ΤΕΙ ΠΑΤΡΩΝ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗ&ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΥ ΣΧΕΔΙΑΣΜΟΥ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

# «Η ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΝΤΙΣΤΡΟΦΩΣ, ΜΙΑ ΑΝΑΣΚΟΠΗΣΗ»

**ΕΠΙΜΕΛΕΙΑ & ΣΥΝΤΑΞΗ**

**ΝΙΚΟΛΑΟΥ ΛΟΥΝΤΟΒΙΚ**
**ΠΑΠΑΘΕΟΧΑΡΗΣ ΘΕΟΧΑΡΗΣ**

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:** ΝΙΚΟΛΑΟΣ ΜΑΣΤΡΟΓΙΑΝΝΗΣ

**Πάτρα, Δεκέμβριος 2011**

**ΤΕΙ ΠΑΤΡΩΝ**
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΥ ΣΧΕΔΙΑΣΜΟΥ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:**ΝΙΚΟΛΑΟΣ ΜΑΣΤΡΟΓΙΑΝΝΗΣ

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

# «Η ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΝΤΙΣΤΡΟΦΩΣ, ΜΙΑ ΑΝΑΣΚΟΠΗΣΗ»

**ΕΠΙΜΕΛΕΙΑ & ΣΥΝΤΑΞΗ**

**ΝΙΚΟΛΑΟΥ ΛΟΥΝΤΟΒΙΚ**
**ΠΑΠΑΘΕΟΧΑΡΗΣ ΘΕΟΧΑΡΗΣ**

**Πάτρα,Δεκέμβριος2011**

# ΕΥΧΑΡΙΣΤΙΕΣ

Τις θερμές μας ευχαριστίες στον καθηγητή μας Κύριο Νικόλαο Μαστρογιάννη για τη συμβολή του στην συγγραφή της πτυχιακής μας καθώς και τις συμβουλές που μας προσέφερε οποτεδήποτε ζητήθηκαν ώστε να μπορέσουμε να διεκπεραιώσουμε την παρούσα πτυχιακή εργασία.

# ΠΡΟΛΟΓΟΣ

Η παρούσα πτυχιακή εργασία αποτελεί μέρος μιας σύγχρονης προσέγγισης χρήσιμη για τον κόσμο των επιχειρήσεων αλλά και της επιστημονικής κοινότητας που ασχολείται με την επιχειρησιακή έρευνα αλλά και γενικότερα με την διευκόλυνση και απλούστευση σχέσεων- διεργασιών, πόρων αλλά και βέλτιστων λύσεων-εξόρυξης και επεξεργασίας δεδομένων.

Σ' έναν κόσμο όλο και πιο «*διασυνδεδεμένο*» με αυξανόμενο όγκο και ταχύτητα μετάδοσης πληροφοριών, σχέσεων μεταξύ ανθρώπινων στελεχών αλλά και ανθρώπων με «*συστήματα*» είναι πολύ σημαντικό τα πάντα να γίνονται **απλά γρήγορα εύχρηστα και σωστά**. Μία προσέγγιση για το πώς μπορούν να γίνουν όλα αυτά με τη βοήθεια της Επιχειρησιακής Έρευνας σε συνδυασμό με την Εξόρυξη δεδομένων θα δούμε παρακάτω.

# Περιεχόμενα

# Κεφάλαιο 1$^{ο}$ - Εισαγωγή

Τα τελευταία χρόνια, στον τομέα της εξόρυξης δεδομένων υπάρχει τρομερό ενδιαφέρον τόσο από τα πανεπιστήμια όσο και από την βιομηχανία.

Το ενδιαφέρον αυτό προκύπτει από το γεγονός ότι η συλλογή και η αποθήκευση των δεδομένων έχει γίνει ευκολότερη και λιγότερο ακριβή, έτσι λοιπόν οι βάσεις δεδομένων είναι πάρα πολλές και υπερφορτωμένες στις σύγχρονες επιχειρήσεις. Αυτό ισχύει ιδιαίτερα σε συστήματα που βασίζονται στο διαδίκτυο και κατά συνέπεια δεν αποτελεί έκπληξη το γεγονός πως η εξόρυξη δεδομένων είναι ιδιαιτέρως χρήσιμη σε τομείς που έχουν σχέση με τις ηλεκτρονικές υπηρεσίες.

Αυτές οι μαζικές βάσεις δεδομένων περιέχουν ένα πλούτο σημαντικών δεδομένων που οι παραδοσιακές μέθοδοι ανάλυσης αποτυγχάνουν να μετατρέψουν σε αξιοποιήσιμη γνώση. Συγκεκριμένα, σημαντική γνώση είναι συχνά κρυμμένη και απροσπέλαστη και οι μέθοδοι με βάση τις υποθέσεις, όπως η online αναλυτική επεξεργασία (OLAP) και οι περισσότερες στατιστικές μέθοδοι, γενικά αποτυγχάνουν να αποκαλύψουν αυτές τις γνώσεις.

Οι επαγωγικές μέθοδοι, με τις οποίες μαθαίνουμε άμεσα από τα δεδομένα χωρίς μια εκ των προτέρων υπόθεση (a priori hypothesis), πρέπει, να χρησιμοποιούνται για να αποκαλυφθεί το κρυμμένα μοντέλο (pattern) και η γνώση. Χρησιμοποιούμε τον όρο εξόρυξη δεδομένων για να αναφερθούμε σε όλες τις πτυχές μιας αυτοματοποιημένης ή ημι-αυτοματοποιημένης διαδικασίας για την εξαγωγή προηγουμένως άγνωστης και ενδεχομένως χρήσιμης γνώσης και μοντέλων που προέρχονται από μεγάλες βάσεις δεδομένων.

Η διαδικασία αυτή αποτελείται από πολυάριθμα βήματα όπως είναι η ενσωμάτωση δεδομένων από πολλές βάσεις δεδομένων, η προ-επεξεργασία των δεδομένων, καθώς και η εισαγωγή ενός μοντέλου με αλγόριθμο. Το μοντέλο χρησιμοποιείται στη συνέχεια για να τον προσδιορισμό και την εφαρμογή ενεργειών που λαμβάνουν χώρα στο πλαίσιο της επιχείρησης. Η εξόρυξη δεδομένων αντλεί, παραδοσιακά σε μεγάλο βαθμό, τόσο από την στατιστική όσο και από την μηχανική μάθηση, **αλλά πολλά προβλήματα στην εξόρυξη δεδομένων μπορούν να διατυπωθούν ως προβλήματα βελτιστοποίησης**.

Η ερευνητική κοινότητα έχει συμβάλλει σημαντικά στον τομέα της εξόρυξης δεδομένων και ειδικότερα στον σχεδιασμό και την ανάλυση των αλγορίθμων των δεδομένων εξόρυξης. Η αρχική συμβολή περιελάμβανε τη χρήση μαθηματικού προγραμματισμού τόσο για τις ταξινομήσεις όσο και για τις συσταδοποιήσεις και η αυξανόμενη δημοτικότητα της εξόρυξης δεδομένων έχει προκαλέσει μια σχετική αύξηση του ενδιαφέροντος στον τομέα αυτό. Υπάρχει μαθηματικός προγραμματισμός για ένα μεγάλο φάσμα προβλημάτων εξόρυξης δεδομένων που περιλαμβάνουν την επιλογή των χαρακτηριστικών, την ταξινόμηση και την συσταδοποίηση των δεδομένων.

Για την επίλυση των προβλημάτων των δεδομένων εξόρυξης χρησιμοποιείται η μεταευρετική. Για παράδειγμα, η επιλογή χαρακτηριστικού έχει γίνει χρησιμοποιώντας προσομοιωμένη ανόπτηση ή αναδιατάξη (simulated annealing) γενετικούς αλγόριθμους και την μέθοδο nested partitions.

Ωστόσο, η τμηματοποίηση της και η εξόρυξη δεδομένων δεν περιορίζεται στον σχεδιασμό αλγορίθμων και η εξόρυξη δεδομένων μπορεί να παίξει σημαντικό ρόλο σε πολλές εφαρμογές επιχειρησιακών δεδομένων. Τεράστιες ποσότητες δεδομένων δημιουργούνται τόσο σε παραδοσιακά πεδία εφαρμογής, όπως είναι ο σχεδιασμός της παραγωγής, καθώς και σε νεότερους τομείς όπως είναι η διαχείριση πελατειακών σχέσεων και η εξατομίκευση.

Η **εξόρυξη δεδομένων** όπως και τα παραδοσιακά εργαλεία **επιχειρησιακής έρευνας** μπορούν να χρησιμοποιηθούν για την καλύτερη αντιμετώπιση τέτοιου είδους προβλημάτων. Η παρούσα πτυχιακή εργασία, παρουσιάζει μια έρευνα της επιχειρησιακής έρευνας και της εξόρυξης δεδομένων, εστιάζοντας στις παραπάνω διασταυρώσεις. Η αναφορά της χρήσης των τεχνικών επιχειρησιακής έρευνας στην εξόρυξη δεδομένων εστιάζει στο **πώς τα πολυάριθμα προβλήματα της εξόρυξης δεδομένων μπορούν να διατυπωθούν και να λυθούν ως προβλήματα βελτιστοποίησης**.

Αυτό γίνεται χρησιμοποιώντας μια σειρά από μεθοδολογίες βελτιστοποίησης, συμπεριλαμβανομένων τόσο της μεταευρετικής όσο και του μαθηματικού προγραμματισμού. Το τμήμα εφαρμογών αυτής της εργασίας εστιάζει την προσοχή του σε ένα συγκεκριμένο τύπο εφαρμογών, δηλαδή δύο τομείς που σχετίζονται με ηλεκτρονικές υπηρεσίες: την διαχείριση πελατειακών σχέσεων και την εξατομίκευση.

Οι πρωταρχικοί στόχοι του κειμένου είναι να περιγράψει περιληπτικά τι είναι, πού χρησιμεύουν, ποιοι είναι οι τομείς και οι λειτουργίες της εξόρυξη δεδομένων και της επιχειρησιακής έρευνας, να καταδείξει το εύρος των διασταυρώσεων των δύο πεδίων να δώσει κάποια λεπτομερή παραδείγματα της έρευνας που πιστεύουμε ότι δείχνει ότι η συνέργεια λειτουργεί καλά, να υπάρξουν αναφορές σε άλλα σημαντικά έργα και τέλος, να προταθούν κάποιες κατευθύνσεις για μελλοντική έρευνα στον τομέα αυτό.

# Κεφάλαιο 2$^{ο}$ - Εξόρυξη δεδομένων

Στις μέρες μας όπου η πληροφορία κυριαρχεί στον κόσμο των επιχειρήσεων αλλά και στην επιστημονική κοινότητα είναι ιδιαιτέρως αναγκαία και απαραίτητη η ύπαρξη ενός εργαλείου για την ανάλυση και ερμηνεία τεραστίων ποσοτήτων δεδομένων που είναι καταχωρημένα σε αρχεία, βάσεις δεδομένων και άλλα μέσα αποθήκευσης, με σκοπό την εξαγωγή της γνώσης που θα προωθήσει την ουσιαστική και ασταμάτητη διαδικασία λήψης αποφάσεων, σε μια σειρά από προβλήματα της καθημερινότητας.

## 2.1 Ορισμοί εξόρυξης δεδομένων

Καλούμε Εξόρυξη Δεδομένων (Data mining) την εξεύρεση (σημαντικών, άγνωστων και πιθανόν χρήσιμων) *πληροφοριών ή επαναλαμβανόμενων Προτύπων* (patterns) σε τεράστιες βάσεις 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).

- **Τι δεν είναι η εξόρυξη δεδομένων**

  - Αναζήτηση ενός αριθμού τηλεφώνου στον χρυσό οδηγό
  - Ερώτηση σε μια μηχανή αναζήτησης πληροφορία για το “teipatras”

- **Τι είναι η εξόρυξη δεδομένων**

  - Είναι αποδοτικές τεχνικές για να την ανάλυση πολύ μεγάλων συλλογών από δεδομένα και να η εξαγωγή χρήσιμων πληροφοριών από αυτά. Κάποια επώνυμα εμφανίζονται πιο συχνά σε κάποιες περιοχές στην Ελλάδα. (π.χ. Πετρίδης, Δημοκίδης,… στην Βόρεια Ελλάδα επίσης η κατάληξη «ακης» στην Κρήτη).

Ομαδοποίηση ομοίων κειμένων που εμφανίζει μια μηχανή αναζήτησης με βάση τα συμφραζόμενα (π.χ. σπουδέςPATRA, teipat.gr)

## 2.2 Γιατί εξόρυξη;

Κάθε χρόνο ενώ τα δεδομένα διπλασιάζονται, οι χρήσιμες πληροφορίες γίνονται όλο και πιο δυσεύρετες. Αυτό είναι και το βασικό πρόβλημα που θέλει να επιλύσει ο τομέας αυτός και αποτελεί πρόκληση στην εποχή μας.

**Συνεπώς, το πρόβλημα είναι ότι ενώ υπάρχει τεράστιος όγκος πληροφοριών δεν έχουμε την απαιτούμενη γνώση**.

Συνήθως υπάρχουν πληροφορίες στα δεδομένα που δεν είναι προφανές, είναι «κρυμμένες».

Οι ειδικοί που τις αναλύουν μπορεί να χρειαστούν μήνες για να ανακαλύψουν χρήσιμες πληροφορίες επίσης πολλά δεδομένα δεν αναλύονται ποτέ. Ακόμα, η υπολογιστική ισχύ των υπολογιστών αυξάνεται με αρκετά μικρότερο ρυθμό σε σχέση με την χωρητικότητα δεδομένων.

Έτσι δημιουργείται είναι μια απόσταση στις δύο αυτές τάσεις η οποία αυξάνεται εκθετικά και καλείται κενό δεδομένων (data gap) ή νόμος της αποθήκευσης (storage law).

Γεγονός είναι πως το κενό μεταξύ της απόδοσης του υλικού και της ποσότητας των δεδομένων που θέλουμε να επεξεργαστούμε αποτελεί ένα σημαντικό πρόβλημα.

Άλλο ένα σημαντικό ζήτημα ήταν το γεγονός πως στη δεκαετία του ‘90 τα δεδομένα προερχόμενα από επιχειρήσεις και επιστημονικούς οργανισμούς αυξάνονταν ραγδαία. Νέες βάσεις δεδομένων και συστήματα αποθήκευσης, μαζί με συστήματα συλλογής δεδομένων άρχισαν να συγκεντρώνουν όλο και περισσότερα δεδομένα μέρα με τη μέρα. Γι’ αυτό το λόγο έγινε επιτακτική η ανάγκη για ένα μοντέλο ώστε να περάσουμε από τα **απλά δεδομένα στη χρήσιμη πληροφορία.**

Τρόπους αντιμετώπισης μας δίνει η αποθήκευση δεδομένων και Εξόρυξη δεδομένων: Η αποθήκη πληροφοριών και η απευθείας αναλυτική επεξεργασία. Αποτελεί εξαγωγή χρήσιμης γνώσης (επαναλαμβανόμενα σχέδια, κανόνες, κανονικότητα, περιορισμούς από μεγάλες βάσεις δεδομένων. Ο λόγος για τον οποίο χρησιμοποιούμε την Εξόρυξη Δεδομένων είναι για να αναλύουμε βάσεις δεδομένων και να υποβοηθούμε στη λήψη αποφάσεων.

**Γιατί Εξόρυξη Δεδομένων (από εμπορική πλευρά)**

Πολλά δεδομένα συγκεντρώνονται και εισάγονται σε αποθήκες δεδομένων:

- Δεδομένα ιστού του διαδικτύου, e-εμπόριο
- Συναλλαγές με Τραπεζικά συστήματα/πιστωτικές κάρτες
- Αγορές σε πολυκαταστήματα/αλυσίδες
- Οι υπολογιστές γίνονται όλο και φτηνότεροι, όλο και πιο ισχυροί
- Μεγάλος ανταγωνισμός
- Παροχή καλύτερων, προσωπικών υπηρεσιών σε κάποιο πεδίο (fraud detection , targeting marketing)

**Γιατί Εξόρυξη Δεδομένων (από επιστημονική πλευρά)**

Τα δεδομένα συλλέγονται και αποθηκεύονται σε τρομερές ταχύτητες(GB/hour)

- Απομακρυσμένοι αισθητήρες(remote sensors) σε δορυφόρους
- Τηλεσκόπια στον ουρανό Microarrays που παράγουν γονιδιακά δεδομένα
- Επιστημονικές προσομοιώσεις που παράγουν terabytes δεδομένων
- Η εξόρυξη δεδομένων μπορεί να βοηθήσει τους επιστήμονες: Στην κατηγοριοποίηση και την τμηματοποίηση των δεδομένων και στην Διατύπωση Υποθέσεων

## 2.3 Ανακάλυψη γνώσης από βάσεις δεδομένων (KDD διαδικασία)

### 2.3.1 Ορισμός

« *Η KDD διαδικασία αποτελεί μια ντετερμινιστική επαναληπτική διαδικασία αναγνώρισης καινοτόμων ,έγκυρων, ενδεχομένως κατανοητών και χρήσιμων προτύπων στα δεδομένα που εξετάζονται.*»(A. Gionis, H. Mannila, T. Mielikainen, and P. Tsaparas, Assessing Data Mining Results via Swap Randomization, ACM Transactions on Knowledge Discovery from Data (TKDD)).

- Αποθήκη Δεδομένων

Η διαδικασία αυτή, ξεκινά από την αποθήκη των δεδομένων από όπου και διαχωρίζουμε τα δεδομένα που μας ενδιαφέρουν και που πρόκειται να αναλύσουμε.

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

## 2.3.2 Προ-επεξεργασία Δεδομένων

Η προ-επεξεργασία των δεδομένων περιλαμβάνει βασικές διαδικασίες όπως είναι η αφαίρεση του θορύβου και των outliers από τα δεδομένα και η συλλογή των απαραίτητων πληροφοριών για τη διαμόρφωση ή τη μέτρηση του θορύβου.

**Ανακάλυψη γνώσης κατά τη προ-επεξεργασία**

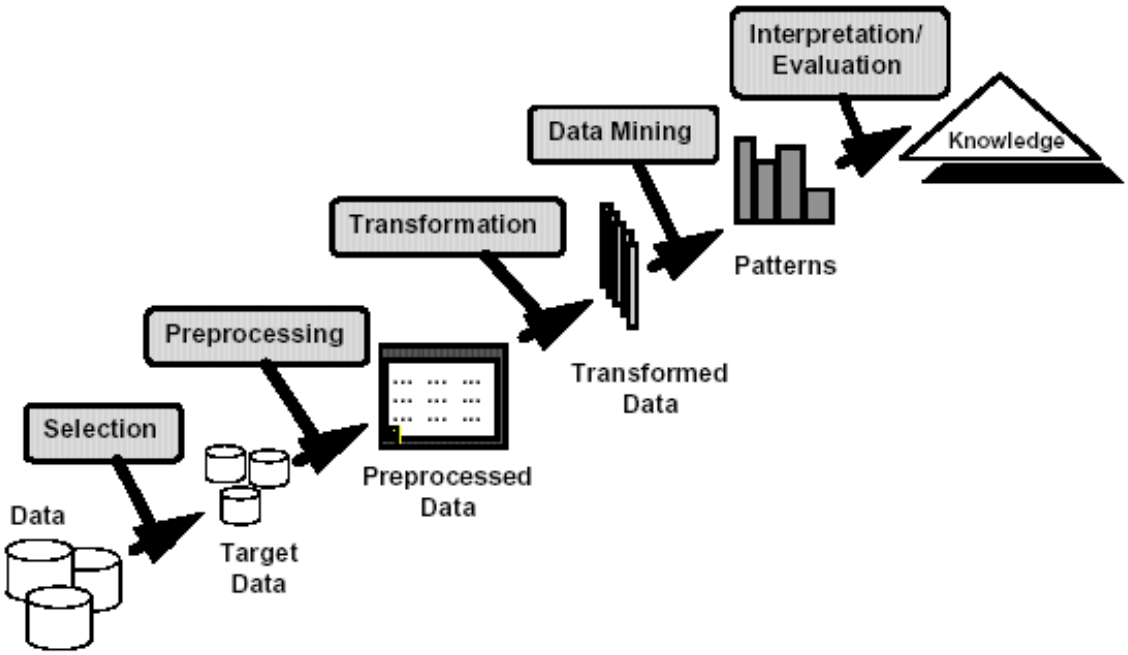
Η KDD διαδικασία είναι μια επαναληπτική και διαλογική που περιέχει μια σειρά από βήματα:

- Data Cleaning – Καθαρισμός Δεδομένων
- Data Integration – Ενοποίηση Δεδομένων
- Data Transformation – Μετασχηματισμοί Δεδομένων

Τα δεδομένα στο πραγματικό κόσμο είναι συνήθως «βρώμικα» ( incomplete): ενδέχεται να απουσιάζουν κάποιες τιμές γνωρισμάτων (να μην καταγράφηκαν, να καταγράφηκαν με λάθη λόγω κακής συνεννόησης ή λανθασμένης λειτουργίας), να απουσιάζουν κάποια αξιόλογα γνωρίσματα (που να μην θεωρήθηκαν αρκετά σημαντικά ή να μην ήταν διαθέσιμα), ή να έχουν μόνο συναθροιστικά (aggregate) δεδομένα.

