαράσταση του αλγόριθμου K-Means

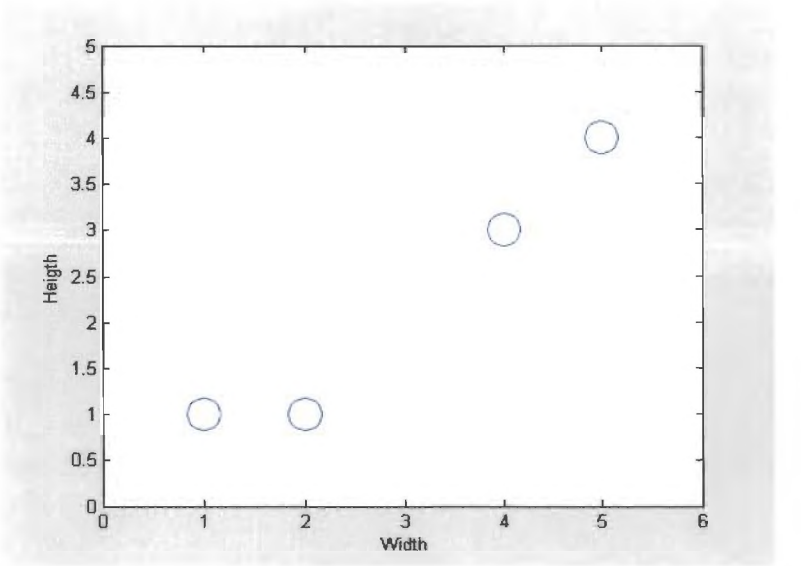
### Παράδειγμα που κατηγοριοποιεί τέσσερα ορθογώνια παραλληλόγραμμα με βάση τις διαστάσεις τους

Η σχηματική αναπαράσταση του αλγορίθμου είναι:



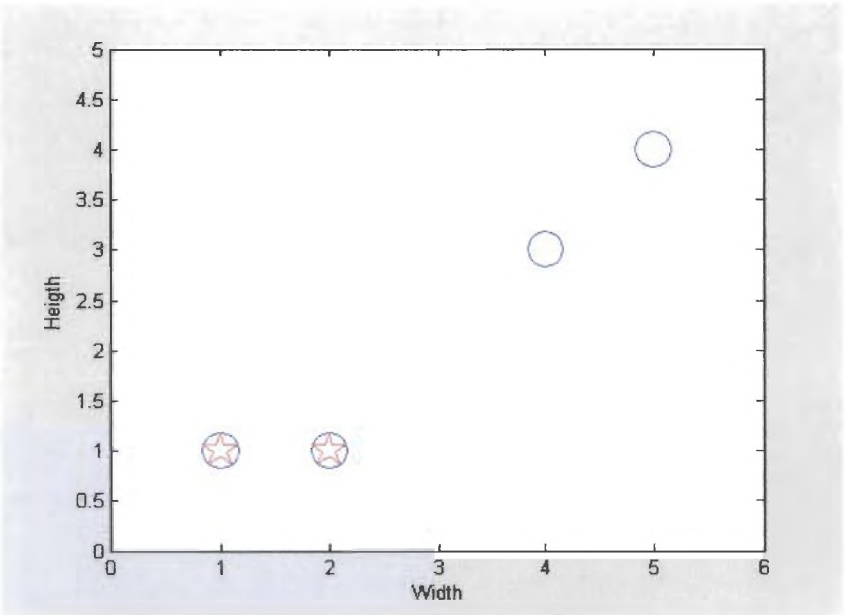

Θα παρουσιάσουμε ένα παράδειγμα που κατηγοριοποιεί τέσσερα ορθογώνια παραλληλόγραμμα με βάση τις διαστάσεις τους:

| ΑΝΤΙΚΕΙΜΕΝΑ | ΠΛΑΤΟΣ | ΜΗΚΟΣ |
|---|---|---|
| Παραλληλόγραμμο 1 | 1 | 1 |
| Παραλληλόγραμμο 2 | 2 | 1 |
| Παραλληλόγραμμο 3 | 4 | 3 |
| Παραλληλόγραμμο 4 | 5 | 4 |