Για να συμπληρώσουμε γνωρίσματα και τιμές που λείπουν με θόρυβο - noisy: Περιέχουν λάθη ή outliers (περιθωριακές τιμές - τιμές που διαφέρουν πολύ από την πλειοψηφία)

- Εύρεση των περιθωριακών τιμών και απομάκρυνση θορύβου.
- Ασυνεπή - inconsistent: περιέχουν ασυνέπειες, διπλότιμα
- Διόρθωση ασυνεπών τιμών




***Πως φτάνουμε από τα δεδομένα στη γνώση ( βήμα- βήμα)***

### 2.3.3 Μέθοδοι της προ-επεξεργασίας δεδομένων

#### 1. Καθαρισμός των δεδομένων:

Μερικά από τα δεδομένα που εξετάζονται πολλές φορές μπορεί να μην υπάρχουν ή ορισμένες από τις τιμές που συλλέγονται να μην είναι συμβατές με άλλες οπότε έτσι διαγράφονται αυτομάτως, ή ακόμη να μην έχουν καταγραφεί δεδομένα λόγω κακής συνεννόησης μεταξύ των υπευθύνων μιας εταιρίας.

Επίσης είναι πιθανό σε κάποια από τα δεδομένα που έχουμε να εντοπίζεται *θόρυβος* δηλαδή το τυχαίο λάθος ή η απόκλιση από μία μεταβλητή.

#### 2. Ενσωμάτωση δεδομένων (Data integration):

Στο αυτό το βήμα, τα δεδομένα από διαφορετικές πηγές (σε αρκετές περιπτώσεις ανομοιογενή), προστίθενται σε μια βάση δεδομένων.

#### 3. Επιλογή δεδομένων (Data selection):

Από όλα τα διαθέσιμα δεδομένα, επιλέγονται εκείνα που σχετίζονται με την ανάλυση που έπεται.

### 4. Τροποποίηση δεδομένων (Data transformation):

Τα δεδομένα που έχουν επιλεχθεί τροποποιούνται έτσι ώστε η μορφή τους να είναι προσαρμοσμένη για την διαδικασία της εξόρυξης.

### 5. Εξόρυξη δεδομένων (Data Mining):

Αποτελεί το σημαντικότερο από τα βήματα της διαδικασίας, εδώ διάφορες εξελιγμένες τεχνικές εφαρμόζονται για την εξαγωγή δυνητικά χρήσιμων προτύπων.

### 6. Αξιολόγηση προτύπων (Pattern evaluation):

Σε αυτό το βήμα, εμφανίζονται χρήσιμα πρότυπα που αναπαριστούν γνώση, βάσει μέτρων αξιολόγησης(*evaluation measures*).

### 7. Αναπαράσταση γνώσης (Knowledge representation):

Στο τελικό βήμα, η γνώση που εξορύχτηκε, εμφανίζεται στον χρήστη, δίνοντάς του τη δυνατότητα να ερμηνεύσει και να κατανοήσει τα αποτελέσματα της εξόρυξης δεδομένων.

## 2.3.4 Μοντελοποίηση Εξόρυξης

Όταν καταλήξουμε στον στόχο της διαδικασίας KDD επιλέγουμε τα απαιτούμενα μοντέλα και τους απαιτούμενους αλγόριθμους Εξόρυξης Δεδομένων, όπως και τη μετατροπή των δεδομένων (εφόσον χρειάζεται για το μοντέλο ) που θα εφαρμοστούν στην εξόρυξη των δεδομένων μας ώστε να καταλήξουμε σε χρήσιμα συμπεράσματα και έγκυρες αναλύσεις.

*Εξόρυξη δεδομένων*

Με τη χρήση κατάλληλων μεθόδων ερευνούμαι για τα πρότυπα με τα οποία θα ασχοληθούμε. Τα πρότυπα θα πρέπει να είναι συγκεκριμένης μορφής όπως κανόνες συσχέτισης, ταξινόμηση, δέντρα αποφάσεων, παλινδρόμηση, ομαδοποίηση κ.λπ. Η βέλτιστη απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από αυτά τα βήματα.

## 2.3.5 Αξιολόγηση των προτύπων

Αξιολογούμε τα εξαγόμενα πρότυπα με κάποια μέτρα, έτσι ώστε να προσδιοριστούν τα πρότυπα τα οποία αντιπροσωπεύουν τη γνώση, δηλαδή τα πραγματικά σημαντικά για μας πρότυπα. Η αξιολόγηση του μοντέλου έχει να κάνει και με την εγκυρότητα των προτύπων, την μέτρηση της ακρίβειας, της χρησιμότητας και του τρόπου κατανόησης του μοντέλου.

### 2.3.6 Σταθεροποίηση και παρουσίαση της γνώσης

Η εξορυγμένη γνώση ενσωματώνεται στο σύστημα και εφαρμόζονται διάφορες τεχνικές αντιπροσώπευσης γνώσης και εμφανίζουν την εξορυγμένη γνώση στους χρήστες. Τέλος, ελέγχουμε για επίλυση τυχόν επιπλοκών με παλιότερη εξορυγμένη γνώση .

## 2.4 Εφαρμογές εξόρυξης δεδομένων

### 2.4.1.Εξόρυξη στο διαδίκτυο

Είναι από τα πιο ενδιαφέροντα ερευνητικά πεδία του τομέα της εξόρυξης των δεδομένων.

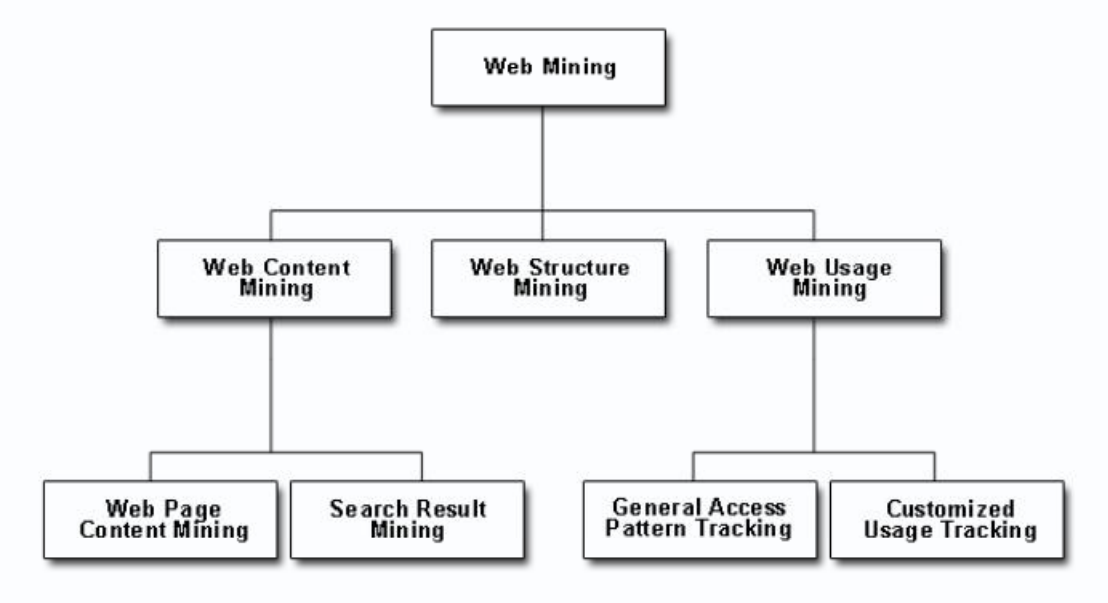
Όπως είναι γνωστό το να υπολογίσουμε το ακριβές μέγεθος δεδομένων του παγκόσμιου ιστού δεν είναι δυνατό. Τον Ιούνιο του 2011 έχει υπολογιστεί πως υπάρχουν περίπου 346,004,403σελίδες με ρυθμό αύξησης τεσσάρων περίπου 4,5 εκατομμυρίων σελίδων το μήνα.*(news.netcraft.com)* Η δημοφιλής μηχανή αναζήτησης Yahoo είχε ανακοινώσει πρόσφατα μέσα από την σελίδα της πως έχει στο ευρετήριο της περίπου 20 εκατομμύρια αντικείμενα από τα οποία τα 19 εκατομμύρια είναι δεδομένων κειμένου. Ο παγκόσμιος ιστός είναι πια η μεγαλύτερη και σημαντικότερη βάση δεδομένων που είναι ελεύθερη και διαθέσιμη στον καθένα και καθημερινά αντιμετωπίζει προκλήσεις τόσο σε θέματα σχεδιασμού-design όσο και πραγματικής ποιότητας δεδομένων.

Ο όρος βάση δεδομένων εδώ χρησιμοποιείται θεωρητικά μιας και δεν υπάρχει συγκεκριμένη δομή ή σχήμα σε αυτό που αποκαλούμε παγκόσμιο ιστό. Αυτό καθιστά περισσότερο σημαντική την ανάγκη για εξόρυξη δεδομένων στον παγκόσμιο ιστό δίνοντας ευχρηστία σε κάθε άνθρωπο. Με τον όρο εξόρυξη γνώσης στο διαδίκτυο δεν αναφερόμαστε μόνο σε δεδομένα που περιέχονται σε ιστοσελίδες αλλά και σε δεδομένα που σχετίζονται με τη δράση ενός χρήστη σε αυτό. Τα δεδομένα διαδικτύου μπορούν να χωριστούν στις ακόλουθες κατηγορίες:

- Περιεχόμενο των ιστοσελίδων.
- Ενδοπληροφορία ιστοσελίδων (HTML/XML κώδικας)
- Διασύνδεση Ιστοσελίδων
- Δεδομένα χρήσης που περιγράφουν πως οι επισκέπτες προσπελαύνουν τις ιστοσελίδες.

- Προφίλ χρηστών που περιλαμβάνουν δημογραφικά δεδομένα και πληροφορίες εγγραφών (εδώ περιέχονται και πληροφορίες από cookies αρχεία).

Στο παρακάτω σχήμα βλέπουμε μια απεικόνιση της εξόρυξης γνώσης από τον παγκόσμιο ιστό.




Εξόρυξη Δεδομένων Χρήσης

Εξόρυξη αρχείων καταγραφής Web, για να ανακαλύψουν μορφές πρόσβασης των χρηστών στις σελίδες του Ιστού. (Mining Web log records to discover user access patterns of Web pages)

Εφαρμογές (Applications)

- Στόχευση πιθανών πελατών για ηλεκτρονικό εμπόριο (Target potential customers for electronic commerce)
- Βελτίωση της ποιότητας και της παροχής υπηρεσιών πληροφόρησης στο Διαδίκτυο προς τον τελικό χρήστη. (Enhance the quality and delivery of Internet information services to the end user)
- Βελτίωση της απόδοσης του διακομιστή του ιστού. (Improve Web server system performance) Αναγνώριση πιθανών προνομιούχων θέσεων διαφήμισης (Identify potential prime advertisement locations)
- Αρχεία καταγραφής του ιστού παρέχουν πλούσιες πληροφορίες σχετικά με τη δυναμική του Ιστού (Web logs provider ich information about Web dynamics)

Τεχνικές Εξόρυξης Δεδομένων Χρήσης

- Κατασκευή πολυδιάστατης άποψης για τον Ιστό βάσης δεδομένων καταγραφής(Construct multi dimensional view on the Web log database)

- Απόδοση πολυδιάστατης ανάλυσης OLAP για να βρίσκει τους κορυφαίους N χρήστες, κορυφαίες N προσβάσιμες ιστοσελίδες, οι πιο συχνά προσβάσιμες ανά χρονικές περιόδους. (Perform multidimensional OLAP analysis to find the top N users, top N accessed Web pages, most frequently accessed time periods, etc.)

- Απόδοση εξόρυξης δεδομένων σε weblog αρχεία (Perform data mining on Weblog records)

- Εύρεση προτύπων σύνδεσης, διαδοχικά σχέδια, και τις τάσεις του Ιστού πρόσβασης (Find association patterns, sequential patterns, and trends of Web Accessing)

- Μπορεί να χρειαστούν επιπλέον πληροφορίες, περιήγηση των χρηστών των Ιστοσελίδων στον διακομιστή buffer του Ιστού. (May need additional information, e.g., user browsing sequences of the Web pages in the Web server buffer)

- Ανάλυση της απόδοσης του συστήματος, βελτίωση του σχεδιασμού του συστήματος από web-caching,προ-αναζήτηση, ιστοσελίδων και εναλλαγή ιστοσελίδων (Analyze system performance, improve system design by Web caching, Web page prefetching, and Web page swapping)

*(ΑΛΓΟΡΙΘΜΟΙ ΚΑΙ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΡΟΕΣ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ* ***ΝΙΚΟΛΑΟΣ Χ. ΤΣΙΡΑΚΗΣ)***

## 2.4.2. Επιστήμη

- NASA, EOS project: 50 GB την ώρα
- Δεδομένα για το περιβάλλον

### 2.4.3.Μάρκετινγκ-Επένδυση

Ένας αρκετά γνωστός αλγόριθμος εξόρυξης δεδομένων είναι ο A-Priori Για παράδειγμα αλγόριθμοι που εξάγουν χρήσιμη πληροφορία από μη δομημένα κείμενα, έτσι ώστε να προβλεφθούν οι τάσεις σε μετοχές.

Μεγάλες εταιρίες

WALMART: 20M συναλλαγές την ημέρα

MOBIL: 100 TB γεωλογικά σύνολα δεδομένων

AT&T 300 M κλήσεις την ημέρα

Εταιρίες πιστωτικών κρατών

### 2.4.4.Πρόληψη και Ασφάλεια

Η εξόρυξη δεδομένων έχει με επιτυχία εφαρμοστεί στην πρόληψη και αποφυγή διάφορων τύπων απάτης.

## 2.5 Οι «ρίζες» της Εξόρυξης Δεδομένων

### 2.5.1 Στατιστική

Σημαντικό μέρος της ερευνητικής βάσης για την εξόρυξη δεδομένων βασίζεται στη στατιστική. Η στατιστική έχει ανάλογη σκοπιμότητα με την εξόρυξη δεδομένων αφού και οι δύο αποσκοπούν στην αναγνώριση χρήσιμων πληροφοριών και προτύπων στα δεδομένα.

### 2.5.2 Τεχνητή Νοημοσύνη- Μηχανική Μάθησης

Δύο τομείς που σχετίζονται με την εξόρυξη δεδομένων είναι η τεχνητή

Νοημοσύνη και η μηχανική μάθησης. Στόχος της τεχνητής νοημοσύνης είναι να εξάγει λογικά συμπεράσματα από ακατέργαστα δεδομένα, κάτι που κάνει και ο τομέας της εξόρυξης δεδομένων. Επίσης ο τομέας της εξόρυξης δεδομένων κάνει εκτεταμένη χρήση εργαλείων τεχνητής νοημοσύνης και μηχανικής μάθησης, όπως φαίνεται και στα επόμενα κεφάλαια της εργασίας.

Η μηχανική μάθησης είναι ένας τομέας της τεχνητής νοημοσύνης η οποία εξετάζει τη δημιουργία προγραμμάτων τα οποία μπορούν να μαθαίνουν.

Στην εξόρυξη δεδομένων, η μηχανική μάθησης χρησιμοποιείται για τεχνικές πρόβλεψης ή κατηγοριοποίησης. Οι δύο αυτοί τομείς χρησιμοποιούνται από την εξόρυξη δεδομένων για μεθόδους βελτιστοποίησης.

### 2.5.3 Βάσεις δεδομένων

Μια βάση δεδομένων είναι μια συλλογή από δεδομένα. Αντίθετα με ένα απλό σύνολο, τα δεδομένα σε μια βάση έχουν μια δομή και σχήμα με το οποίο είναι σχετιζόμενα. Έτσι, η εξόρυξη δεδομένων μπορεί να φιλτράρει τα δεδομένα και να τα επεξεργαστεί με περισσότερη ευκολία.

## 2.6 Στόχοι και διαδικασίες της Εξόρυξης δεδομένων

Η εξόρυξη δεδομένων έχει σαν βασικούς της στόχους την εφαρμογή τεχνικών πρόβλεψης (prediction) και περιγραφής (description) σε μεγάλες βάσεις δεδομένων(Fayyad et al., 1996a; 1996b; Hegland, 2003). Ειδικότερα:

- Η **πρόβλεψη** εμπεριέχει την χρήση κάποιων μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Αλλιώς, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (predictive data mining tasks), επιδιώκουν να κάνουν εκτιμήσεις εξάγοντας συμπεράσματα από τα δεδομένα που διαθέτουμε.
- Η **περιγραφή** επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας περίπλοκης βάσης δεδομένων με τον καλύτερο δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο, δηλαδή οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων (descriptive data mining tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.

## 2.7 Τεχνικές-εργασίες εξόρυξης

1. Κατηγοριοποίηση- Ταξινόμηση (classification)
2. Ομαδοποίηση-Συσταδοποίηση (clustering)
3. Κανόνες Συσχέτισης (association rule mining)
4. Παλινδρόμηση (regression)

## 2.7.1 Κατηγοριοποίηση- Ταξινόμηση (classification)

### 2.7.1.1 Κατηγοριοποίηση

Είναι η εκμάθηση μια συνάρτησης, η κατασκευή ενός μοντέλου που απεικονίζει ένα στοιχείο από ένα σύνολο από προκαθορισμένες κλάσεις.

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

Τα βήματα που ακολουθούμε ώστε να ταξινομήσουμε τα δεδομένα είναι η **εποπτευμένη μάθηση ή εκμάθηση** και η **ταξινόμηση-κατηγοριοποίηση**.

- Η ανάθεση αντικειμένων σε προκαθορισμένες κλάσεις
  Ιδιότητες $X_1, \ldots, X_k$
- Μοντέλο κατηγοριοποίησης
  $f : \Pi(X_1) \times \ldots \times \Pi(X_k) \rightarrow \Pi(C)$
- Εκπαίδευση από υπάρχοντα δεδομένα (σύνολο εκμάθησης)

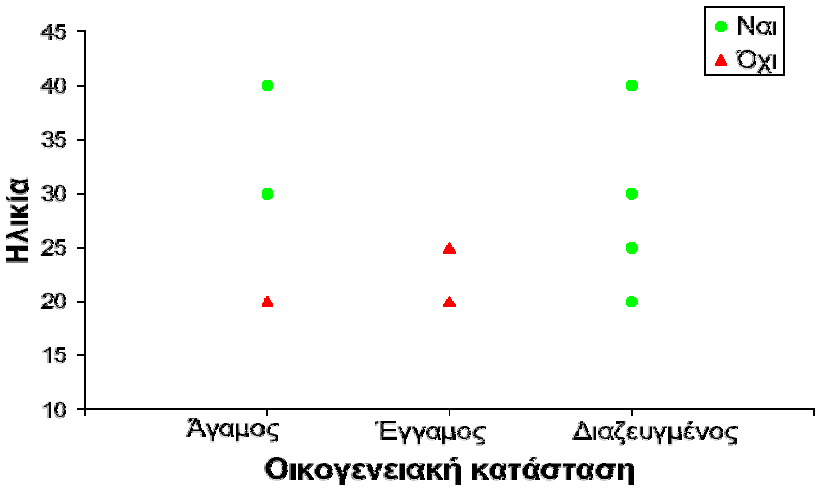
Παράδειγμα:

| *Ηλικία* | *Οικογενειακή κατάσταση* | *Αγοραστής* |
|---|---|---|
| 20 | Διαζευγμένος | Ναι |
| 30 | Διαζευγμένος | Ναι |
| 25 | Έγγαμος | Όχι |
| 30 | Άγαμος | Ναι |
| 40 | Άγαμος | Ναι |
| 20 | Έγγαμος | Όχι |
| 30 | Διαζευγμένος | Ναι |
| 25 | Διαζευγμένος | Ναι |
| 40 | Διαζευγμένος | Ναι |
| 20 | Άγαμος | Όχι |

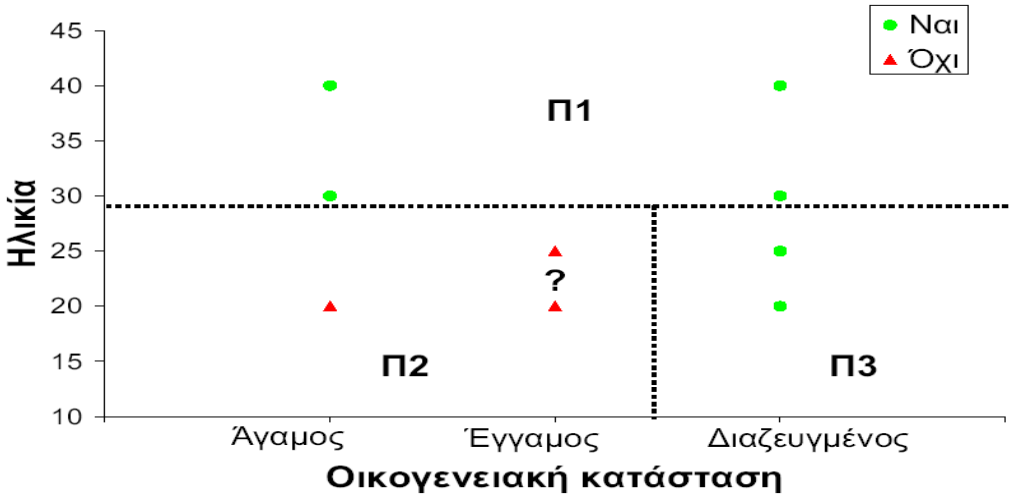
$f : [20 \ldots 40] \times \{\text{Άγγαμος, Έγγαμος, Διαζευγμένος}\} \rightarrow \{Ναι, Όχι\}$

Κατηγοριοποιητής:

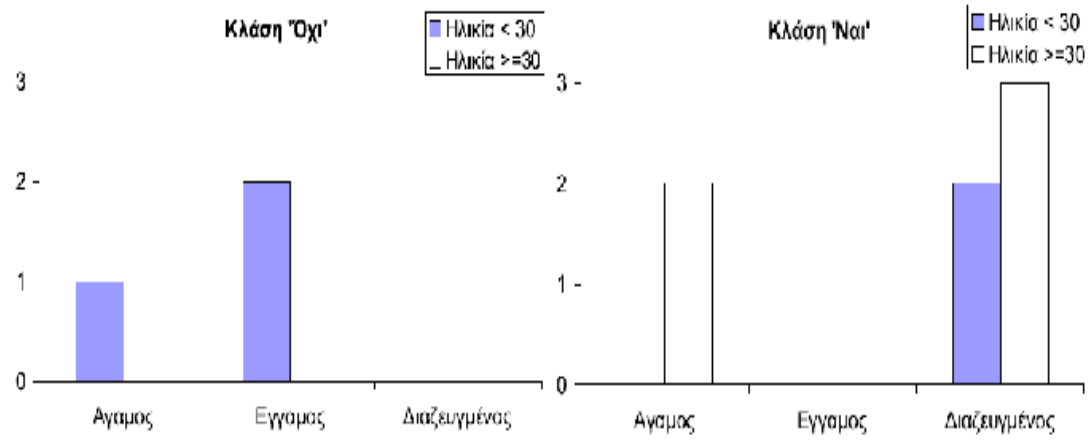
- Αλγόριθμος κατασκευής μοντέλου



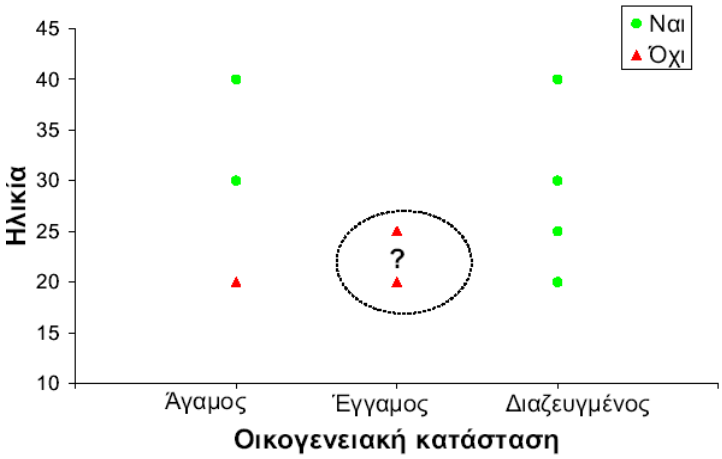

- Διαμερισμός σε περιοχές




- Εξέταση κατανομών πιθανότητας




- Εξέταση πλησιέστερων αντικειμένων




**Κριτήρια αξιολόγησης κατηγοριοποιητών**

- Ακρίβεια πρόβλεψης του μοντέλου

$$\text{Ακρίβεια} = \frac{\text{Αριθμός σωστών προβλέψεων}}{\text{Συνολικός αριθμός προβλέψεων}}$$

- Ευκολία στην κατανόηση του μοντέλου
- Κλιμάκωση στο μέγεθος του συνόλου εκμάθησης
- Ανοχή στο θόρυβο και στις ελλιπείς τιμές

Η κατηγοριοποίηση (classification) είναι μια από τις βασικές εργασίες της εξόρυξης δεδομένων. Ουσιαστικά εξετάζει τα χαρακτηριστικά ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων.

Τα αντικείμενα που θα κατηγοριοποιηθούν αναπαριστάνονται από τις εγγραφές της βάσης δεδομένων και η διαδικασία της κατηγοριοποίησης είναι να αναθέτει κάθε εγγραφή σε κάποιες από τις προκαθορισμένες κατηγορίες.

Η κατηγοριοποίηση χαρακτηρίζεται από έναν καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για το μοντέλο αποτελείται από προ-κατηγοριοποιημένα παραδείγματα.

Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιήσει δεδομένα που δεν έχουν ακόμα ανατεθεί σε κάποια από τις κατηγορίες.

Συνήθως, υπάρχει ένας περιορισμένος αριθμός κατηγοριών και πρέπει να τοποθετηθεί κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές.

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

- Bayesian ταξινόμηση
- Δένδρα Απόφασης
- Νευρωνικά Δίκτυα
- Τεχνική των κοντινότερων γειτόνων
- Support Vector Machines

**Χαρακτηριστικά Αφελών(Naïve)Bayesian**

- Η **ακρίβεια πρόβλεψης** των αφελών Bayesian κατηγοριοποιητών επηρεάζεται αρνητικά από το γεγονός ότι σε πραγματικά δεδομένα σχεδόν πάντοτε υπάρχουν εξαρτήσεις μεταξύ των μεταβλητών
- Το μοντέλο που προκύπτει είναι απλό και σχετικά **εύκολο στην κατανόηση**.
- Η κατασκευή των ιστογραμμάτων για τους υπολογισμούς των πιθανοτήτων, απαιτεί μόνο μία ανάγνωση του συνόλου δεδομένων. Επομένως, οι Bayesian κατηγοριοποιητές **κλιμακώνονται** σε μεγάλους όγκους δεδομένων.

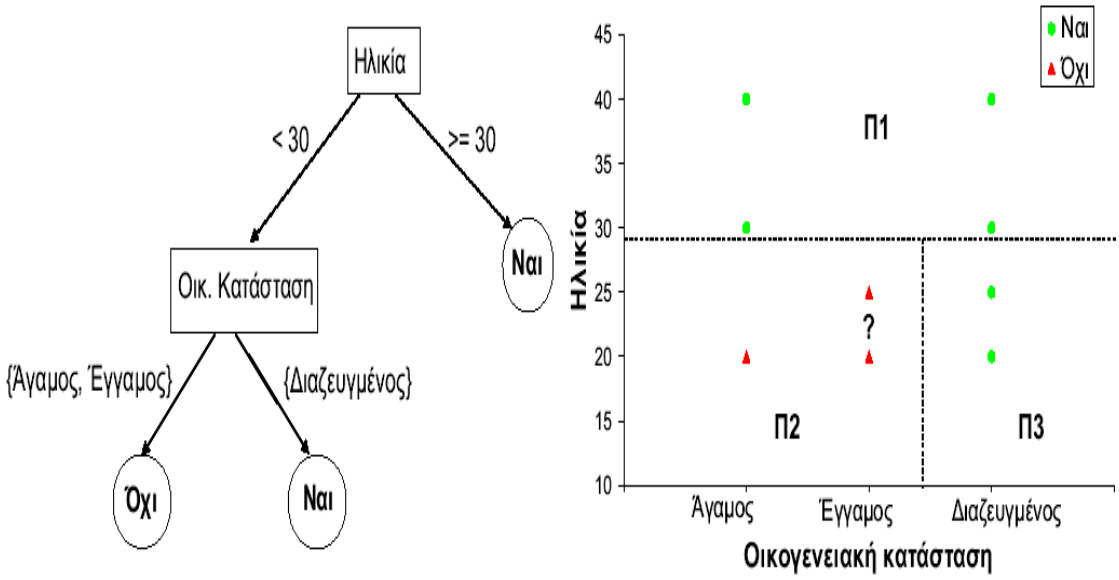
- Οι Bayesian κατηγοριοποιητές έχουν **καλή ανοχή στο θόρυβο**, επειδή οι θορυβώδεις τιμές εξομαλύνονται από τις υπόλοιπες κατά τους υπολογισμούς των εν μέρει πιθανοτήτων.
- Οι Bayesian κατηγοριοποιητές **δεν επηρεάζονται από τις ελλιπείς τιμές**, επειδή μπορούν να αγνοηθούν.

### 2.7.1.2 Δένδρα απόφασης

Τα δέντρα απόφασης (decision trees) μελετώνται συχνά ως ένα ζήτημα μηχανικής μάθησης. Για παράδειγμα, αν έχουμε ένα σύνολο εγγραφών και καθεμία από αυτές έχει μια λίστα χαρακτηριστικών. Ένα δέντρο απόφασης στο σύνολο των εγγραφών είναι ένα δέντρο όπου σε κάθε κόμβο του (που δεν είναι φύλλο) υπάρχει ένα ερώτημα που αναφέρεται στα χαρακτηριστικά των εγγραφών και κάθε ερώτημα καταλήγει σε ένα συγκεκριμένο παιδί ενός κόμβου. Τα φύλλα του δηλώνουν τις κλάσεις. Έτσι ένα δέντρο απόφασης εκτελεί κατηγοριοποίηση χρησιμοποιώντας ερωτήματα σχετικά με τα χαρακτηριστικά των εγγραφών. *(Νικόλαος Τσιράκης Αλγόριθμοι και Τεχνικές Εξόρυξης Δεδομένων 26 από Ροές Δεδομένων στον Παγκόσμιο Ιστό)*

Οι εφαρμογές που χρησιμοποιούν δέντρα απόφασης είναι παρόμοιες με αυτές που κάνουν κατηγοριοποίηση.

**Παράδειγμα**



Κατασκευή δένδρου απόφασης (brute-force)

- Κατασκευή κάθε δυνατού πιθανού δένδρου
- Επιλογή του ακριβέστερου
- NP-complete

**Χαρακτηριστικά Δένδρων Απόφασης**

- Η κατασκευή του βέλτιστου δένδρου απόφασης απαιτεί αποτρεπτικό χρόνο (είναι NP-complete πρόβλημα). Για το λόγο αυτό χρησιμοποιούνται ευρετικοί αλγόριθμοι, οι οποίοι είναι άπληστοι και δεν χρησιμοποιούν οπισθοδρόμηση. Τα ευρετικά μειώνουν κατά πολύ το χρόνο κατασκευής. Το αποτέλεσμα είναι ότι τα δένδρα απόφασης **κλιμακώνονται σε μεγάλους όγκους δεδομένων**
- Η **ακρίβεια πρόβλεψης** των δένδρων απόφασης είναι αποδεκτή για τις περισσότερες περιπτώσεις, συγκρίσιμη με την ακρίβεια άλλων κατηγοριοποιητών
- Το μοντέλο που προκύπτει είναι πολύ **εύκολο στην κατανόηση**.
- Τα δένδρα απόφασης έχουν καλή **ανοχή στο θόρυβο**, ειδικά όταν εφαρμόζεται ψαλιδισμός

### 2.7.1.3 Μηχανές διανυσμάτων Υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης (support vector machines) αποτελούν την πιο πρόσφατη μέθοδο μηχανικής μάθησης.

Ας υποθέσουμε πως έχουμε στην διάθεση μας n αντικείμενα εκπαίδευσης, τα οποία αποτελούνται από ένα διάνυσμα n xi ∈R και μια τιμή κλάσης j y . Μια μηχανή διανυσμάτων υποστήριξης παράγει ένα ταξινομητή, το βέλτιστο υπέρ-επίπεδο διαχωρισμού, μέσα από την μη γραμμική απεικόνιση των εισερχόμενων ανωτέρω διανυσμάτων στον πολύ-διάστατο χώρο των χαρακτηριστικών που περιγράφουν τα διανύσματα αυτά (Shin et al., 2005).Δηλαδή, μια μηχανή διανυσμάτων υποστήριξης αντιστοιχεί τα δεδομένα μιας βάσης σε ένα πολύ-διάστατο χώρο, καθορίζοντας στον χώρο αυτό ένα βέλτιστο υπέρ-επίπεδο διαχωρισμού τους.

Η βασική έννοια γύρω από την οποία δομείται μια μηχανή διανυσμάτων υποστήριξης είναι αυτή του περιθωρίου (margin), σε κάθε μια από τις πλευρές ενός καθορισμένου υπέρ-επιπέδου που χωρίζει τα δεδομένα ενός συνόλου εκπαίδευσης. Συγκεκριμένα, μια μηχανή διανυσμάτων υποστήριξης κατασκευάζει ένα γραμμικό μοντέλο για την εκτίμηση του συνόλου των παραμέτρων $a$ της συνάρτησης απόφασης $f(x, a)$ , έτσι ώστε η τελευταία να πραγματοποιήσει την αντιστοίχιση $i\ j\ x \rightarrow y$ (η $f(x, a)$ ονομάζεται *μηχανή εκπαίδευσης*) (Βαζιργιάννης & Χαλκίδη, 2003).

Αν τα δεδομένα εκπαίδευσης είναι γραμμικά διαχωρισμένα, τότε η μηχανή διανυσμάτων υποστήριξης εκπαιδεύει μηχανές για την εκτίμηση ενός βέλτιστου υπέρ-επιπέδου που διαχωρίζει τα δεδομένα, στην μεγαλύτερη δυνατή απόσταση ανάμεσα σε αυτό και τα πλησιέστερα αντικείμενα εκπαίδευσης. Τα αντικείμενα εκπαίδευσης που είναι πλησιέστερα στο βέλτιστο υπέρ-επίπεδο διαχωρισμού ονομάζονται *διανύσματα υποστήριξης* (***support vectors***), επίσης η λύση αναπαρίσταται ως ένας γραμμικός συνδυασμός αυτών των αντικειμένων. Σε πιο γενικές καταστάσεις που τα δεδομένα είναι μη γραμμικά διαχωρισμένα, η μηχανή διανυσμάτων υποστήριξης χρησιμοποιεί μη γραμμικές μηχανές για την εύρεση ενός

υπέρ-επιπέδου που ελαχιστοποιεί το πλήθος των λαθών για το σύνολο εκπαίδευσης (Cristianini & Shawe-Taylor, 2000; Shin et al., 2005).

Η μεγιστοποίηση του περιθωρίου και κατ' επέκταση η δημιουργία της μεγαλύτερης δυνατής απόστασης ανάμεσα στο υπέρ-επίπεδο και τα αντικείμενα των δεδομένων εκπαίδευσης σε κάθε πλευρά του πρώτου, έχει αποδειχτεί ότι ελαχιστοποιεί το άνω όριο του σφάλματος γενίκευσης. Αυτή η ελαχιστοποίηση επιτυγχάνεται με την εκπαίδευση του $a$, ώστε η $f(x, a)$ να ικανοποιεί την ιδιότητα του *μέγιστο υπεριθωρίου*, δηλαδή το όριο απόφασης που αντιπροσωπεύει να έχει την μέγιστη απόσταση από το κοντινότερο αντικείμενο εκπαίδευσης.

## 2.7.2.Ομαδοποίηση-Συσταδοποίηση (clustering)

Είναι εύρεση ενός συνόλου από ομάδες με όμοια στοιχεία - χωρίζουμε τα δεδομένα σε ομάδες από «όμοια» σύνολα.

Αρχικά με τη μέθοδο της ομαδοποίησης, αναφερόμαστε σε μια διεργασία στην διαδικασία εξόρυξης γνώσης, βάση της οποίας εξάγουμε χρήσιμες πληροφορίες από τα δεδομένα που μελετούμε. Η μέθοδος αυτή στοχεύει στην τμηματοποίηση μιας ομάδας δεδομένων σε ομάδες που περιέχουν στοιχεία με περισσότερα κοινά χαρακτηριστικά μεταξύ σε σχέση με στοιχεία των άλλων ομάδων.

Η ομαδοποίηση θεωρείται ως μια από τις σημαντικότερες διεργασίες στη διαδικασία της μη-εποπτευμένης μάθησης, από την οποία μπορούν να περιγραφούν κατανομές ή πρότυπα που παρουσιάζουν ενδιαφέρον στα υπό μελέτη δεδομένα μας.

Η συσταδοποίηση-ομαδοποίηση (clustering) είναι η εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων (clusters). Αυτό που διαφοροποιεί τη συσταδοποίηση από την κατηγοριοποίηση είναι ότι η συσταδοποίηση δε βασίζεται σε προκαθορισμένες κατηγορίες.

Στην κατηγοριοποίηση, ο πληθυσμός διαιρείται σε κατηγορίες αναθέτοντας κάθε στοιχείο ή εγγραφή σε μια προκαθορισμένη κατηγορία με βάση ένα μοντέλο που αναπτύσσεται μέσω της εκπαίδευσης του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων.

Όπως και στην κατηγοριοποίηση έτσι και στη συσταδοποίηση υπάρχουν πολλές εφαρμογές. Για παράδειγμα, ας θεωρήσουμε πως έχουμε διαθέσιμα τα δεδομένα πελατών μιας επιχείρησης. Χρησιμοποιώντας τεχνικές συσταδοποίησης, μπορούμε να βρούμε τον καταμερισμό των πελατών και της αγοράς, π.χ. μπορούμε να δούμε ποιοι πελάτες ψωνίζουν για την οικογένεια τους και ποιοι για τον εαυτό τους ή ποιοι έχουν υψηλό εισόδημα και ποιοι όχι.

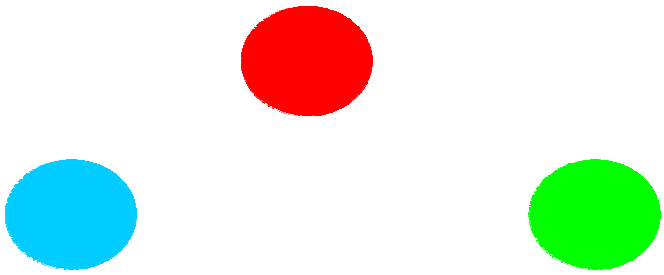
Πεδία εφαρμογή της συσταδοποίησης:

- Ελάττωση δεδομένων
- Παραγωγή υπόθεσης
- Έλεγχος υπόθεσης
- Πρόβλεψη βασισμένη σε συστάδες

## ΕΙΔΗ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ:

### Διαιρετική Ομαδοποίηση

Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη για την συμπίεση και την εμφάνιση μεγάλων βάσεων δεδομένων. Κυρίως βασίζεται στην άμεση αποσύνθεση του συνόλου των δεδομένων που εξετάζονται σε ένα σύνολο ομάδων οι οποίες είναι ασυσχέτιστες μεταξύ τους.



Τρεις διαιρεμένες - χωρισμένες συστάδες

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 φτιάχνονται.

ΔΙΑΙΡΕΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ

- *K-Means*
- *Αλγόριθμος ISODATA*

### Ιεραρχική ομαδοποίηση

Σκοπός αυτής της μεθόδου είναι είτε η συνένωση μικρότερων ομάδων συνόλων δεδομένων σε μεγαλύτερες, ή η διαχώριση μεγαλύτερων ομάδων σε μικρότερες.

Είναι σημαντικό να διαπιστωθεί ποιες από αυτές ήταν μεγάλες και διασπάστηκαν και ποιες μικρές και άρα συνενώθηκαν.

Υπάρχουν δυο είδη ιεραρχικών αλγορίθμων:

Συσσωρευτικοί

Οι αλγόριθμοι αυτοί είναι αποτελεσματικοί σε διάφορα πεδία όπως η αναγνώριση οπτικών χαρακτήρων, η ομαδοποίηση εγγράφων και η εικόνα ιατρικής γνωμάτευσης.

*Διαιρετικοί*

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

Ιεραρχικοί Αλγόριθμοι:

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

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

- *CURE*

Ο αλγόριθμος αυτός καθορίζει τις ομάδες που έχουν ακαθόριστο σχήμα και σημαντικές διαφορές στο μέγεθος.

- *ROCK (RObust Clustering using linKs)*

Ο αλγόριθμος αυτός αξιολογεί την έννοια των συνδέσεων (links) μεταξύ των ζευγών των σημείων που εξετάζονται ενός στη τύχη επιλεγμένου δείγματος. Αυτό έχει αποτέλεσμα χάρη σε μια συνάρτηση ομοιότητας μεταξύ δύο γειτονικών σημείων.