1) Επιλογή των αρχικών κέντρων

Επιλέγουμε τα δύο πρώτα στοιχεία




C1 = (1,1) και C2=(2,1)

2) Υπολογίζουμε τις αποστάσεις από τα κέντρα

D=[ 0 1 3.61 5 C1

1 0 2.83 4.24] C2

1ο 2ο 3ο 4ο παραλληλόγραμμο

3) Υπολογίζουμε με βάση την ελάχιστη απόσταση σε ποια ομάδα ανήκει το κάθε αντικείμενο:

G = [ 0 0 0 0 ομάδα 2

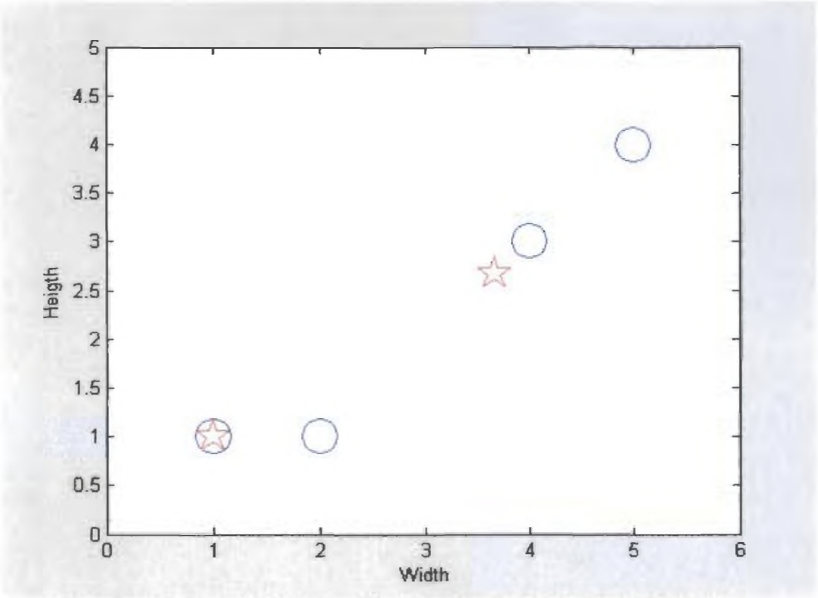
0 1 1 1] ομάδα 1

1ο 2ο 3ο 4ο παραλληλόγραμμο

Με 1 συμβολίζουμε το «ανήκει» και με 0 το «δεν ανήκει»

4) Υπολογίζουμε τα νέα κέντρα με τον τύπο της μέσης τιμής

Οπότε C1=(1,1) και C2=((2+4+5)/3, (1+3+4)/3) = (11/3, 8/3)




5) Υπολογίζουμε τις νέες αποστάσεις

D=[ 0 1 3.61 5 C1

3.14 2.36 0.47 1.89] C2

1ο 2ο 3ο 4ο παραλληλόγραμμο

6) Υπολογίζουμε με βάση την ελάχιστη απόσταση σε ποια ομάδα ανήκει το κάθε αντικείμενο:

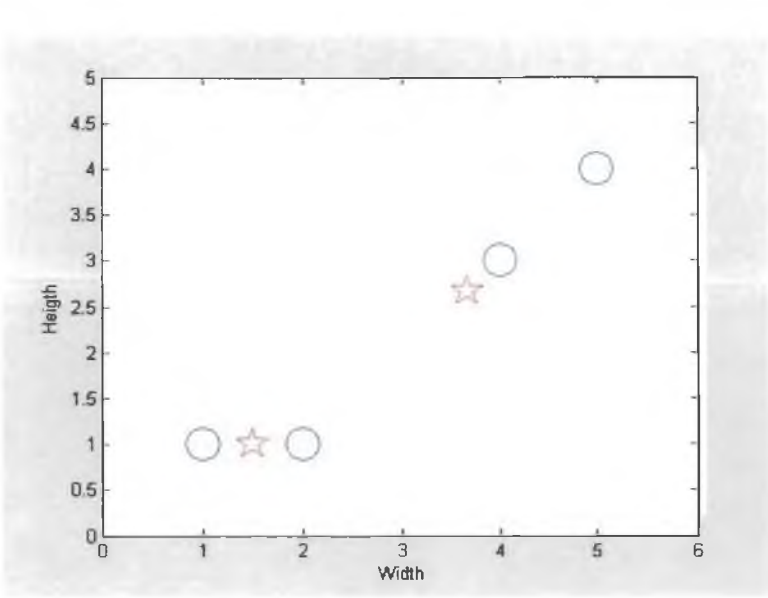
G = [ 1 1 0 0 ομάδα 2

0 0 1 1] ομάδα 1

1ο 2ο 3ο 4ο παραλληλόγραμμο

7) Υπολογίζουμε τα νέα κέντρα με τον τύπο της μέσης τιμής

Οπότε C1=((1+2)/2,(1+1)/2) και C2=((4+5)/2, (3+4)/2) = (9/2, 7/2)




8) Υπολογίζουμε τις νέες αποστάσεις

D=[ 0.5 0.5 3.20 4.61 C1

4.30 3.54 0.71 0.71] C2

1ο 2ο 3ο 4ο παραλληλόγραμμο

9) Υπολογίζουμε με βάση την ελάχιστη απόσταση σε ποια ομάδα ανήκει το κάθε αντικείμενο:

G = [ 1 1 0 0 ομάδα 2

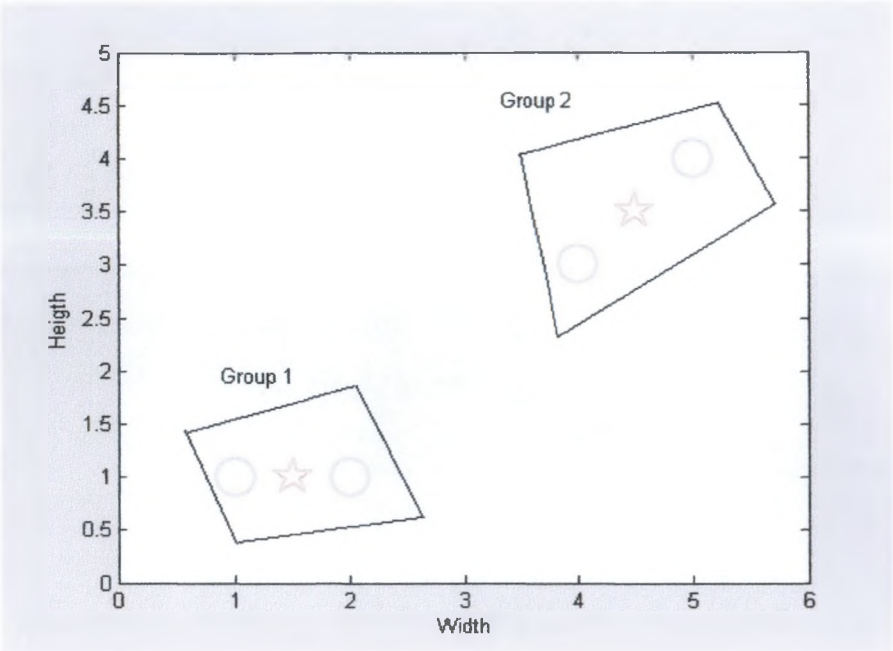
0 0 1 1] ομάδα 1

1ο 2ο 3ο 4ο παραλληλόγραμμο

10) Παρατηρούμε πως κανένα αντικείμενο δεν άλλαξε ομάδα οπότε τερματίζει ο αλγόριθμός με τις ομάδες :

Ομάδα 1 = {$1^{ο}$, $2^{ο}$ παραλληλόγραμμο}

Ομάδα 2 = {$3^{ο}$, $4^{ο}$ παραλληλόγραμμο}




# 4.ΣΥΣΧΕΤΙΣΗ

Μια από τις σημαντικότερες τεχνικές εξόρυξης από δεδομένα είναι η προτυποποίηση αλληλεξαρτήσεων(dependency modeling).Σκοπός της είναι η εύρεση των σημαντικών αλληλοεξαρτήσεων μεταξύ των διαφόρων πεδίων/χαρακτηριστικών του συνόλου εκπαίδευσης . Η πιο διαδεδομένη μέθοδος για παραγωγή αλληλοεξαρτήσεων είναι η εύρεση κανόνων συσχέτισης (association rules) . Το πρόβλημα της εξαγωγής κανόνων συσχέτισης παρουσιάστηκε αρχικά το 1993 ως μια προσπάθεια εξαγωγής χρήσιμων συσχετισμών μεταξύ των πεδίων μιας βάσης δεδομένων .