**Ομαδοποίηση βασισμένη σε γράφους:**

Αυτή η μέθοδος βασίζεται στο ότι ένας αλγόριθμος ομαδοποίησης βασισμένης σε γράφους όταν εφαρμοστεί σε ένα γράφο διατηρεί μόνο τις γειτονικές συνδέσεις ενός στοιχείου με τα πλησιέστερα του και καθορίζει ένα βαθμό ομοιότητας για τα στοιχεία το οποίο καθορίζεται από τον αριθμό των κοντινότερων γειτόνων τους. Ακόμα καθορίζει τα στοιχεία-πυρήνες που μας ενδιαφέρουν περισσότερο και δημιουργεί ομάδες περίγυρα από αυτά και τέρμα, αξιοποιεί τις πληροφορίες που συνθέτουν το γράφο έτσι ώστε να αξιολογήσει αν δύο ομάδες πρέπει να συγχωνευθούν.

**Ομαδοποίηση βασισμένη στην πυκνότητα:**

Αυτός ο τρόπος της ομαδοποίησης βασίζεται στην οργάνωση ομάδων των γειτονικών στοιχείων ενός συνόλου δεδομένων με αφετηρία κάποια κριτήρια πυκνότητας.

**Ομαδοποίηση υποχώρων:**

Η ομαδοποίηση υποχώρων εφαρμόζεται σε προβλήματα που προκύπτουν από τα δεδομένα υψηλών διαστάσεων. Συμβαίνει αυτό λόγω της ύπαρξης των πολλών διαστάσεων, συγχρόνως και αυτών που αντιστοιχούν σε θόρυβο, σχεδόν κάθε υποσύνολο εμφανίζει χαμηλή πυκνότητα σημείων**.**

**Ομαδοποίηση βασισμένη σε πλέγμα (Grid based) αλγόριθμοι:**

Στις μέρες μας διάφοροι αλγόριθμοι ομαδοποίησης έχουν παρουσιαστεί για χωρικά δεδομένα, οι οποίοι κβαντοποιούν το διάστημα σε έναν πεπερασμένο αριθμό κελιών και κάνουν έπειτα όλες τις διαδικασίες στο κβαντοποιημένο αυτό διάστημα.

## 2.7.3 Κανόνες συσχέτισης

Η εξαγωγή κανόνων συσχέτισης (association rules) είναι κατά πολλούς μια από τις πιο σημαντικές διεργασίες εξόρυξης δεδομένων.

Έχουν μεγάλη σημασία επειδή παρέχουν έναν περιληπτικό τρόπο για να εκφραστούν οι χρήσιμες πληροφορίες, συνήθως εύκολα κατανοητές από τους τελικούς χρήστες.

Βρίσκουμε συσχετίσεις αναμεταξύ των δεδομένων, π.χ. ποια δεδομένα

εμφανίζονται συχνά μαζί σε εγγραφές.

Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των χαρακτηριστικών ενός συνόλου των δεδομένων. Αυτοί οι συσχετισμοί παρουσιάζονται στην ακόλουθη μορφή:

A→B όπου το A και το B αναφέρονται στα σύνολα χαρακτηριστικών που εμφανίζονται στα αναλυόμενα δεδομένα.

Οι κανόνες συσχέτισης είναι μία πρόσφατη μέθοδος που ανακαλύφθηκε στις αρχές της δεκαετίας του ’90 και προέρχεται από την ανάγκη των ισχυρών αγορών να καταχωρήσουν τις συναλλαγές κάθε πελάτη με ηλεκτρονική μορφή αναλύοντας το «καλάθι αγοράς» του (market basket analysis). Οι ισχυρές αυτές αγορές (υπεραγορές) με τη χρήση της μεθόδου αυτής συγκεντρώνουν ένα σημαντικότατο όγκο πληροφοριών που σχετίζονται με τις κινήσεις-αγορές των πελατών τους.

Εύρεση Συχνών Προτύπων, Εξαρτήσεων και Συσχετίσεων –Dependencies and associations: εύρεση σημαντικών/συχνών εξαρτήσεων μεταξύ γνωρισμάτων

Εύρεση Κανόνων Συσχέτισης [Descriptive]

Παράδειγμα:

{pen} ⇒ {ink}

*Συνδυαστικοί κανόνες*

*Αν ένα στοιχείο* **pen** *αγοράζεται σε μια συναλλαγή, τότε είναι πιθανό ότι αγοράζεται και το στοιχείο* **ink**

*Γενικά, συνδυαστικός κανόνας* ***(association rule).***

**Αλγόριθμος:**

**Apriori:** Διαβάζει τον αρχικό πίνακα D διαδοχικές φορές.

**Ποσοτικοί κανόνες συσχέτισης:**

Ισο-βαθύς κατάτμηση (Equi-depth Partitioning)

Κανόνες μα βάση την απόσταση (Distance-bases Rules)

## 2.7.4 Παλινδρόμηση

Η παλινδρόμηση είναι μια μέθοδος της εξόρυξης δεδομένων η οποία χρησιμοποιείται ως μία συνάρτηση πρόβλεψης στην οποία μέσα έχουν καταχωρηθεί τα εξεταζόμενα δεδομένα και η οποία προβλέπει ένα ρεαλιστικό αριθμό.

Με τη χρήση αυτής της μεθόδου έχουμε τη δυνατότητα να προβλέψουμε το ύψος των κερδών μιας επιχείρησης, το ύψος των τιμών ακινήτων, τις πωλήσεις, τους βαθμούς κελσίου, το τετραγωνικό μήκος σε μέτρα, τις αποστάσεις κλπ.

Αποτελεί μία μέθοδος η οποία έχει μελετηθεί σημαντικά κυρίως στη στατιστική αλλά και στα νευρωνικά δίκτυα. Βασικός σκοπός είναι η πρόβλεψη της τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν. Πολλές φορές χρησιμοποιούμε ένα μοντέλο για την κάθε μεταβλητή. Η παλινδρόμηση καλύπτει ένα μεγάλο τμήμα του τομέα της εξόρυξης δεδομένων που έχει να κάνει με προβλέψεις.

## Κεφάλαιο 3$^{ο}$ - Επιχειρησιακή έρευνα

Η Επιχειρησιακή Έρευνα, η οποία ασχολείται με την αποδοτική κατανομή πόρων που είναι διαθέσιμοι σε ανεπαρκείς ποσότητες, αποτελεί τόσο τέχνη όσο και επιστήμη. Η τέχνη αφορά το σκέλος της ικανότητας αναπαράστασης των εννοιών *αποδοτική κατανομή* και *ανεπάρκεια* σε ένα καλώς ορισμένο μαθηματικό μοντέλο το οποίο αναφέρεται σε μια δεδομένη περίπτωση. Η επιστήμη υπεισέρχεται στη διατύπωση των υπολογιστικών μεθόδων επίλυσης αυτών των μοντέλων. *(SHAUM'S Επιχειρησιακή έρευνα Δεύτερη αμερικάνικη έκδοση εκδόσεις κλειδάριθμος)*

Εφόσον η καλύτερη-βέλτιστη κατανομή των οικονομικών πόρων των ενεργειακών πόρων, του ανθρώπινου δυναμικού, όπως και διάφορων άλλων πόρων σε ανεπάρκεια, καθίσταται σημαντική σε πολλές παραδοσιακούς επιστημονικούς τομείς, η ΕΕ είναι χρήσιμη σε άτομα που προέρχονται από διάφορα γνωστικά αντικείμενα.


### 3.1 Ορισμοί επιχειρησιακής έρευνας

Η Επιχειρησιακή Έρευνα είναι η εφαρμογή της σύγχρονης επιστήμης πάνω σε πολύπλοκα προβλήματα που ανακύπτουν στη διεύθυνση και διοίκηση μεγάλων συστημάτων, αποτελούμενων από ανθρώπους, μηχανές, υλικά και κεφάλαια στις επιχειρήσεις.

Η χαρακτηριστική της μεθοδολογία συνίσταται στην ανάπτυξη επιστημονικού μοντέλου του υπό μελέτη συστήματος που περιλαμβάνει μετρήσεις τυχαίων παραγόντων και με το οποίο προβλέπει και συγκρίνει τα αποτελέσματα εναλλακτικών αποφάσεων, στρατηγικών και ελέγχων.

Ο σκοπός της είναι να βοηθήσει τη διοίκηση να καθορίσει την πολιτική και τις ενέργειές της επιστημονικά.*(Εταιρείας Επιχειρησιακής Έρευνας UK)*

- Η συστηματική εφαρμογή ποσοτικών μεθόδων, τεχνικών και εργαλείων στην ανάλυση προβλημάτων που εμπεριέχουν την λειτουργία συστημάτων.

*(Daellen bachand George,1978)*

Η Επιχειρησιακή Έρευνα μπορεί να θεωρηθεί ότι είναι:

- Η εφαρμογή επιστημονικών μεθόδων από μικτές ομάδες σε προβλήματα που αφορούν τον έλεγχο οργανωμένων συστημάτων (αποτελούμενων

από ανθρώπους και μηχανές) κατά τρόπο ώστε να παρέχουν λύσεις που εξυπηρετούν κατά τον καλύτερο δυνατό τρόπο τους σκοπούς του οργανισμού ή της επιχείρησης ως συνόλου.

*(Fundamentals of Operations Research, Ackoff and Sasienni)*

**και ο Ελληνικός Ορισμός...**

- Επιχειρησιακή Έρευνα είναι η επιστημονική προετοιμασία των αποφάσεων της Διοικήσεως (με την επιστημονική ανάλυση των δεδομένων και τη δημιουργία μαθηματικών προτύπων).

## 3.2 Ορολογία επιχειρησιακής έρευνας

- OR: Operational Research (Ευρωπαϊκή & Αμερικάνικη ονομασία)
- MS: Management Science - Διοικητική επιστήμη (εναλλακτική ονομασία)
- DS: Decision Science - Επιστήμη Αποφάσεων(σπανιότερα χρησιμοποιούμενο)
- ΕΕ: Επιχειρησιακή Έρευνα (Ελληνική ονομασία)

## 3.3 Ιστορία επιχειρησιακής έρευνας

Η Επιχειρησιακή Έρευνα πρωτοεμφανίστηκε στη δεκαετία του 1930 στην Μεγάλη Βρετανία και προσπάθησε να βρει λύσεις σε καθαρά λειτουργικά προβλήματα. Ξεκίνησε σαν έναν τρόπος εύρεσης του αποδοτικότερου τρόπου εξολόθρευσης στρατιωτών εν καιρώ πολέμου.

Οι πρώτες απόπειρες σε μη στρατιωτικά προβλήματα:

Μελέτη προβλημάτων σε σχέση με το χρόνο που μιλάει κανείς στα τηλεφώνα, Erlang1917, μελέτες ζητημάτων εμπορίου και πωλήσεων, Horace Levenson1920.

Πρώτη χρήση του όρου «Επιχειρησιακή Έρευνα» είναι για την ανάπτυξη και τελειοποίηση διαφόρων τύπων radars, Κέντρο Ερευνών Αγγλικού Υπουργείου Άμυνας 1937-39.

Ορισμένα ενδιαφέροντα στοιχεία:

Οι πρώτοι μελετητές ΕΕ προέρχονταν από διάφορες επιστημονικές περιοχές (Μαθηματικοί, Φυσικοί, Ψυχολόγοι, Τοπογράφοι).

Η συνεισφορά τους ήταν διάφορες μελέτες μέσω ελέγχων υποθέσεων, λογικών συγκρίσεων, αξιολόγηση αποτελεσμάτων, πραγματοποίησης πειραμάτων, συλλογής και ανάλυσης δεδομένων, μετατροπής δεδομένων σε πληροφορίες.

Τέσσερις τουλάχιστον από τους μελετητές του OR Section (ΕΕ) κέρδισαν βραβείο Nobel μετά τον πόλεμο.

Μετά την ολοκλήρωση του πολέμου, η ΕΕ ακολούθησε το δρόμο της και απέκτησε ευρύτερο πεδίο εφαρμογής, ανάλογα με τις ανάγκες της ανθρωπότητας σε καιρό ειρήνης.

Στη Μεγάλη Βρετανία, οι μελετητές του OR Section γύρισαν στις παλιές τους δουλειές και τελικώς, η ΕΕ δεν διαδόθηκε τόσο όσο θα έπρεπε, εκτός από συγκεκριμένες εξαιρέσεις (βιομηχανία μετάλλων, ορυχεία).

Στην Αμερική, η ΕΕ αναπτύχθηκε σημαντικά μέσω των Πανεπιστημιακών Ιδρυμάτων με αποτέλεσμα η χρήση της να διαδοθεί πολύ περισσότερο και η εκπαίδευση σε σχετικά ζητήματα να είναι προηγμένη.

## 3.4 Βασικά χαρακτηριστικά της επιχειρησιακής έρευνας

**1.Είναι σχετικά καινούργια μορφή έρευνας που αξιοποιείται σε συνεργασία με διοικητικά στελέχη.**

Για παράδειγμα παλαιότερα, πολλά προβλήματα που τώρα μελετούνται από την ΕΕ λύνονταν με τη λογική:

- Η ΕΕ αποφορτίζει τα διοικητικά στελέχη και τα διευκολύνει στη λήψη σημαντικών αποφάσεων
- Η ΕΕ πραγματοποιείται στον εργασιακό χώρο και όχι σε κάποιο ερευνητικό εργαστήριο
- Η συνεργασία των διοικητικών στελεχών και οι πληροφορίες που παρέχουν καθορίζουν την επιτυχία των έργων ΕΕ

**2. Αναφέρεται σε προβλήματα λήψης αποφάσεων και ελέγχου ενεργών συστημάτων.**

Η λήψη αποφάσεων σε μεγάλες και διαρκώς αναπτυσσόμενες επιχειρήσεις γίνεται πολύ δύσκολη υπόθεση

- Η αποκλειστική υιοθέτηση εμπειρικών τρόπων λήψης αποφάσεων κρύβει κινδύνους
- Η λήψη αποφάσεων θα πρέπει να περιλαμβάνει την οπτική των χαμηλότερων λειτουργικών επιπέδων, κάτι που συνήθως παραμελείται

**3. Εφαρμόζει επιστημονική μεθοδολογία για την ποσοτική εκτίμηση της βέλτιστης λύσης προβλημάτων με βάση αντικειμενικά κριτήρια και χρησιμοποιεί μοντέλα.**

**4. Διεξάγεται από μικτές ομάδες επιστημόνων.**

- Χρησιμοποιούνται ομάδες επιστημόνων πολλών ειδικοτήτων όπως:
- Μηχανικοί
- Οικονομολόγοι
- Ψυχολόγοι

**5. Κάθε ειδικότητα εξασφαλίζει μια διαφορετική οπτική του προβλήματος που πρέπει να επιλυθεί.**

**6. Έχει υιοθετήσει προσέγγιση μέσω συστημάτων (systems approach)**

## 3.5 Πρότυπα (ή μοντέλα)

**Εικονικά:**

- Πιστές αναπαραστάσεις ενός υπό μελέτη συστήματος (π.χ. χάρτες, πρότυπα πλοίων και αεροπλάνων δοκιμαζόμενα σε κατάλληλους χώρους).
- Πολύ συγκεκριμένα και ειδικά.
- Πολύ δύσχρηστα.

**Αναλογικά:**

- Οι ιδιότητες του συστήματος παριστάνονται από άλλες ιδιότητες (π.χ. οι γεωδαιτικές γραμμές σε χάρτη παριστάνουν υψόμετρο, οι ηλεκτρικές ροές, ένταση ή τάση, παριστάνουν αντίστοιχα οικονομικά μεγέθη).
- Γενικότερα, λιγότερο συγκεκριμένα
- Πιο εύχρηστα από τα εικονικά

**Συμβολισμοί:**

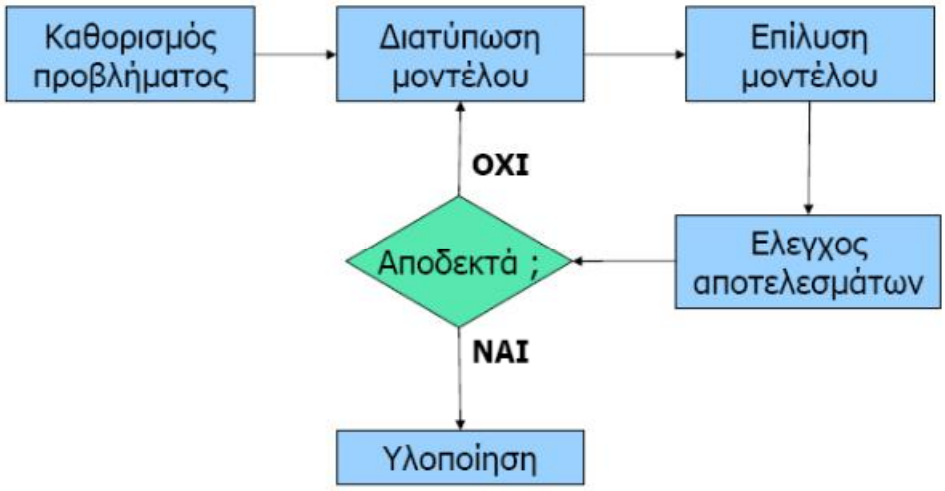
- Χρησιμοποίηση γραμμάτων, αριθμών, συμβόλων (π.χ. +, -, >=) για την αναπαράσταση των παραμέτρων ενός προβλήματος και των σχέσεων μεταξύ τους

- Πιο γενικά
- Πιο εύχρηστα
- Περισσότερο χρησιμοποιούμενα

## 3.6 Μεθοδολογία

Βασικά στάδια της μεθοδολογίας της Επιχειρησιακής Έρευνας:

1. **Διαμόρφωση του προβλήματος**
2. **Κατασκευή του μαθηματικού προτύπου**
3. **Επίλυση του μαθηματικού προτύπου**
4. **Έλεγχος της λύσεως**
5. **Υλοποίηση και διατήρηση της λύσεως**




### 3.6.1 Διαμόρφωση του προβλήματος

- Διάγνωση του προβλήματος (συνήθως, από τα συμπτώματά του όταν δεν είναι προφανές)
- Εντοπισμός υπό-προβλήματος που θα επιλυθεί
- Ορισμός στόχου, περιορισμών και απαιτήσεων

**Βήματα διαμόρφωσης του προβλήματος**

- Διάγνωση συμπτωμάτων μη ικανοποιητικής λειτουργίας
- Καθορισμός στοιχείων προβλήματος, όπως:
  - Κατανόηση οργανωτικής δομής
  - Εντοπισμός ατόμων που λαμβάνουν τις αποφάσεις (decision makers)
  - Ορισμός αντικειμενικών στόχων (ελαχιστοποίηση κόστους, μεγιστοποίηση κέρδους, αύξηση ποσοστού αγοράς κλπ)
  - Εύρεση εναλλακτικών τρόπων δράσεως με την επισήμανση των ελεγχόμενων μεταβλητών
  - Εντοπισμός του περιβάλλοντος του υπό μελέτη συστήματος (μη ελεγχόμενες μεταβλητές)
- Προσδιορισμός ενός κριτηρίου επιλογής της βέλτιστης λύσης, με κριτήρια:
  - Αιτιοκρατικά (π.χ. κόστος λειτουργίας)
  - Στοχαστικά (π.χ. τιμή κόστους όταν υπεισέρχονται τυχαίες μεταβλητές των οποίων μπορούν να προσδιορισθούν οι κατανομές)

## 3.6.2 Κατασκευή μαθηματικού προτύπου

- Εντοπισμός εναλλακτικών τρόπων μοντελοποίησης
- Εντοπισμός διαθεσίμων δεδομένων
- Υπολογισμός δυσκολίας απόκτησης δεδομένων
- Επιλογή τελικού τρόπου μοντελοποίησης
- Επισημοποίηση μοντέλου
- Το μοντέλο είναι δυνατό να περιλαμβάνει ατέλειες όπως:
  - Να περιλαμβάνει άσχετες μεταβλητές
  - Να μην περιλαμβάνει σχετικές μεταβλητές
  - Ορισμένες μεταβλητές να μην αξιολογούνται σωστά
  - Η δομή του να είναι εσφαλμένη (η συνάρτηση που συνδέει το κριτήριο λειτουργίας του με τις ελεγχόμενες μεταβλητές)
- Η εξακρίβωση ατελειών πραγματοποιείται με στατιστικές αναλύσεις (συσχετίσεως, παλινδρομήσεως, εκτιμήσεως, δειγματοληψίας)
- Οι απαραίτητες πληροφορίες για την κατασκευή του μοντέλου εξασφαλίζονται με την ανάλυση συστήματος

## 3.6.3 Επίλυση του μαθηματικού προτύπου

- Χρησιμοποίηση αλγορίθμου ή και χρησιμοποίηση ηλεκτρονικού υπολογιστή
- Εύρεση λύσεων (ανάλυση ευαισθησίας, what-if analysis)

**Μέθοδοι Επίλυσης Μαθηματικού Μοντέλου:**

Μέθοδοι και Θεωρίες Επιχειρησιακής Έρευνας:

Γραμμικός και μη γραμμικός προγραμματισμός, δυναμικός προγραμματισμός, ακέραιος και συνδυαστικός προγραμματισμός, χρονικός προγραμματισμός, προσομοίωση, ευρετικές μέθοδοι, θεωρίες ελέγχου αποθεμάτων, αναμονής, συντηρήσεως και αντικαταστάσεως, μηχανικού εξοπλισμού, αξιοπιστίας, παιγνίων.

**Υλοποίηση και διατήρηση της λύσεως**

- Έγκριση των αποτελεσμάτων από την επιχείρηση
- Υλοποίηση των αποτελεσμάτων που προέκυψαν από την μελέτη
- Χρησιμοποίηση του αλγορίθμου/ εργαλείου για την επίλυση αντίστοιχων λειτουργικών προβλημάτων

## 3.6.4 Διακρίσεις Προβλημάτων

- **Επίπεδο διοικήσεως**

Τρία επίπεδα: τακτικό, τεχνικό και στρατηγικό.

- **Περιεχόμενο**

Υπάρχουν αναρίθμητα προβλήματα ΕΕ από πλευράς περιεχομένου όπως:

- Παραγωγή: επιλογή θέσεως εργοστασίου
- Εμπορία: Καθορισμός βέλτιστης σύνθεσης παραγωγής
- Οικονομικά: Χρηματοοικονομικός προγραμματισμός, καθορισμός πιστωτικής πολιτικής, προϋπολογισμός
- Προσωπικό: Ανάλυση & αξιολόγηση προσωπικού

- **Τον Τύπο ή τη Μορφή**

Ο τύπος του προβλήματος αναφέρεται στον τρόπο με τον οποίο οι παράμετροί του σχετίζονται μεταξύ τους:

- Το μαθηματικό πρότυπο ή μοντέλο καθορίζει τον τύπο του προβλήματος

- Υπάρχουν συγκεκριμένες κατηγορίες προβλημάτων ΕΕ, χωρίς αυτό να σημαίνει ότι δεν υπάρχουν προβλήματα που δεν μπορούν να ενταχθούν σε καμία από τις κατηγορίες αυτές
- Η κατηγοριοποίηση αυτή διευκολύνει τη διδασκαλία των μοντέλων που ανήκουν στην ίδια κατηγορία

## 3.7 Βελτιστοποίηση στην επιχειρησιακή έρευνα

Η επιχειρησιακή έρευνα μοντελοποιεί προβλήματα βελτιστοποίησης μέσω των τεχνικών που προσφέρει επιτρέποντας την μετέπειτα επίλυσή τους.

Στην ιστορία της επιχειρησιακής έρευνας περιλαμβάνονται πολλές επιτυχημένες εφαρμογές σε δύσκολα προβλήματα βελτιστοποίησης συνήθως με σημαντικότατη πρακτική σημασία. Οπότε είναι λογικό πως αξιοποιήθηκε η επιτυχία του γραμμικού προγραμματισμού από τα προβλήματα συνεχών μεταβλητών στα προβλήματα συνδυαστικής βελτιστοποίησης χρησιμοποιώντας τον ακέραιο προγραμματισμό (Integer Programming= IP).

Αυτός ο τρόπος προσέγγισης χρειάζεται την εκφορά μαθηματικών μοντέλων που να περιγράφουν το πρόβλημα κατά ενδεχόμενο.

***Πρόβλημα βελτιστοποίησης*** *είναι ένα πρόβλημα στο οποίο ζητείται η καλύτερη δυνατή λύση ανάμεσα σε όλες τις διαθέσιμες λύσεις*.(ΑΛΓΟΡΙΘΜΟΙ ΣΥΝΔΥΑΣΤΙΚΗΣ *ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΜΕ ΕΜΦΑΣΗ ΣΕ ΜΕΤΑΕΥΡΕΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΧΡΗΣΤΟΣ Γ. ΓΚΟΓΚΟΣ)*

Τα προβλήματα βελτιστοποίησης διαιρούνται σε προβλήματα με μεταβλητές απόφασης που λαμβάνουν συνεχείς τιμές (συνεχή) και σε προβλήματα με μεταβλητές απόφασης που λαμβάνουν διακριτές τιμές (συνδυαστικά).

Τα συνεχή προβλήματα βελτιστοποίησης χωρίζονται σε **γραμμικά** και σε **μη γραμμικά προβλήματα**. Η επίλυση των γραμμικών προβλημάτων γίνεται αποτελεσματικά με τον γραμμικό προγραμματισμό ενώ η επίλυση των μη γραμμικών προβλημάτων συχνά επεξεργάζεται με μεθόδους καθολικής βελτιστοποίησης ή με ευρετικές μεθόδους.

### Γειτονιά της λύσης

Η αναζήτηση της βέλτιστης λύσης ενός προβλήματος θεωρείται μια διαδικασία στην οποία κατευθυνόμαστε από μια αρχική λύση προς την τελική λύση μέσω ενδιάμεσων βημάτων-λύσεων. Η γειτονιά μια λύσης x είναι το σύνολο των πιθανών λύσεων στις οποίες μπορούμε να μεταβούμε από την x με μια απλή λειτουργία τροποποίησης ενός χαρακτηριστικού του x. Πολλές φορές χρησιμοποιούνται γειτονιές με σχετικά μικρό μέγεθος λόγω του ότι δίνουν τη δυνατότητα σε κάθε βήμα να ερευνούνται όλες οι γειτονικές λύσεις της τρέχουσας λύσης. Μια από τις πλέον κοινές μορφές γειτονιάς είναι η γειτονιά διπλής εναλλαγής (two-exchange) στην οποία επιλέγονται δύο χαρακτηριστικά της τρέχουσας λύσης και αλλάζουν θέση μεταξύ τους.

Εναλλακτικά ενδείκνυται να αξιοποιηθούν γειτονιές πολύ μεγάλου μεγέθους

(VLSN=Very Large Scale Neighborhoods). Βέβαια ισχύει πως όσο πιο μεγάλη είναι η γειτονιά τόσο μεγαλύτερη είναι και η δυνατότητα εντοπισμού σωστών λύσεων. Η αργοπορία

σε κάθε επανάληψη η οποία παρατηρείται στην περίπτωση των VLSN ενδέχεται να καταλήγει σε χειρότερες λύσεις σε σχέση με μια μικρότερη γειτονιά. Γι ’αυτόν τον λόγο χρησιμοποιείται τελικά η χρήση εξειδικευμένων αλγόριθμων αναζήτησης.

## 3.8 Τεχνικές επίλυσης προβλημάτων βελτιστοποίησης

Υπάρχουν διάφορες τεχνικές που επιλύουν προβλήματα βελτιστοποίησης. Οι σημαντικότερες κατηγορίες τεχνικών βελτιστοποίησης είναι ο **μαθηματικός προγραμματισμός και οι μεταευρετικές τεχνικές**.

Δεν μπορούμε να πούμε πως έχει διαπιστωθεί ότι μια μέθοδος ή μια κατηγορία μεθόδων αντιμετωπίζει καλύτερα σε σχέση με τις υπόλοιπες το φάσμα των προβλημάτων βελτιστοποίησης. *Όπως έχει θεωρητικά αποδειχθεί από τους Wolpert και MacReady ισχύει υπό περιορισμούς ότι η μέση απόδοση όλων των αλγορίθμων βελτιστοποίησης επί όλων των προβλημάτων βελτιστοποίησης είναι ισοδύναμη. Το θεώρημα αυτό είναι γνωστό ως “No Free Lunch Theorem” και εκμηδενίζει τις ελπίδες να βρεθεί μια τεχνική η οποία να δίνει καλύτερα αποτελέσματα επί του συνόλου όλων των προβλημάτων βελτιστοποίησης.* (ΑΛΓΟΡΙΘΜΟΙ ΣΥΝΔΥΑΣΤΙΚΗΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΜΕ ΕΜΦΑΣΗ ΣΕ ΜΕΤΑΕΥΡΕΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΧΡΗΣΤΟΣ Γ. ΓΚΟΓΚΟΣ)

### 3.8.1 Μαθηματικός προγραμματισμός

Με την έννοια μαθηματικός προγραμματισμός απευθυνόμαστε σε τεχνικές βελτιστοποίησης οι οποίες βασίζονται σε μαθηματικά υψηλού επιπέδου και οι οποίες χρησιμοποιώντας τον γραμμικό προγραμματισμό έφτιαξαν μια κατηγορία υπολογιστικών μεθόδων κατάλληλη να επιλύει πολλά δύσκολα προβλήματα βελτιστοποίησης.

### 3.8.2 Γραμμικός Προγραμματισμός

Ο Γραμμικός Προγραμματισμός είναι ο κλάδος της Επιχειρησιακής Έρευνας που έχει ως στόχο τη μεγιστοποίηση η ελαχιστοποίηση γραμμικών συναρτήσεων υποκείμενους από ορισμένους περιορισμούς για τις μεταβλητές.

Ο καθένας από μας λύνει προβλήματα γραμμικού προγραμματισμού συνήθως χωρίς να το καταλαβαίνει. Όταν αγοράζουμε από το Πολυκατάστημα και υπολογίζουμε πως θα αγοράσουμε τα τρόφιμα που θέλουμε με τα χρήματα που διαθέτουμε, τότε λύνουμε ένα απλό ΠΓΠ. Όταν αποφασίζουμε ποια μέσα μεταφοράς θα χρησιμοποιήσουμε για να μην αργήσουμε στη δουλειά μας (ελαχιστοποίηση του χρόνου), ξανά τότε λύνουμε ένα μικρό ΠΓΠ.

Επίσης όταν ρυθμίζουμε την δουλειά μας ώστε να μένει χρόνος και για κάτι άλλο (διασκέδαση, ταξίδια, οικογένεια κλπ.) και τότε λύνουμε ένα πρόβλημα βελτιστοποίησης σύμφωνα με τα δεδομένα που έχουμε, την αντίληψή μας και τον στόχο που θέλουμε να καταφέρουμε.

Για να λύσουμε τα προβλήματα αυτά πιο συστηματικά, με μαθηματικές και υπολογιστικές μεθόδους πρώτα πρέπει να ορίσουμε μια μαθηματική διατύπωση, δηλαδή να φτιάξουμε το μοντέλο του προβλήματος. Τα προβλήματα που διατυπώνονται με γραμμικές σχέσεις μεταξύ των μεταβλητών (περιορισμοί, αντικειμενικός στόχος, ...) καλούνται Προβλήματα Γραμμικού Προγραμματισμού.

Ο Γραμμικός προγραμματισμός:

- Επιλύει, υπό ορισμένες προϋποθέσεις, το πρόβλημα κατανομής πεπερασμένων πόρων κατά τον καλύτερο δυνατό τρόπο (allocation problem).
- Επιλέγει την απόλυτη ομοιομορφία κάθε δραστηριότητας κατά τρόπο ώστε να επιτυγχάνεται η βελτιστοποίηση του αποτελέσματος.
- Επιχειρεί να πετύχει έναν σαφώς διατυπωμένο στόχο, ο οποίος εκφράζεται με τη βελτιστοποίηση της λεγόμενης «αντικειμενικής συνάρτησης».

Οι περιορισμοί:

- Υφίστανται τόσο στους διατιθέμενους πόρους (αγαθά), όσο και στις απαιτούμενες στάθμες των δραστηριοτήτων.
- Δεν προκαθορίζουν πλήρως ένα τρόπο ενέργειας αλλά αφήνουν περιθώρια για περισσότερες από μία εναλλακτικές δυνατότητες δράσης (λύσεις).

**Ο Γραμμικός Προγραμματισμός είναι χρήσιμος διότι:**

- Διάφορα πρακτικά προβλήματα είναι δυνατόν να μορφοποιηθούν σε προβλήματα γραμμικού προγραμματισμού.
- Υφίσταται αλγόριθμος (Simplex) ο οποίος επιτρέπει την επίλυση προβλημάτων γραμμικού προγραμματισμού αρκετά εύκολα.

**Περιοχές Εφαρμογής Γραμμικού Προγραμματισμού**

- Μίξη υλικών
- Προγραμματισμός παραγωγής
- Διαχείριση διύλισης πετρελαίου
- Διανομή
- Χρηματοοικονομικός προγραμματισμός
- Προγραμματισμός ανθρώπινου δυναμικού
- Προγραμματισμός καλλιέργειας
- Πολιτική Αγορών & Πωλήσεων

**Προϋποθέσεις εφαρμογής ΓΠ**

Γραμμικότητα:

- Η αντικειμενική συνάρτηση να είναι γραμμική συνάρτηση, δηλαδή η τιμή της να είναι ποσοτικά ανάλογη ως προς τις ποσότητες κάθε μιας από τις δραστηριότητες.

- Όλοι οι περιορισμοί ενός ΓΠ είναι γραμμικής μορφής, δηλαδή η χρησιμοποίηση των διαθέσιμων μέσων είναι ανάλογη ως προς τις ποσότητες κάθε μιας από τις δραστηριότητες.

Προσθετικότητα:

- Το συνολικό μέτρο αποτελεσματικότητας και κάθε συνολική χρησιμοποίηση διαθέσιμων μέσων, εφ' όσον προκύπτουν από την κοινή λειτουργία των δραστηριοτήτων, πρέπει να είναι ίσα προς το άθροισμα των αντίστοιχων ποσοτήτων, που προκύπτουν από κάθε μια δραστηριότητα όταν αυτή λειτουργεί χωριστά και ανεξάρτητα.

Διαιρετότητα:

- Συνήθως έχει φυσικό νόημα μόνο σε ακέραιες τιμές των μεταβλητών αποφάσεων.
- Δεν υπάρχει εγγύηση των μεθόδων επιλύσεως προβλημάτων ΓΠ για ακέραια λύση.
- Οι μεταβλητές αποφάσεων μπορούν να πάρουν πραγματικές τιμές.

Προσδιορισμένοι συντελεστές:

- Όλοι οι συντελεστές ενός προβλήματος ΓΠ θεωρούνται ως γνωστές σταθερές.

**ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ Γ.Π.**

- Λύση προβλήματος Γ.Π. : διάνυσμα των τιμών όλων των μεταβλητών απόφασης
- Εφικτή λύση (feasible solution): μία λύση όπου ικανοποιεί τους περιορισμούς του προβλήματος
- Πεδίο εφικτών λύσεων ή επιτρεπτή περιοχή: το σύνολο των εφικτών λύσεων
- Σύνορο εφικτού πεδίου: το κυρτό πολύγωνο που ορίζουν οι περιορισμοί του προβλήματος.
- Λύση ακραίου σημείου: η λύση που αντιστοιχεί σε μία γωνία του κυρτού πολυγώνου (τομή n περιορισμών)
- Άριστη λύση (optimal solution): η λύση που αριστοποιεί (μεγιστοποιεί ή ελαχιστοποιεί) την αντικειμενική συνάρτηση

### 3.8.3 Μέθοδος Simplex

Σε ρεαλιστικές εφαρμογές οι μεταβλητές και οι περιορισμοί των προβλημάτων γραμμικού προγραμματισμού είναι εκατοντάδες και μερικές φορές ακόμα και χιλιάδες. Υπάρχει η ανάγκη δηλαδή για μια συστηματική μέθοδο επίλυσης των προβλημάτων Γραμμικού Προγραμματισμού η οποία να καταφέρνει να εφαρμοστεί μέσω καταλλήλων προγραμμάτων των υπολογιστών για την επίλυση προβλημάτων Γραμμικού Προγραμματισμού κάθε μεγέθους. Η συστηματική αυτή μέθοδος είναι η **μέθοδος simplex**.

Λόγο ότι οι εισαγόμενες μεταβλητές αποκλίσεως έχουν μοναδιαίες αξίες ή μηδενικές, λέμε ότι οι αντίστοιχες δραστηριότητες είναι ουδέτερες. Το πρόβλημα που προκύπτει με τους περιορισμούς των εξισώσεων έχει το ίδιο σύνολο βέλτιστων λύσεων με το αρχικό. Η φυσική ερμηνεία των μεταβλητών αποκλίσεως είναι ότι παριστάνουν το μέρος του διαθεσίμου μέσου που μένει αχρησιμοποίητο.

Το 1947 ο George Dantzig ανέπτυξε τον αλγόριθμο Simplex για την βέλτιστη επίλυση προβλημάτων γραμμικού προγραμματισμού. Η επιτυχία του αλγορίθμου επιβεβαιώνεται από σχετική μελέτη η οποία ανέδειξε το γεγονός ότι 85% των εταιρειών που βρίσκονται στην λίστα Fortune 500 έχει χρησιμοποιήσει τον εν λόγω αλγόριθμο.(ΑΛΓΟΡΙΘΜΟΙ ΣΥΝΔΥΑΣΤΙΚΗΣ *ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΜΕ ΕΜΦΑΣΗ ΣΕΜΕΤΑΕΥΡΕΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΧΡΗΣΤΟΣ Γ. ΓΚΟΓΚΟΣ)*

Ο αλγόριθμος Simplex εφαρμόζεσαι όταν υπάρχει ένα πρόβλημα Γ.Π.και πρέπει να μετατραπεί σε κανονική μορφή δηλαδή κάθε μεταβλητή απόφασης να είναι μη αρνητική και όλοι οι περιορισμοί ναείναι ισότητες. Αυτό για να γίνει πρέπει να προστεθούν νέες μεταβλητές slack ή excess σε κάθε περιορισμό που ήταν ανισότητα έτσι ώστε να μετατραπεί σε ισότητα. Δηλαδή από ένα πρόβλημα με n μεταβλητές απόφασης και m περιορισμούς έχουμε μια αναπαράσταση με n+m μεταβλητές όλες με περιορισμό μη αρνητικότητας. Ο αλγόριθμος αρχικά εντοπίζει με κάποιο τρόπο (π.χ. μέθοδο του μεγάλου Μ)μια κορυφή του εφικτού συνόλου η οποία αποτελεί την βασική εφικτή λύση (basic feasible solution). Σε αυτή τη λύση ένα υποσύνολο από τις μεταβλητές του προβλήματος με πλήθος ίσο με τον αριθμό των περιορισμών m αποτελούν τις βασικές μεταβλητές ενώ οι υπόλοιπες μεταβλητές n σε πλήθος ονομάζονται ελεύθερες.

Οι n ελεύθερες μεταβλητές λαμβάνουν την τιμή μηδέν και προσδιορίζεται η τιμή των βασικών μεταβλητών από τις m διαθέσιμες εξισώσεις των περιορισμών λύνοντας ένα σύστημα εξισώσεων m x m.

Στην επόμενη φάση ο αλγόριθμος πηγαίνει από κορυφή σε κορυφή κατά μήκος των ακμών του εφικτού συνόλου. Αυτό συμβαίνει επαναλαμβάνοντας την αντικατάσταση μιας μη βασικής μεταβλητής με μια βασική μεταβλητή με στόχο την αποκόμιση του μεγαλύτερου δυνατού κέρδους για την συνάρτηση στόχο από αυτή την ενέργεια. Η επιλογή της μη βασικής μεταβλητής γίνεται εξετάζοντας το συντελεστή που έχει στην συνάρτηση στόχο ενώ η επιλογή της βασικής μεταβλητής που θα αντικατασταθεί επιλέγεται με βάση τον συντελεστή που έχει στον περιορισμό που συμμετέχει. Στο τέλος κάθε επανάληψης η βασική μεταβλητή που αντικαθίσταται λαμβάνει οριακή τιμή.

**Κανόνες μεθόδου SIMPLEX**

- Κριτήριο εισόδου της βάσης.

Επιλέγεται η μη-βασική μεταβλητή που θα αυξήσει περισσότερο την ΑΣ (κέρδος). Είναι αυτή που έχει το μεγαλύτερο συντελεστή στον υπολογισμό του P (ή μικρότερο αρνητικό στον πίνακα).

- Κριτήριο εξόδου από τη βάση

Επιλέγεται η βασική μεταβλητή που περιορίζει περισσότερο τη μη-βασική μεταβλητή που διαλέξαμε. Η εξερχόμενη μεταβλητή είναι αυτή που έχει τον μικρότερο θετικό λόγο.

- Κριτήριο Τερματισμού διαδικασίας

Όλα τα στοιχεία της γραμμής της αντικειμενικής συνάρτησης είναι θετικά ή μηδέν.

## 3.8.4 Ανάλυση ευαισθησίας

Η ανάλυση ευαισθησίας (sensitivity analysis) επιτρέπει να εξάγουμε με την μέθοδο Simplex επιπλέον πληροφορίες για την λύση που παράγεται αλλάζοντας οριακά τους συντελεστές των μεταβλητών απόφασης της συνάρτησης στόχου καθώς και τους όρους που βρίσκονται στα δεξιά μέρη των ανισοτήτων.