## 4.1 Εφαρμογή των κανόνων συσχέτισης

Η κλασική εφαρμογή των κανόνων συσχέτισης είναι η <<Ανάλυση του Καλαθιού της Νοικοκυράς >> (market basket analysis) .Σκοπός είναι να αναγνωριστούν τα προιόντα που αγοράζονται μαζί . Έστω για παράδειγμα ο παρακάτω πίνακας με τα δεδομένα που σε κάθε γραμμή του απεικονίζει τα προιόντα που αγοράσθηκαν σ΄ένα καλάθι, σε μια υπεραγορά(super market).

| | | | | |
|---|---|---|---|---|
| κόκα-κόλα | γάλα | τυρί | φρυγανιές | κρέας |
| τυρί | νερό | γάλα | κρέας | |
| κρέας | γάλα | ψωμί | | |
| .... | .... | .... | .... | .... |

Ένας κανόνας συσχέτισης θα μπορούσε να είναι ότι το γάλα πωλείται μαζί με το τυρί, με την προφανή αξιοποιήση της πληροφορίας που είναι η γειτνίαση του σημείου πώλησης γάλακτος μ΄αυτό του τυριού.

Μερικές συνήθεις πρακτικές εφαρμογές τους είναι : η εύρεση των προιόντων που πωλούνται μαζί σε μία συναλλαγή, η επεξεργασία ερωτηματολογίων, η εύρεση των προιόντων που διακινούνται μαζί σε μια αποθήκη για πρόβλεψη προμήθειας, η εύρεση ιστοσελίδων που επισκέπτεται μαζί ένας χρήστης, η εύρεση των λέξεων που συναντώται μαζί σ΄ένα κείμενο.

Ένας κανόνας συσχέτισης είναι μία έκφραση της μορφή X=>Y, όπου X και Y είναι σύνολα τιμών των πεδίων , όπως για παράδειγμα σύνολα προιόντων (items). Η σημασία ενός τέτοιου κανόνα είναι ότι οι περισσότερες εγγραφές του συνόλου εκπαίδευσης που περιέχουν το X περιέχουν και το Y .Πρακτικά, δηλαδή, οι κανόνες συσχέτισης έχουν στόχο την εύρεση συσχετίσεων μεταξύ των στηλών ενός πίνακα με δεδομένα . Αναφερόμενοι στο παραπάνω παράδειγμα, ένας κανόνας συσχέτισης θα μπορούσε να είναι <<το 98% των πελατών που αγοράζουν γάλα και κρέας ,αγοράζουν επίσης και τυρί>>.

## 4.2 Η σπουδαιότητα ενός κανόνα συσχέτισης

Η σπουδαιότητα ενός κανόνα συσχέτισης καθορίζεται αναλογικά από το ποσοστό εφαρμογής του κανόνα επί του συνόλου εκπαίδευσης .Συγκεκριμένα, οι αλγόριθμοι συσχέτισης που έχουν προταθεί και εφαρμόζονται πρακτικά, εξάγουν κανόνες συσχέτισης της μορφής :<<το 98% των πελατών που αγοράζουν γάλα και κρέας αγοράζουν επίσης και τυρί. Αλλά και στο 70% των αγορών έχουν αγοραστεί γάλα , κρέας και τυρί>>.Το πρώτο ποσοστό αναφέρεται ως αξιοπιστία (confidence) του κανόνα ενώ το δεύτερο ως επιβεβαίωση(support). Η επιβεβαίωση αφορά στο ποσοστό που εμφανίζονται και τα τρία προιόντα μαζί επί του όλου του συνόλου εκπαίδευσης ενώ η αξιοπιστία αφορά στο ποσοστό που εμφανίζονται και τα τρία προιόντα μαζί επί του αριθμού των αγορών που περιέχουν γάλα και κρέας .Το

πρόβλημα της εύρεσης κανόνων συσχέτισης εστιάζεται στην εύρεση όλων των κανόνων που έχουν μια καθορισμένη από τον χρήστη ελάχιστη τιμή επιβεβαίωσης και αξιοπιστίας .