Κάθε γραμμικό πρόβλημα αναφέρεται ως πρωτογενές (primal) και έχει ένα συμπληρωματικό πρόβλημα που ονομάζεται δυϊκό (dual). Η λύση του ενός από τα δύο συνεπάγεται την λύση και του άλλου.

Η δυϊκή τιμή ενός περιορισμού που ονομάζεται και σκιώδης τιμή (shadow price) είναι ο ρυθμός αλλαγής της συνάρτησης στόχου καθώς το δεξί μέρος της ανισότητας χαλαρώνει. Η οικονομική ερμηνεία της δυϊκής τιμής ενός περιορισμού αναφέρει ότι υποδηλώνει το ποσό που είμαστε διατεθειμένοι να πληρώσουμε για χαλάρωση μιας μονάδας στον περιορισμό. Η δυϊκή τιμή μιας μεταβλητής που ονομάζεται και μειούμενο κόστος (reduced cost) είναι ο ρυθμός αλλαγής της συνάρτησης στόχου εάν αναγκάσουμε την μεταβλητή που είχε τιμή 0 στην βέλτιστη λύση να εισαχθεί σε μια νέα λύση.

(ΑΛΓΟΡΙΘΜΟΙ ΣΥΝΔΥΑΣΤΙΚΗΣ *ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΜΕ ΕΜΦΑΣΗ ΣΕ ΜΕΤΑΕΥΡΕΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΧΡΗΣΤΟΣ Γ. ΓΚΟΓΚΟΣ)*

## 3.8.5 Γραμμικός Προγραμματισμός & Δυαδικό Πρόβλημα

### Ορισμένες Γενικές Αρχές

Το δυαδικό πρόβλημα είναι ένα πρόβλημα ελαχιστοποιήσεως σε κανονική μορφή.

Όταν το πρωτεύον πρόβλημα έχει n μεταβλητές απόφασης, τότε το δυαδικό έχει n περιορισμούς.

Ο $1^{ος}$ περιορισμός του δυαδικού συσχετίζεται με την μεταβλητή x1 του πρωτεύοντος, ο 2ος με τη μεταβλητή x2 κ.ο.κ.

Όταν το πρωτεύον έχει m περιορισμούς, το δυαδικό έχει m μεταβλητές. Η μεταβλητή του δυαδικού u1συσχετίζεται με τον πρώτο περιορισμό του πρωτεύοντος κ.ο.κ.

Τα δεξιά τμήματα των περιορισμών του πρωτεύοντος προβλήματος γίνονται οι σταθερές της αντικειμενικής συνάρτησης στο δυαδικό πρόβλημα.

Οι σταθερές της αντικειμενικής συνάρτησης στο δυαδικό πρόβλημα γίνονται τα δεξιά τμήματα των περιορισμών του δευτερεύοντος.

Οι σταθερές των περιορισμών της μεταβλητής i του πρωτεύοντος προβλήματος γίνονται οι σταθερές στον περιορισμό i του δυαδικού προβλήματος.

### Σύνοψη Συσχέτισης Πρωτεύοντος & Δυαδικού

**Πρωτεύον Πρόβλημα**

-Πρόβλημα μεγιστοποιήσεως

-Συντελεστές αντικειμενικής συνάρτησης

-Σταθεροί όροι περιορισμών

-Τεχνολογική μήτρα

-Περιορισμοί

- i στη ανισότητα <=
- i στη ανισότητα =

Μεταβλητές

- xj>=0
- xj><0

**Δυαδικό Πρόβλημα**

-Πρόβλημα ελαχιστοποιήσεως

-Σταθεροί όροι περιορισμών Συντελεστές αντικειμενικής συνάρτησης

-Ανάστροφη Τεχνολογική μήτρα

-Μεταβλητές

- uj>=0
- uj><0

-Περιορισμοί •j

- στη ανισότητα >=•j

**Χρησιμότητα Θεωρίας Δυαδικότητας**

- Διευκολύνει την επίλυση του αντίστοιχου πρωτεύοντος προβλήματος σε ορισμένες περιπτώσεις.
- Παρέχει οικονομικές πληροφορίες για τα μεγέθη του πρωτεύοντος.
- Χρησιμεύει στην πραγματοποίηση αναλύσεων ευαισθησίας.

## 3.9 Ευρετικές τεχνικές

Στηρίζονται στην αρχή του Simon, δηλαδή πως οι επιχειρηματίες δεν αναζητούν πάντα την άριστη λύση σε μια απόφαση, αλλά μια αρκετά καλή λύση (satisficing).Ειδικά στη περίπτωση όπου για να φτάσει κανείς στην άριστη λύση συναντά πολύ μεγάλες δυσκολίες, τότε η ανάγκη να καταφύγει σε προσεγγιστικές (ή ευρετικές) τεχνικές που δίνουν μια καλή (και όχι αναγκαστικά άριστη) λύση είναι μεγάλη.

Οι ευρετικές τεχνικές είναι ιδιαίτερα κατάλληλες για μια μεγάλη κατηγορία προβλημάτων, τα οποία καλούνται ως μοντέλα του Ακέραιου Προγραμματισμού και ανήκουν στην κατηγορία των προβλημάτων που ονομάζονται np-complete. Οι περισσότερες από τις τεχνικές που επιλύουν προβλήματα εφοδιαστικής αλυσίδας ανήκουν στην κατηγορία των ευρετικών τεχνικών. Εφαρμόζεται στις περιπτώσεις που δεν μας ενδιαφέρει η άριστη αλλά μία «αρκετά καλή» στρατηγική.

Ο εντοπισμός του ευρετικού αλγορίθμου για να λυθεί το πρόβλημα Μαθηματικού (ακεραίου) προγραμματισμού εξαρτάται από:

- Τη φύση του προβλήματος

- Την εμπειρία του αναλυτή

Τυχαίνει πολλές φορές να υπάρχουν δύο ή περισσότεροι ευρετικοί αλγόριθμοι για ένα πρόβλημα.

### 3.9.1 Μεταευρετικές τεχνικές

Μια τεχνική επίλυσης προβλημάτων θεωρείται ως ευρετική (heuristic) όταν αναζητά καλές λύσεις που μπορούν να επιτευχθούν διαθέτοντας λογική υπολογιστική ισχύ. *(ΑΛΓΟΡΙΘΜΟΙ ΣΥΝΔΥΑΣΤΙΚΗΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΜΕ ΕΜΦΑΣΗ ΣΕΜΕΤΑΕΥΡΕΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΧΡΗΣΤΟΣ Γ. ΓΚΟΓΚΟΣ)*

Μας προβληματίζει ο παραπάνω ορισμός σε σχέση με την καταλληλότητα χρήσης ευρετικών τεχνικών για την επίλυση προβλημάτων βελτιστοποίησης γιατί δεν μπορούμε να πούμε ότι οι λύσεις που έχουμε είναι βέλτιστες.

Τελικά όμως αποτελούν τον πιο διαδεδομένο και πετυχημένο τρόπο αντιμετώπισης δύσκολών προβλημάτων.

Η βασική υπόθεση στην οποία βασίζονται οι ευρετικές τεχνικές είναι πως η αποτελεσματική λύση ορισμένων συνδυαστικών προβλημάτων επιζητά περισσότερη ευελιξία από ότι μπορεί να επιτευχθεί με τεχνικές που εξ αποτελέσματος παρουσιάζουν ιδιότητες σύγκλισης.

Ως ένα βαθμό θεωρούνται ως αλγόριθμοι αναζήτησης που εφαρμόζουν επαναληπτικά μια ιδέα επιλογής για την συνέχεια της αναζήτησης. Αυτό συμβαίνει μέσω κάποιας απλής λογικής «συνέχειας» ενεργειών καθώς και με τεχνικές πειραματισμού και διόρθωσης (trial and error).

Κάποιες ευρετικές τεχνικές που αξιοποιούνται για την εύρεση καλών λύσεων σε συγκεκριμένα προβλήματα ανεξαρτητοποιούνται από τα κοινά προβλήματα που επικαλούνται να επιλύσουν. Αυτές οι τεχνικές ονομάζονται **μεταευρετικές** και πολλές από αυτές βασίζονται από μηχανισμούς που έχουν παρατηρηθεί στην φύση. Οι μεταευρετικές τεχνικές έχουν τη δυνατότητα να ομαδοποιηθούν με βάση τον τρόπο με τον εξέτασης των νέων πιθανών λύσεων.

Μέθοδοι όπως η αναζήτηση ταμπού και η προσομοιωμένη ανόπτηση αναλύουν μόνο μια καινούργια λύση σε κάθε βήμα ενώ μέθοδοι όπως οι γενετικοί αλγόριθμοι και οι διασκορπισμένη αναζήτηση επεξεργάζονται παράλληλα ένα μεγάλο αριθμό από λύσεις.

Ένας άλλος τρόπος με τον οποίον κατηγοριοποιούνται οι μεταευρετικές τεχνικές είναι σε αυτές που χρησιμοποιούν και σε τεχνικές που δεν χρησιμοποιούν μνήμη. Στις πρώτες καταγράφονται τα χαρακτηριστικά ενδιάμεσων λύσεων που παράγονται με την εφαρμογή της μεθόδου με σκοπό να υποβοηθηθεί η λήψη μελλοντικών αποφάσεων ενώ στις τεχνικές που ανήκουν στην δεύτερη κατηγορία οι ενέργειες που εκτελούνται εξαρτώνται μόνο από την τρέχουσα κατάσταση.

### 3.9.2 Προσομοιωμένη ανόπτηση

Η πρώτη ιστορικά μεταευρετική τεχνική είναι η προσομοιωμένη ανόπτηση (SA= Simulated Annealing).

Είναι ανάλογη με αυτή της ανόπτησης όπως χρησιμοποιείται κατά την διαδικασία ψύχρανσης ενός υλικού το οποίο πρώτα έχει περιέλθει σε κατάσταση τήξεως. Συγκεκριμένα δομικά χαρακτηριστικά του υλικού μπορούν να επιτευχθούν ανάλογα με τον τρόπο με τον οποίο θα γίνει η ψύχρανση του υλικού. Αν το υλικό μετατραπεί σε κρύο νερό τότε η απότομη μείωση της θερμοκρασίας προκαλεί ατέλειες στο υλικό. Στην περίπτωση όμως που η μείωση

της θερμοκρασίας γίνει σταδιακά τότε σχηματίζονται μεγάλοι συμπαγείς κρύσταλλοι και το υλικό βρίσκεται σε κατάσταση ελάχιστης ενέργειας (ground state). Με αυτό το πρόβλημα ασχολήθηκαν οι Metro polis et al. όπου δημιούργησαν έναν αλγόριθμο για να προσομοιώσουν την διαδικασία.

Η αντιστοίχηση στα προβλήματα αντοχής υλικών και στα προβλήματα βελτιστοποίησης έγινε από τους Kirkpatrick et al. αναγνωρίζοντας την αναλογία μεταξύ του φορτίου ενέργειας ενός υλικού και στο κόστος μιας λύσης ενός συνδυαστικού προβλήματος.

Όπως στα υλικά η θερμοκρασία που είναι μια παράμετρος ελέγχου της διαδικασίας βοηθάει το σύστημα να οργανωθεί σε καταστάσεις μειωμένης ενέργειας έτσι και στα προβλήματα βελτιστοποίησης η ίδια παράμετρος οδηγεί σε εξερεύνηση καταστάσεων με ελαττωμένο κόστος.

Η μέθοδος προσομοιωμένης ανόπτησης αποτελεί μια μέθοδος τοπικής αναζήτησης. Παρόλα αυτά, η επιλογή μιας λύσεις δεν γίνεται μόνο με κριτήριο μείωσης του κόστους αλλά δεχόμαστε και λύσεις που άμεσα δεν βελτιώνουν το κόστος.

Το αν θα αποδεχθούμε υποδεέστερες λύσεις εξαρτάται από την απόσταση της καινούργιας λύσης από την καλύτερη λύση όπως και από το χρονικό διάστημα που έχει περάσει από την έναρξη της αναζήτησης.

Συνήθως στην εφαρμογή της προσομοιωμένης ανόπτησης χρησιμοποιείται ο τύπος exp(-δ/t) για να καθοριστεί η πιθανότητα με την οποία μια χειρότερη λύση κατά δ μονάδες σχετικά με την τρέχουσα καλύτερη λύση θα γίνει δεκτή όταν η θερμοκρασία έχει την τιμή t. Ακόμα, εάν ο αλγόριθμος λειτουργεί κάνοντας αποδεκτή με πιθανότητα 1 οποιαδήποτε νέα λύση είναι καλύτερη από την τρέχουσα καλύτερη λύση. Συμπεραίνουμε πως πρόκειται για μια στοχαστική μέθοδο με κύριο στόχο την αποφυγή εγκλωβισμού της αναζήτησης σε τοπικά βέλτιστα. Είναι αξιοσημείωτο ότι η συγκεκριμένη μέθοδος έχει μελετηθεί σημαντικά σε σχέση με την στατιστική της συμπεριφορά μέσω μαρκοβιανών αλυσίδων. Είναι αποδειγμένο ότι η μέθοδος συγκλίνει στην βέλτιστη λύση όμως απαιτείται πάρα πολύς χρόνος σε σχέση με το μέγεθος του προβλήματος.

## ΜΕΘΟΔΟΣ ΚΑΤΑΚΛΥΣΜΟΥ

Η μέθοδος κατακλυσμού (GD=GreatDeluge) είναι πολύ σχετική με την μέθοδο προσομοιωμένης ανόπτησης.

Δέχεται μόνο δύο παραμέτρους, τον χρόνο που ο χρήστης θέλει να διαθέσει και την ποιότητα της λύσης που απαιτείται.

Η μέθοδος GD αποδέχεται μια υποψήφια λύση αν ικανοποιούνται οι ακόλουθες συνθήκες:

P'<=B όταν το P < B

P'<=P όταν το P>=B

όπου P είναι το «κόστος» της τρέχουσας λύσης, P' είναι το «κόστος» της υποψήφιας λύσης και B είναι το τρέχον πάνω όριο. Αρχικά το όριο B είναι ίσο με την αρχικό «κόστος» και σε κάθε βήμα μειώνεται με κάποιο ρυθμό που αντιστοιχεί στην ταχύτητα αναζήτησης.

### 3.9.3 Αναζήτηση ταμπού TABU SEARCH

Η tabu search (TS) μέθοδος είναι ένας ευφυής τρόπος αναζήτησης που χρησιμοποιεί δομές δεδομένων την αποθήκευση πληροφοριών σε σχέση με τις αποφάσεις που παρατηρούνται κατά την αναζήτηση.

Αξιοποιεί αυτές τις πληροφορίες για να προχωρήσει την αναζήτηση σε καινούργιες περιοχές. Ως μέθοδος αποδίδεται στον F. Glover.

Το όνομα tabu ή taboo προέρχεται από το γεγονός πως οριοθετεί ορισμένες λύσεις που παρουσιάζονται κατά την αναζήτηση ως απαγορευμένες επιδιώκοντας την αποφυγή «κολλήματος» της διαδικασίας σε κάποια τοπικά βέλτιστη λύση. Είναι αξιοσημείωτο ότι η μέθοδος αναζήτησης ταμπού είναι κυρίως μια ντετερμινιστική μέθοδος αν και έχουν υπάρξει στοχαστικές παραλλαγές της όπως η μέθοδος Probabilistic Tabu Search.

Ο αλγόριθμος TS ξεκινά με μια αρχική λύση που παράγεται είτε τυχαία είτε ως αποτέλεσμα κάποιου άλλου αλγορίθμου, η οποία λαμβάνει ταυτόχρονα την θέση της τρέχουσας και της καλύτερης λύσης. Ταυτόχρονα μπαίνει σε αρχείο και η μνήμη tabu η οποία ενδέχεται να αποτελείται από πολλές επιμέρους λίστες.

Η μνήμη δεν αποτρέπει τις κινήσεις που έχουν ήδη γίνει από το να ξαναγίνουν αλλά αποτρέπει την αντιστροφή τους. Ένα σύνολο από γειτονικές λύσεις που προκύπτουν από αυτή μέσω ενός μετασχηματισμού καθορίζεται για την τρέχουσα λύση. Αφού εξεταστούν οι λύσεις αυτές επιλέγεται η λύση που ελαχιστοποιεί την συνάρτηση κόστους στηριζόμενες στην μνήμη ταμπού που αν χρειαστεί αποκλείει κάποιες λύσεις ή να εισάγει ποινή γι’ αυτές στην συνάρτηση κόστους περιθωριοποιώντας έτσι την επιλογή τους.

Η επιλεχθείσα λύση αντικαθιστά την καλύτερη λύση αν έχει καλύτερη τιμή συνάρτησης κόστους από αυτή. Σε κάθε βήμα επιλέγεται μια νέα τρέχουσα λύση ακόμα και αν δεν έχει καλύτερο κόστος σε σχέση με την τρέχουσα. Η μνήμη tabu ενημερώνεται αναλύοντας επιμέρους χαρακτηριστικά της τρέχουσας λύσης.

Τα χαρακτηριστικά αυτά μπορούν να είναι ιδιότητες της λύσης που αλλάζουν κατά τον μετασχηματισμό από την παλιά τρέχουσα λύση στην νέα. Οι πληροφορίες που εισάγονται στην μνήμη tabu δεν διατηρούνται συνεχώς αλλά καθώς η διαδικασία εξελίσσεται νέες τιμές αντικαθιστούν παλιότερες. Το μέγεθος της μνήμης είναι ιδιαίτερα σημαντικό για την απόδοση της μεθόδου ενώ έχουν προταθεί και μεταβλητά μεγέθη μνήμης.*(ΑΛΓΟΡΙΘΜΟΙ ΣΥΝΔΥΑΣΤΙΚΗΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΜΕ ΕΜΦΑΣΗ ΣΕΜΕΤΑΕΥΡΕΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΧΡΗΣΤΟΣ Γ. ΓΚΟΓΚΟΣ)*

Ο αλγόριθμος ταμπού προσέχει ιδιαίτερα στη περίπτωση όπου εντοπίζεται μια καλύτερη λύση σε σχέση με άλλες αλλά η μνήμη ταμπού την απορρίπτει. Σε αυτή την περίπτωση η λύση τις περισσότερες φορές γίνεται δεκτή ενώ η διαδικασία ονομάζεται aspiration. Το θέμα στην διαδικασία aspiration είναι ότι αν η επιστροφή σε ήδη υπάρχουσα λύση δεν μπορεί να συμβεί τότε η πληροφορία tabu μπορεί να αγνοηθεί.

Η μέθοδος TS έχει εφαρμοστεί με επιτυχία σε πολλά και σημαντικά προβλήματα. Τα τελευταία χρόνια πληθώρα άρθρων έχουν βγει στη δημοσιότητα προτείνοντας παραλλαγές της μεθόδου και την εφαρμογή της σε διάφορα προβλήματα συνδυαστικής βελτιστοποίησης.

Ακόμα παρατηρούμε πολλές φορές πως οι λύσεις που προκύπτουν βρίσκονται κοντά στη αποδεδειγμένη βέλτιστη λύση όταν αυτή υπάρχει ενώ η μέθοδος συχνά δίνει καλύτερα αποτελέσματα σε σχέση με εκείνα που λαμβάνονται από άλλες ευρετικές τεχνικές.

Τέλος, η μέθοδος TS θεωρείται ως μια μεταευρετική τεχνική που μπορεί να αξιοποιηθεί σε συνδυασμό με άλλες μεθόδους προκειμένου να οδηγήσει την αναζήτηση σε αποδοτικότερες περιοχές. Η μέθοδος TS έχει συνδυαστεί παλιότερα από διάφορους ερευνητές με την προσομοιωμένη ανόπτηση, τους γενετικούς αλγορίθμους, τα νευρωνικά δίκτυα, τον προγραμματισμό με περιορισμούς, και άλλες.

### 3.9.4 Άπληστη τυχαιοποιημένη προσαρμοστική διαδικασία αναζήτησης GRASP

Η μέθοδος GRASP (GRASP=Greedy Randomized Adaptive Search Procedure) είναι ένας μεταευρετικός τρόπος αναζήτησης που εφαρμόζεται μέσω της επαναληπτικής εκτέλεσης δύο φάσεων, της φάσης κατασκευής και της φάσης αναζήτησης. Στην φάση κατασκευής φτιάχνεται μια υποψήφια λύση που αποτελείται από τμήματα της λύσης που εντοπίζονται σε μια ειδικά διαμορφωμένη λίστα τμημάτων την Περιορισμένη Λίστα Υποψηφίων (Restricted Candidate List= RCL).

Στην φάση της αναζήτησης εφαρμόζεται τοπική αναζήτηση με σημείο έναρξης την υποψήφια λύση και το αποτέλεσμα της χρησιμοποιείται προκειμένου να ενημερωθεί η RCL. Οι δύο φάσεις επαναλαμβάνονται και η βέλτιστη λύση που παρατηρείται στο σύνολο είναι η προτεινόμενη λύση. Η μέθοδος GRASP συστήθηκε από τους Fe ο και Resente το 1995 ενώ οι πρώτες απόπειρες εφαρμογής της αφορούσαν το πρόβλημα set covering.

Τα πλεονεκτήματα της μεθόδου GRASP είναι η ευκολία αξιοποίησής της και η έλλειψη μεγάλου αριθμού παραμέτρων που πρέπει να προσδιοριστούν έτσι ώστε να χρησιμοποιηθεί κατάλληλα σε διάφορα προβλήματα βελτιστοποίησης.

### 3.9.5 Υβριδικές μεταευρετικές τεχνικές

Υβριδικές μεταευρετικές τεχνικές καλούνται οι τεχνικές εκείνες που συνδυάζουν περισσότερες από μια τεχνικές βελτιστοποίησης μια τουλάχιστον εκ των οποίων είναι μεταευρετικής φύσεως. Ο συνδυασμός τεχνικών έχει οφέλη στην ποιότητα και την ταχύτητα εύρεσης της παραγόμενης λύσης του προβλήματος. Παλιότερα επιστημονικοί κλάδοι οι οποίοι παρήγαγαν και εξέλισσαν μεθόδους βελτιστοποίησης όπως η επιχειρησιακή έρευνα και η τεχνητή νοημοσύνη δεν συνήθιζαν να ενσωματώσουν μεταευρετικές τεχνικές στις μεθόδους τους. Αυτό είναι πιθανό να οφείλεται στην ανωριμότητα των ίδιων των τεχνικών, στην έλλειψη υπολογιστικής ισχύος, κ.α. Ωστόσο πλέον στις μέρες μας πιστεύεται ότι τα οφέλη είναι πολλαπλά από τη χρήση υβριδικών τεχνικών.

Οι υβριδικές μεταευρετικές τεχνικές χωρίζονται στις εξής δύο κατηγορίες ανάλογα με τη προέλευση των επιμέρους τεχνικών που συνθέτουν. Η πρώτη κατηγορία έχει να κάνει με τον συνδυασμό δύο ή περισσοτέρων μεταευρετικών τεχνικών. Για παράδειγμα συνδυάζοντας μια τεχνική που διατηρεί ένα πληθυσμό λύσεων (π.χ. γενετικοί αλγόριθμοι, διασκορπισμένη αναζήτηση, κ.α.) με μια τεχνική που είναι ικανή να εξερευνήσει αποδοτικά τον χώρο γύρω από τις λύσεις (π.χ. επαναλαμβανόμενη αναζήτηση, αναζήτηση μεταβαλλόμενης γειτονιάς κ.α.) πετυχαίνεται από τη μια ο εντοπισμός νέων «καλών» περιοχών του χώρου αναζήτησης

και από την άλλη η αναγνώριση των τοπικών ελαχίστων των περιοχών στις οποίες δρα η διαδικασία.

Η δεύτερη κατηγορία υβριδικών τεχνικών έχει να κάνει με τον συνδυασμό μεταευρετικών τεχνικών και κλασσικών τεχνικών από τον μαθηματικό προγραμματισμό και την τεχνητή νοημοσύνη. Κυρίως για τον μαθηματικό προγραμματισμό προσφάτως διάφορα επιστημονικά γεγονότα όπως τα workshops “Mathematical Contributions to Metaheuristics” του 2006 και του 2008 δείχνουν την πορεία που υπάρχει πλέον στις επιστήμες. Επίσης έχει εισαχθεί ο όρος “matheuristics” για την περιγραφή συνδυασμών μεταξύ μεταευρετικών τεχνικών και μαθηματικού προγραμματισμού.

Παράδειγμα συνδυασμού μεταευρετικών τεχνικών και μαθηματικού προγραμματισμού είναι ο εντοπισμός άνω ορίων της λύσης μέσω του εντοπισμού καλών πλήρων λύσεων του προβλήματος. Ο ταχύς εντοπισμός άνω ορίων από μια μεταευρετική τεχνική κατευθύνει στην αποκοπή τμημάτων του Branchand Bound δένδρου κατά την απαρίθμηση των κόμβων του και συνεπώς στον γρηγορότερο εντοπισμό λύσεων.

Το παραπάνω παράδειγμα δεν είναι ο μοναδικός τρόπος συνεργασίας μεταευρετικών και κλασσικών τεχνικών. Αναλυτικότερα σε σχέση με διάφορους τρόπους που συνδιάζονταιοι μεταευρετικές τεχνικές και ο μαθηματικός προγραμματισμός γίνεται από τους Raidl και Puchinger.

Μία άλλη κατηγορία των υβριδικών μεταευρετικών τεχνικών χωρίζει σε τεχνικές που στηρίζονται στην απλή συνεργασία επιμέρους τεχνικών (collaborative combinations) και σε τεχνικές που συνθέτουν ένα ολοκληρωμένο σύνολο στο οποίο κάθε τεχνική αποτελεί αναπόσπαστο κομμάτι της διαδικασίας επίλυσης επιφορτισμένο με αυστηρά καθορισμένο ρόλο (integrative combinations).Στην πρώτη κατηγορία η εκτέλεση των επιμέρους τεχνικών ενδέχεται να γίνεται σειριακά ή και παράλληλα. Στην δεύτερη κατηγορία μπορεί είτε ένας επακριβής αλγόριθμος να ενσωματώνεται σε μια μεταευρετική τεχνική ή το αντίθετο.

Ο τομέας της επιχειρησιακής έρευνας είναι ευρύς και έχει μελετηθεί και αναλυθεί από πολλούς επιστήμονες και επαγγελματίες. Εδώ έγινε μία σύνοψη για το τι είναι και πως αναλύεται και χωρίζεται κυρίως σε σχέση με τις μεθόδους βελτιστοποίησης που αφορούν αυτήν εδώ την εργασία στον συνδυασμό δηλαδή με τον τομέα της εξόρυξης δεδομένων.

# Κεφάλαιο 4ο - Η συνεισφορά της επιχειρησιακής έρευνας στην εξόρυξη δεδομένων

Η εξόρυξη δεδομένων ξεκινά με ένα σύνολο δεδομένων που ονομάζεται σύνολο εκπαίδευσης (training set) και αποτελείται από τις περιπτώσεις που περιγράφουν τις παρατηρούμενες τιμές συγκεκριμένων μεταβλητών ή χαρακτηριστικών. Οι περιπτώσεις αυτές χρησιμοποιούνται στη συνέχεια για να μάθουν μια συγκεκριμένη έννοια ή μοτίβο και, ανάλογα με τη φύση αυτής της έννοιας, εφαρμόζονται διαφορετικοί επαγωγικοί αλγόριθμοι. Οι πιο κοινές έννοιες που μαθαίνει κάποιος στην εξόρυξη δεδομένων είναι η κατηγοριοποίηση, η συσταδοποίηση-ομαδοποίηση δεδομένων και οι κανόνες συσχέτισης.

Στην κατηγοριοποίηση, επισημαίνονται τα δεδομένα της εκπαίδευσης, δηλαδή, κάθε παράδειγμα προσδιορίζεται ότι ανήκει σε μία από τις δύο ή περισσότερες κατηγορίες (classes) και ένας επαγωγικός αλγόριθμος χρησιμοποιείται για να δημιουργηθεί ένα μοντέλο που εισάγει την διάκριση μεταξύ των τιμών μιας κατηγορίας. Το μοντέλο μπορεί να χρησιμοποιηθεί για να ταξινομεί κάθε νέα περίπτωση, σύμφωνα με το χαρακτηριστικό της κατηγορίας. Συνήθως, ο κύριος στόχος είναι η ταξινόμηση να είναι όσο το δυνατόν ακριβέστερη, αλλά τα ακριβή μοντέλα δεν είναι απαραιτήτως χρήσιμα ή ενδιαφέροντα και άλλα μέτρα, όπως η απλότητα και η καινοτομία είναι επίσης σημαντικά.

Στην συσταδοποίηση των δεδομένων και στους κανόνες συσχέτισης δεν υπάρχει χαρακτηριστικό της κατηγορίας και τα δεδομένα είναι χωρίς επισήμανση (unlabelled). Για τις δύο αυτές προσεγγίσεις, διδάσκονται μοτίβα σε μία από τις δύο διαστάσεις της βάσης δεδομένων, δηλαδή, το χαρακτηριστικό της διάστασης και το παράδειγμα της διάστασης. Συγκεκριμένα, η συσταδοποίηση δεδομένων περιλαμβάνει τον προσδιορισμό του ποια παραδείγματα δεδομένων ανήκουν σε φυσικές ομάδες ή συστάδες, ενώ οι κανόνες συσχέτισης διδάσκουν τις σχέσεις ανάμεσα στα χαρακτηριστικά.

## 4.1. Μέθοδοι βελτιστοποίησης για την εξόρυξη δεδομένων

Ένα βασικό σημείο τομής της εξόρυξης δεδομένων και της επιχειρησιακής έρευνας είναι η χρήση αλγορίθμων βελτιστοποίησης, είτε άμεσα, ως αλγόριθμοι εξόρυξης δεδομένων, ή για να συντονίσουν παραμέτρους άλλων αλγορίθμων. Η βιβλιογραφία στον τομέα αυτό χρονολογείται από το 1965 με το έργο του Mangasarian (1965) όπου το πρόβλημα του διαχωρισμού δύο κατηγοριών είχε διατυπωθεί ως ένα γραμμικό πρόγραμμα. Εξακολουθεί να είναι ένας ενεργός τομέας της έρευνας από τότε και το ενδιαφέρον έχει αυξηθεί ραγδαία τα τελευταία χρόνια με την αυξανόμενη δημοτικότητα της εξόρυξης δεδομένων (βλ. π.χ., Glover, 1990, Mangasarian, 1994, Bennett και Bredensteiner,1999, Borosetal, 2000?. Felici &Truemper, 2002, Street, 2005).

Στην ενότητα αυτή, θα εξετάσουμε εν συντομία διαφορετικούς τύπους μεθόδων βελτιστοποίησης που χρησιμοποιούνται συνήθως για την εξόρυξη δεδομένων μεταξύ των οποίων τη χρήση του μαθηματικού προγραμματισμού για να διαμορφωθούν μηχανές διανυσμάτων και της μεταευρετικής (γενετικοί αλγόριθμοι).

### 4.1.1. Μαθηματικός προγραμματισμός και μηχανές διανυσμάτων υποστήριξης

Ένα από τα γνωστά σημεία διασταύρωσης της βελτιστοποίησης και της εξόρυξης δεδομένων είναι η διαμόρφωση των μηχανών διανυσμάτων υποστήριξης (support vector machines SVM), ως πρόβλημα βελτιστοποίησης. Οι μηχανές διανυσμάτων υποστήριξης έχουν την προέλευσή τους στην δημιουργική εργασία των Vapnik και Lerner (1963), αλλά μόλις πρόσφατα αποτέλεσαν το ενδιαφέρον των κοινοτήτων εξόρυξης δεδομένων και μηχανικής μάθησης (machine learning).

Στην πρώτη εργασία μαθηματικού προγραμματισμού που σχετίζεται με αυτόν τον τομέα, ο Mangasarian (1965) δείχνει πώς να χρησιμοποιείται ο γραμμικός προγραμματισμός για την απόκτηση γραμμικών και μη γραμμικών διακριτών μοντέλων μεταξύ διαχωρίσιμων σημείων των δεδομένων και αρκετοί συγγραφείς έχουν δουλέψει πάνω σε αυτό. Η ιδέα της απόκτησης μιας γραμμικής διάκρισης απεικονίζεται στο σχήμα 1. Το πρόβλημα εδώ είναι ο καθορισμός του καλύτερου μοντέλου για το διαχωρισμό των δύο κατηγοριών.

Εάν τα δεδομένα μπορούν να διαχωριστούν από ένα υπερεπίπεδο Η, όπως στο σχήμα. 1, το πρόβλημα μπορεί να λυθεί σχετικά εύκολα. Για να το διατυπώσουμε μαθηματικά, υποθέτουμε ότι το χαρακτηριστικό της κατηγορίας $y_i$ παίρνει δύο τιμές, -1 ή +1. Υποθέτουμε ότι όλα τα χαρακτηριστικά εκτός από το χαρακτηριστικό της κατηγορίας έχουν πραγματικές τιμές και υποδηλώνουν τα στοιχεία κατάρτισης (training data), που αποτελούνται από n περιπτώσεις, όπως $\{(a_J, y_J)\}$, όπου $j = 1, 2, \ldots, n$, $y_J \in \{-1, +1\}$ και $a_j \in R^m$.

Εάν υπάρχει ένα διαχωριστικό υπερεπίπεδο τότε υπάρχουν γενικά πολλά τέτοια υπερεπίπεδα και ορίζουμε το βέλτιστο υπερεπίπεδο που χωρίζει ως αυτό που μεγιστοποιεί το άθροισμα των αποστάσεων από το επίπεδο στο πλησιέστερο θετικό παράδειγμα και στο πλησιέστερο αρνητικό παράδειγμα. Για να μάθουμε αυτό το βέλτιστο υπερεπίπεδο σχηματίζουμε τα κυρτά τμήματα των δύο ομάδων δεδομένων (τα θετικά και τα αρνητικά παραδείγματα), βρίσκουμε τα πλησιέστερα σημεία cκαι d στα κυρτά τμήματα και στη συνέχεια αφήνουμε το βέλτιστο υπερεπίπεδο να αποτελεί το υπερ επίπεδο που τέμνει την ευθεία γραμμή μεταξύ c καιd. Αυτό μπορεί να διατυπωθεί ως εξής:

$$\min \frac{1}{2} t\|c-d\|^2$$

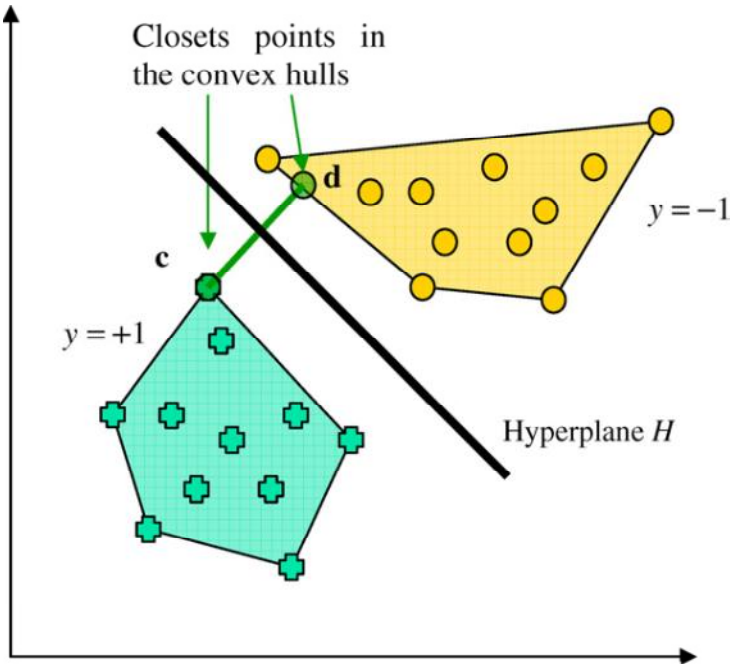
$$s.tc = \sum_{i:yi=+1} aiαi$$

$$d = \sum_{i:yi=-1} aiαi$$

$$\sum_{i:yi=+1} aiαi = 1$$

$$\sum_{i:yi=+1} aiαi = -1$$

$$ai \geq 0$$



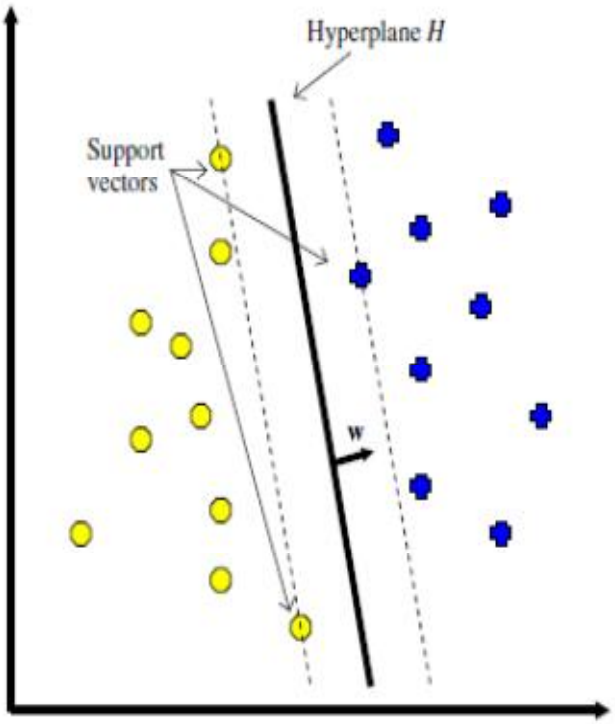

Εικ. 1. Ένα υπερεπίπεδο για τον διαχωρισμό των δύο κατηγοριών.

Το υπερεπίπεδο H μπορεί να οριστεί με όρους της μονάδας w και της απόστασης b από την αφετηρία (σχήμα. 1). Με άλλα λόγια, H = {x € $R^m$: x w + b = 0}, όπου x w είναι το γινόμενο μεταξύ των δύο αυτών ανυσμάτων. Για να δώσουμε μια εικόνα των διανυσμάτων υποστήριξης και των μηχανών διανυσμάτων υποστήριξης μπορούμε να φανταστούμε τα δύο υπερεπίπεδα, παράλληλα στο αρχικό επίπεδο και αφού έχουν την ίδια κανονική (normal), ωθούνται προς τη μια ή την άλλη κατεύθυνση μέχρι που απαντάται το κυρτό τμήμα από τα σύνολα όλων των περιπτώσεων με κάθε ταξινόμηση.

Αυτό θα συμβεί σε ορισμένες περιπτώσεις ή σε ή διανύσματα, γι' αυτό το λόγο ονομάζονται υποστηρικτές διανυσμάτων (βλ. σχήμα. 2). Αυτή η διαδικασία αποτυπώνεται μαθηματικά, απαιτώντας τους ακόλουθους περιορισμούς για να πραγματοποιηθεί:

$a_i$ w + b ≥ +1, $\forall_i$: $y_i$ = + 1

$a_i$ w + b ≤ -1, $\forall_i$: $y_i$ = - 1




Εικ. 2 Απεικόνιση μηχανών διανυσμάτων υποστήριξης (support vector machines SVM)

Με αυτή τη διατύπωση, η απόσταση μεταξύ των δύο επιπέδων που ονομάζεται περιθώριο, φαίνεται ότι είναι 2/||w|| και το βέλτιστο επίπεδο μπορεί επομένως να βρεθεί επιλύοντας το ακόλουθο μαθηματικό πρόβλημα βελτιστοποίησης που μεγιστοποιεί το περιθώριο:

$\text{Max}_{w,b} \|w\|^2$

$a_i\, w + b \geq +1, \forall_i: y_i = +1$

$a_i w + b \leq -1, \quad \forall_i: y_i = -1 (3)$

Όταν τα δεδομένα είναι μη διαχωρίσιμα, το πρόβλημα αυτό δεν έχει καμία εφικτή λύση και οι περιορισμοί για την ύπαρξη των δύο υπερεπιπέδων πρέπει να είναι χαλαροί.

Ένας τρόπος για να επιτευχθεί αυτό είναι με την εισαγωγή μεταβλητών σφάλματος $e_j$, για κάθε περίπτωση $a_j$, j =1, 2,. . . , n. Ουσιαστικά, αυτές οι μεταβλητές μετρούν την παραβίαση κάθε περίπτωσης και χρησιμοποιώντας αυτές τις μεταβλητές, οι παρακάτω τροποποιημένοι περιορισμοί χρησιμοποιούνται στο πρόβλημα

$a_i\, w + b \geq +1 - \varepsilon_i, \forall_i: y_i = +1$

$a_i\, w + b \leq -1 + \varepsilon_i, \forall_i: y_i = -1 (3)$

$\varepsilon_i \geq 0, \forall_I$

Δεδομένου ότι οι μεταβλητές $e_j$ αντιπροσωπεύουν το σφάλμα της κατάρτισης, ο στόχος θα μπορούσε να είναι η ελαχιστοποίηση του $\|w\|/2 + C\sum jej$ όπου C είναι σταθερά που μέτρα πόση ποινή δίνεται. Ωστόσο, αντί να διατυπώνεται το πρόβλημα άμεσα φαίνεται πως είναι καλύτερα να διατυπώνεται το ακόλουθο δυαδικό πρόβλημα:

$\max \sum_i ai - \frac{1}{2} \sum ij \alpha_i \alpha_j y_i y_j a_i a_j$

$0 \leq ai \leq C$

$\sum_i \alpha iyi = 0$

Η λύση στο πρόβλημα αυτό είναι οι διπλές μεταβλητές α και η επίτευξη της αρχικής λύσης, δηλαδή, το μοντέλο που ταξινομεί περιπτώσεις που καθορίζονται από το υπερεπίπεδο, υπολογίζουμε

$w = \sum_i \alpha iyia_i$

Το όφελος από τη χρήση του διπλού είναι ότι οι περιορισμοί είναι πολύ πιο απλοί και πιο εύκολοι στο χειρισμό και τα δεδομένα της εκπαίδευσης μπαίνουν μόνο μέσω του γινομένου aiaj. Το τελευταίο αυτό σημείο είναι σημαντικό για την επέκταση της προσέγγισης σε μη γραμμικό μοντέλο. Η απαίτηση ενός υπερεπιπέδου ή μιας γραμμικής διάκρισης σημείων είναι σαφώς πάρα πολύ περιοριστική για τα περισσότερα προβλήματα.