## 4.3 Ο αλγόριθμος συσχέτισης Apriori

Ο αλγόριθμος Apriori παρουσιάστηκε αρχικά το 1994.Σχεδόν όλοι οι αλγόριθμοι συσχέτισης βασίζονται στην αρχική ιδέα του 1994. Σχεδον όλοι οι αλγόριθμοι συσχέτισης βασίζονται στην αρχική ιδέα του ενώ τον χρησιμοποιούν τα περισσότερα εμπορικά συστήματα εξόρυξης από δεδομένα .

Ο αλγόριθμος Apriori δέχεται σαν είσοδο ένα σύνολο αγορών (transactions) που αποτελεί και το σύνολο εκπαίδευσης . Κάθε αγορά είναι ουσιαστικά μία λίστα (itemset) από προιόντα (items) που αγοράστηκαν μαζί. Συγκεκριμένα , έστω I={ι1,ι2,....,im} ένα σύνολο από προιόντα . Εστω D, ένα σύνολο από αγορές , όπου κάθε αγορά T είναι μία λίστα από προιόντα όπου T I.

Γενικά, κάθε υποσύνολο από βάση για το επόμενο βήμα. Η αποτελεσματικότητα στην εύρεση των μεγάλων λιστών από προιόντα αποτελεί κριτήριο για την αποτελεσματικότητα συνολικά ενός αλγόριθμου εύρεσης κανόνων συσχέτισης, λόγω της μεγάλης πολυπλοκότητά της.

### 4.3.1 Τα βήματα του αλγορίθμου συσχέτισης Apriori και παράδειγμα εφαρμογής του

Βήματα του αλγορίθμου A priori:

1.Βρες τα προϊόντα που εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση, δηλαδή το σύνολο:

$L_1$ = μεγάλες λίστες από το 1-προϊόν

2.Από κ = 2 και για όσο το $L_{κ-1}$ δεν είναι κενό κάνε:

a. Βρες το σύνολο $C_κ$ των υποψηφίων μεγάλων λιστών από κ-προϊόντα με βάση το $L_{κ-1}$

b. Βρες ποια από αυτά εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση και φτιάξε το σύνολο:

$L_κ$ = μεγάλες λίστες από το κ-προϊόν

c.Για κάθε στοιχείο των $L_1$, $L_2$, ..., $L_n$, βρες ποια ικανοποιούν την ελάχιστη αξιοπιστία

## 4.3.2 Εφαρμογή του αλγόριθμου σε παράδειγμα με οκτώ καλάθια από ένα super market

Από οκτώ (8) καλάθια σε ένα super market προκύπτουν τα σύνολα:

- $B_1$ = { milk, coke, beer }
- $B_2$ = { milk, pepsi, juice }
- $B_3$ = { milk, beer }
- $B_4$ = { coke, juice }
- $B_5$ = { milk, pepsi, beer }
- $B_6$ = { milk, beer, juice, pepsi }
- $B_7$ = { coke, beer, juice }
- $B_8$ = { beer, pepsi }

Θεωρούμε ελάχιστη επιβεβαίωση το 35% και ελάχιστη αξιοπιστία το 50%.
Σύμφωνα με τα παραπάνω από το 1° βήμα προκύπτει (η επιβεβαίωση αυτών):

$L_1$ = { milk (62.5%), coke (37.5%), pepsi (50%), beer(75%), juice(50%) }

Όπου: milk = 5/8 = 62.5%, coke = 3/8 = 37.5%, pepsi = 4/8 = 50%, beer = 6/8 = 75%, juice = 4/8 = 50%.

Στην πρώτη επανάληψη του δευτέρου βήματος οι υποψήφιες λίστες με δύο προϊόντα είναι :

$L_2$ = { beer – milk (50%), beer – milk (37.5%), milk – pepsi (37.5%),
beer – juice (25%), milk – juice (25%), pepsi – juice (25%),
beer – coke (25%), juice – coke (25%)}.

Μετρώντας την επιβεβαίωση αυτών προκύπτει:

$L_1$ = { beer – milk , beer – pepsi, milk – pepsi }

Στην δεύτερη επανάληψη του δευτέρου βήματος οι υποψήφιες μεγάλες λίστες με τρία προϊόντα είναι :

$C_2$ = { beer – milk – pepsi (25%) }

Από την επιβεβαίωσή της προκύπτει:

$L_3$ = { } → κενό και οι επαναλήψεις του δευτέρου βήματος σταματούν.

Στο τρίτο βήμα από όλους τους δυνατούς κανόνες που προκύπτουν από την $L_2$:

- Beer → milk
- Beer → pepsi
- Milk → pepsi

Ο πρώτος κανόνας έχει αξιοπιστία 50%

Ο δεύτερος κανόνας έχει αξιοπιστία 37.5%

Ο τρίτος κανόνας έχει αξιοπιστία 50%

Σύμφωνα με τα παραπάνω εξετάζουμε μόνο τον κανόνα συσχέτισης beer → milk

Βιβλιογραφία


- [Agrawal] Agrawal, R, Mannila, H., Srikant, R., and Verkamo, A.I., "Fast Discovery of Association Rules",in Fayyad, U.M.,Piatetsky-Shapiro, G., and Smyth, p.,(eds), Advances in Knowledge Discovery and Data Mining ,AAAI Press/MIT Press, pp. 307-328,1996.
- [Boutsinas01 ] Boutsinas, b., and Vrahatis , M.N., "Artificial Nonmonotonic Neurai Networks",Artificial Intelligence 132(1), Elsevier Science Publishers B.V., ppc 1-38, 2001.
- [Boutsinas02a] Boutsinas,B., "Accessing Data Mining Rules though Expert Sustems", International Journal of Information Technology and Decision Making,1(4),World Scientific , pp. 657-672,2002.

- [Clark] Clark, P., and Niblett, T."The CN2 induction algorithms", Machine Learning ,3, pp.261-283, 1989.
- [Cohn]Cohn D.L. and Melsa J.L.,''Decision and Estimation Theory'',McGraw-Hill, 1978.
- [Draper] Draper N., Smith H., "Εφαρμοσμένη Ανάλυση Παλινδρόμησης",2η Αγγλική Έκδοση, μετάφραση Ε. Χατζηκωνσταντινίδης , Α. Καλαματιανού ,Εκδόσεις Παπαζήση,1997.
- [Fayyad] Fayyad, U.M. , Piatetsky-Shapiro, G., and Smyth, P.,(eds.),''Advances in Knowledge Discovery and Data Mining '',AAAI press/MIT press, pp 307-328,1996.
- [Gallant] Gallant S.I. , "Neural Network Learning and Expert Systems", MIT press, 1993.
- [Huang] Huang, Z., "Extensions to the k-means algorithm for clustering large data sets with categorigal values",Data mining and Knowledge Discovery , 2, pp.283-304,1998.
- [Jain] jain , A.K., and Dubes, R.C., "Algorithms for Clustering Data" Prentice-Hall, Englewoods Cliffs, NJ., 1988.
- [Kleinbaum] Kleinbaum D., Kupper L.,Muller K., Nizam A., "Applied Regrassion Analysis and Other Multivariate Methods", 3rd Edition, Duxbury Press, 1998.
- [Lucas] Lucas P., van der Gaag L., "Principles of Expert Systems",Addison-Wesley Publishers Ltd. Press, 1991.
- [Michalski] Michalski R.S., Carbonell J.G.., Mitchell T.M., "Machine Learning",Springer-Verlag,1984.
- [Piatetsy-Shapiro] Piatetsy-Shapiro, G., Brachman, R., Khabaza, T., Kloesgen W. and Simoudis, E., "An overview of issues in developing industrial data mining and knowledge discovery applications", Proceedings of the 2nd Int. Conf on Knowledge Discovery and Data Mining , AAAI Press,pp. 89-95,1996.
- [Porter] Porter E.P., "Tax expert systems and future development.(The CPA &the Computer)", The CPA journal Online, jan 1994.