Ευτυχώς, η προσέγγιση μηχανών διανυσμάτων υποστήριξης (SVM) μπορεί να επεκταθεί σε μη γραμμικά μοντέλα με πολύ απλό τρόπο χρησιμοποιώντας αυτό που ονομάζεται λειτουργίες πυρήνα (kernel functions) $K(x, y)=\varphi(x)\ \varphi(y)$, όπου $\varphi = R^m \rightarrow H$ αποτελεί μια χαρτογράφηση (mapping) από το m-διάστατο ευκλείδειο χώρο έως κάποιο χώρο H του Hilbert. Η προσέγγιση αυτή εισήχθη στην SVM βιβλιογραφία από τους Cortes και Vapnik(1995) και λειτουργεί διότι τα δεδομένα a j εισέρχονται στη διπλή μόνο μέσω του γινομένου aiaj, το οποίο κατά συνέπεια μπορεί να αντικατασταθεί με K(ai, αj). Η επιλογή του πυρήνα καθορίζει το μοντέλο. Για παράδειγμα, για να χωρέσει ένα p βαθμού πολυώνυμο ο πυρήνας μπορεί να επιλεγεί ως $K(x, y) =(x\ y+ 1)^p$.Πολλές άλλες επιλογές έχουν ληφθεί υπόψη στη βιβλιογραφία, αλλά δεν θα τις διερευνήσουμε εδώ. Λεπτομερείς εκθέσεις των SVM μπορούν να βρεθούν στο βιβλίο του Vapnik (1995) και στις έρευνες του Burges (1998) και των Bennett και Campbell (2000).

## 4.1.2. Μεταευρετική για συνδυαστική βελτιστοποίηση

Πολλά προβλήματα βελτιστοποίησης που προκύπτουν στην εξόρυξη δεδομένων είναι μάλλον διακριτά και όχι συνεχή και πολυάριθμες διατυπώσεις συνδυαστικής βελτιστοποίησης έχουν προταθεί για την επίλυση τέτοιων προβλημάτων.

Αυτό περιλαμβάνει για παράδειγμα την επιλογή χαρακτηριστικού, δηλαδή, το πρόβλημα του προσδιορισμού του καλύτερου συνόλου γνωρισμάτων ώστε να χρησιμοποιηθεί από τον αλγόριθμο μάθησης, τον καθορισμό της βέλτιστης δομής ενός δικτύου Bayesian στην ταξινόμηση και την εξεύρεση της βέλτιστης συσταδοποίησης περιπτώσεων των δεδομένων. Ειδικότερα, πολλές προσεγγίσεις μεταευρετικής έχουν προταθεί για την αντιμετώπιση τέτοιων προβλημάτων.

Η μεταευρετική είναι η προτιμώμενη μέθοδος σε σχέση με άλλες μεθόδους βελτιστοποίησης κυρίως όταν υπάρχει ανάγκη να βρεθούν λύσεις σε πολύπλοκα προβλήματα βελτιστοποίησης με πολλά τοπικά βέλτιστα και μικρή εσωτερική δομή για να κατευθύνει την έρευνα (Glover και Kochenberger, 2003). Πολλά τέτοια προβλήματα ανακύπτουν στο πλαίσιο εξόρυξης δεδομένων. Η μεταευρετική προσέγγιση για την επίλυση τέτοιων προβλημάτων είναι να ξεκινήσουμε με την απόκτηση μιας αρχικής λύσης ή ενός αρχικού συνόλου λύσεων και στη συνέχεια να ξεκινήσουμε μια βελτιωμένη αναζήτηση η οποία καθοδηγείται από ορισμένες αρχές.

Η δομή της έρευνας έχει πολλά κοινά στοιχεία σε διάφορες μεθόδους. Σε κάθε βήμα του αλγορίθμου αναζήτησης, υπάρχει πάντοτε μια λύση $X_\kappa$ (ή ένα σύνολο

λύσεων), που αντιπροσωπεύει την τρέχουσα κατάσταση του αλγορίθμου. Πολλές μέθοδοι μεταευρετικής αποτελούν μέθοδοι αναζήτησης λύση προς λύση (solution-to-solution search methods), δηλαδή, το $X_κ$ είναι μια λύση ή σημείο $X_κ$ € X σε κάποιο χώρο επίλυσης X, που αντιστοιχεί στις εφικτές περιοχές. Άλλες καθορίζονται από το σύνολο, δηλαδή, σε κάθε βήμα το $X_κ$ αντιπροσωπεύει ένα σύνολο λύσεων $X_κ \subseteq X$ Ωστόσο, η βασική δομή της αναζήτησης παραμένει η ίδια ανεξάρτητα από το αν η μεταευρετική είναι λύση προς λύση ή καθορίζονται από το σύνολο.

Ο λόγος για το μετα-πρόθεμα (meta-prefix) είναι ότι η μεταευρετική δεν καθορίζει όλες τις λεπτομέρειες της αναζήτησης, η οποία έτσι μπορεί να προσαρμοστεί από μία τοπική ευρετική μέθοδο σε μία συγκεκριμένη δεδομένα εφαρμογής εξόρυξης. Αντίθετα, καθορίζει γενικές στρατηγικές για την καθοδήγηση συγκεκριμένων πτυχών της έρευνας. Για παράδειγμα, η αναζήτηση ταμπού χρησιμοποιεί μια λίστα λύσεων ή κινήσεων που ονομάζεται ο κατάλογος ταμπού, το οποίο διασφαλίζει ότι η αναζήτηση δεν επισκέπτεται ξανά πρόσφατα λύσεις ή δεν παγιδεύεται σε τοπικά βέλτιστα. Ο κατάλογος ταμπού μπορεί επομένως να θεωρηθεί ως περιορισμός της γειτονιάς.

Από την άλλη, μέθοδοι όπως ο γενετικός αλγόριθμος προσδιορίζουν τη γειτονιά, όπως όλες οι λύσεις που μπορούν να λαμβάνονται με συνδυασμό των τρεχουσών λύσεων μέσω ορισμένων χειριστών. Άλλες μέθοδοι, όπως η προσομοιωμένη ανόπτηση (simulated annealing), δεν προσδιορίζουν τη γειτονιά κατά κανένα τρόπο, αλλά μάλλον διευκρινίζουν μια προσέγγιση για την αποδοχή ή την απόρριψη λύσεων που επιτρέπουν στην μέθοδο να ξεφεύγει από τοπικά βέλτιστα. Τέλος, η μέθοδος των ένθετων κατατμήσεων (nested partitions)αποτελεί ένα παράδειγμα μιας μεθόδου που βασίζεται στα σύνολα και επιλέγει υποψήφιες λύσεις από την γειτονιά με κατανομή πιθανοτήτων η οποία προσαρμόζεται καθώς η αναζήτηση προχωράει ώστε να επιλέγονται οι καλύτερες λύσεις με μεγαλύτερη πιθανότητα.

Όλη η μεταευρετική μπορεί να θεωρηθεί ότι μοιράζει τα στοιχεία της επιλογής υποψήφιων λύσεων από μια γειτονιά τρεχουσών λύσεων και, στη συνέχεια, είτε γίνονται αποδεκτές, είτε απορρίπτονται.

Με την προοπτική αυτή, κάθε μεταευρετική καθορίζεται προσδιορίζοντας ένα ή περισσότερα από αυτά τα στοιχεία, αλλά επιτρέποντας και σε τρίτους να προσαρμόζονται στην ιδιαίτερη εφαρμογή. Αυτό μπορεί να θεωρηθεί τόσο ως δύναμη όσο και ως αδυναμία. Συνεπάγεται ότι μπορούμε να εκμεταλλευτούμε την ειδική δομή για κάθε εφαρμογή, αλλά επίσης σημαίνει ότι ο χρήστης πρέπει να προσδιορίζει πλευρές της εφαρμογής , πράγμα που μπορεί να είναι περίπλοκο. Στη συνέχεια εξετάζουμε τα πιο συνηθισμένα στοιχεία της μεταευρετικής και πώς ταιριάζουν σε αυτό το πλαίσιο.

Μία από τις πρώτες μεθόδους μεταευρετικής είναι η προσομοιωμένη ανόπτηση (Kirkpatrick et al., 1983), η οποία λειτουργεί με την φυσική διαδικασία ανόπτησης, αλλά εδώ απλά διερευνούμε μια μέθοδο για να καθορίσουμε αν μία λύση θα πρέπει να γίνεται δεκτή. Ως μέθοδος αναζήτησης λύση προς λύση, σε κάθε βήμα επιλέγει έναν υποψήφιο $X^c$ από τη γειτονιά $N(X_K)$ της τρέχουσας λύσης $X_κ$ € X.

Ο ορισμός της γειτονιάς καθορίζεται από το χρήστη. Αν η υποψήφια λύση είναι καλύτερη από την τρέχουσα λύση τότε γίνεται αποδεκτή. Αν είναι χειρότερη δεν απορρίπτεται αυτομάτως, αλλά γίνεται δεκτή με πιθανότητα P P[Accept $X^c$] = $e^{f(xk)-f(xc)}/Tk$,όπουf: X → R είναι μια πραγματικά αντικειμενική συνάρτηση που πρέπει να ελαχιστοποιηθεί και ***Tk*** είναι μια παράμετρος που ονομάζεται θερμοκρασία. Σαφώς, η πιθανότητα αποδοχής είναι υψηλή, αν η διαφορά είναι μικρή και το ***Tk*** είναι μεγάλο. Το κλειδί στην προσομοιωμένη ανόπτηση είναι να οριστεί ένα χρονοδιάγραμμα ψύξης$\{Tk\}^{\infty}_{k=1}$με το οποίο η θερμοκρασία να μειώνεται, έτσι ώστε αρχικά να επιλέγονται

κατώτερες λύσεις με μια αρκετά υψηλή πιθανότητα για να αποφεύγονται τα τοπικά βέλτιστα αλλά τελικά γίνεται αρκετά μικρή ώστε ότι ο αλγόριθμος να συγκλίνει. Η προσομοιωμένη ανόπτηση έχει για παράδειγμα χρησιμοποιηθεί για να λύσει το πρόβλημα επιλογής χαρακτηριστικών στην εξόρυξη δεδομένων .

Άλλες δημοφιλείς λύση προς λύση μέθοδοι μεταευρετικής περιλαμβάνουν την αναζήτηση ταμπού (tabu search), την άπληστη τυχαιοποιημένη προσαρμοστική διαδικασία αναζήτησης (GRASP – greedy randomized adaptive search procedure) και η μεταβλητή αναζήτηση γειτονίας (VNS – variable neighborhood search). Το καθοριστικό χαρακτηριστικό της αναζήτησης ταμπού είναι στο πώς οι λύσεις επιλέγονται από τη γειτονιά. Σε κάθε βήμα του αλγορίθμου, υπάρχει ένας κατάλογος $L_k$ λύσεων που πρόσφατα επισκέφθηκε και ως εκ τούτου αποτελούν ταμπού. Ο αλγόριθμος εξετάζει όλες τις λύσεις της γειτονιάς που δεν είναι ταμπού και επιλέγει τη βέλτιστη. Η καθοριστική ιδιότητα του GRASP είναι η προσέγγιση multistart (πολλαπλά ξεκινήματα), η οποία προετοιμάζει διάφορες διαδικασίες τοπικής αναζήτησης από διαφορετικές αφετηρίες. Το πλεονέκτημα είναι ότι από τη μια πλευρά η αναζήτηση γίνεται όλο και πιο σφαιρική, αλλά από την άλλη πλευρά κάθε αναζήτηση δεν μπορεί να χρησιμοποιήσει ότι έχουν μάθει οι άλλες αναζητήσεις, η οποία εισάγει κάποια αναποτελεσματικότητα. Η VNS είναι ενδιαφέρουσα υπό την έννοια ότι χρησιμοποιεί μια προσαρμοζόμενη διάρθρωση γειτονιάς που αλλάζει με βάση τις επιδόσεις των λύσεων που αξιολογούνται. Περισσότερες πληροφορίες σχετικά με ταμπού αναζήτησης μπορούν να βρεθούν στους Glover και Laguna (1997), η GRASP εξετάζεται από τους Resende και Ribeiro(2003), και για μια εισαγωγή στην προσέγγιση VNS παραπέμπουμε στους Hansen και Mladenovic(1997).

Αρκετές μέθοδοι μεταευρετικής καθορίζονται με βάση τον πληθυσμό παρά με τη μέθοδο λύση προς λύση. Αυτό περιλαμβάνει γενετικούς αλγόριθμους και άλλες εξελικτικές προσεγγίσεις καθώς και την αναζήτηση διασποράς και την μέθοδο των ένθετων κατατμήσεων(nested partitions). Οι πιο δημοφιλείς μέθοδοι μεταευρετικής που χρησιμοποιούνται στην εξόρυξη δεδομένων είναι στην πραγματικότητα γενετικοί αλγόριθμοι και οι παραλλαγές του. Ως προσέγγιση της ολικής βελτιστοποίησης, διαπιστώθηκε ότι οι γενετικοί αλγόριθμοι (GA) εφαρμόζονται σε προβλήματα βελτιστοποίησης που είναι δυσεπίλυτα για ακριβείς λύσεις με συμβατικές μεθόδους(Holland, 1975, Goldberg, 1989).

Είναι ένας αλγόριθμος αναζήτησης, όπου σε κάθε επανάληψη δημιουργούνται ταυτόχρονα μια σειρά λύσεις. Σε κάθε βήμα, ένα υποσύνολο του τρέχοντος συνόλου των λύσεων επιλέγεται με βάση την απόδοση τους και αυτές οι λύσεις συνδυάζονται σε νέες λύσεις. Οι χειριστές που χρησιμοποιούνται για τη δημιουργία του νέων λύσεων είναι της επιβίωσης, όπου η λύση μεταφέρεται στην επόμενη επανάληψη χωρίς αλλαγή. Έτσι έχουμε την ανταλλαγή (crossover) όπου οι ιδιότητες των δύο λύσεων συνδυάζονται σε μία και της μετάλλαξης (mutation), όπου η λύση είναι ελαφρώς τροποποιημένη. Στη συνέχεια, επαναλαμβάνεται η ίδια διαδικασία με ένα νέο σύνολο των λύσεων.

Οι χειριστές ανταλλαγής και μετάλλαξης εξαρτώνται από την αντιπροσώπευση της λύσης και όχι από την αξιολόγηση των επιδόσεών του. Η επιλογή των λύσεων, ωστόσο, εξαρτάται από την απόδοση. Η γενική αρχή είναι ότι οι υψηλής απόδοσης λύσεις πρέπει να έχουν περισσότερες πιθανότητες να επιβιώσουν και να τους επιτραπεί να δημιουργούν νέες λύσεις μέσω της ανταλλαγής. Για τους γενετικούς αλγόριθμους και τις άλλες εξελικτικές μεθόδους, το καθοριστικό στοιχείο είναι ο καινοτόμος τρόπος με τον οποίο η ανταλλαγή και η μετάλλαξη καθορίζουν μια γειτονιά της τρέχουσας λύσης. Αυτό επιτρέπει στην αναζήτηση να διατρέχει γρήγορα και έξυπνα μεγάλα τμήματα του χώρου λύσης.

Στην εξόρυξη δεδομένων, έχουν χρησιμοποιηθεί γενετικοί και εξελικτικοί αλγόριθμοι για να λύσουν ένα πλήθος προβλημάτων, συμπεριλαμβανομένης της επιλογής

χαρακτηριστικού (Yang and Honavar, 1998; Kim et al., 2000) και της ταξινόμησης (Fu et al., 2003a,b; Larranaga et al., 1996;Sharpe and Glover, 1999).

Η αναζήτηση διασποράς είναι άλλη μία μέθοδο μεταευρετικής που σχετίζεται με την έννοια της εξελικτικής αναζήτησης. Σε κάθε στάδιο ένας αλγόριθμος αναζήτησης διασποράς εξετάζει ένα σύνολο λύσεων που ονομάζεται σύνολο αναφοράς (reference set). Παρόμοια, με το γενετικό αλγόριθμο, οι λύσεις αυτές συνδυάζονται σε ένα νέο σύνολο. Ωστόσο, σε αντίθεση με το γενετικό, στην αναζήτηση διασποράς οι λύσεις συνδυάζονται χρησιμοποιώντας γραμμικούς συνδυασμούς, οι οποίες κατά συνέπεια καθορίζουν τη γειτονιά.

Η μέθοδος των ένθετων παραρτημάτων (nested partition)που εισήχθη από τους Shi και Olafsson (2000), αποτελεί άλλη μια μέθοδο μεταευρετικής για συνδυαστική βελτιστοποίηση. Η βασική ιδέα αυτής της μεθόδου έγκειται στη συστηματική κατάτμηση της εφικτής περιοχή σε υποπεριοχές, αξιολογώντας τις δυνατότητες κάθε περιοχής και στη συνέχεια εστιάζοντας στην υπολογιστική προσπάθεια για την πιο πολλά υποσχόμενη περιοχή.

Αυτή η διαδικασία πραγματοποιείται κατ' επανάληψη με κάθε ένθετο διαμέρισμα εντός του τελευταίου. Η υπολογιστική αποτελεσματικότητα της μεθόδου nested partition στηρίζεται σε μεγάλο βαθμό στον κατακερματισμό, που αν πραγματοποιηθεί κατά τρόπο ώστε οι καλές λύσεις να βρίσκονται κοντά, τότε μπορεί να καταλήξουν σε λύσεις κοντά στη βέλτιστη πολύ γρήγορα. Σε δεδομένα εξόρυξης, ο αλγόριθμος nested partition έχει χρησιμοποιηθεί για την επιλογή χαρακτηριστικών (Olafsson και Yang, 2004,Yang και Olafsson, 2006), και της συσταδοποίησης (Kim και Olafsson,2004).

## 4.2. Η διαδικασία εξόρυξης δεδομένων

Όπως περιγράφεται στην εισαγωγή, η εξόρυξη δεδομένων συνεπάγεται τη χρήση ενός επαγωγικού αλγόριθμου για να μάθουμε προηγουμένως άγνωστα σχέδια από μια μεγάλη βάση δεδομένων. Αλλά πριν χρησιμοποιηθεί ο αλγόριθμος, πρέπει να υπάρξει μια μεγάλη προ-επεξεργασία δεδομένων. Ορισμένοι συγγραφείς την διακρίνουν από την επαγωγική μάθηση με την αναφορά στην όλη διαδικασία ως ανακάλυψη της γνώσης και χρησιμοποιούν τον όρο εξόρυξη δεδομένων μόνο για το μέρος της επαγωγικής μάθησης της διαδικασίας. Όπως προαναφέρθηκε, όμως, αναφερόμαστε σε όλη διαδικασία ως εξόρυξη δεδομένων. Τα τυπικά βήματα στη διαδικασία περιλαμβάνουν τα ακόλουθα:

- Όσον αφορά τις αναλύσεις δεδομένων, η εξόρυξη δεδομένων ξεκινά από τον προσδιορισμό των επιστημονικών ή επιχειρησιακών στόχων και την διαμόρφωση αυτών ως πρόβλημα εξόρυξης δεδομένων
- Δεδομένου του προβλήματος που πρέπει να αντιμετωπιστεί, οι κατάλληλες πηγές δεδομένων πρέπει να εντοπιστούν και τα δεδομένα πρέπει να ενσωματωθούν και να υποστούν κάποια προ-επεξεργασία ώστε να είναι κατάλληλα για την εξόρυξη δεδομένων.
- Μόλις τα δεδομένα έχουν ετοιμαστεί το επόμενο βήμα είναι να δημιουργηθούν προηγουμένως άγνωστα μοτίβα από τα δεδομένα χρησιμοποιώντας την επαγωγική μάθηση. Οι πιο συχνοί τύποι μοτίβων είναι μοντέλα ταξινόμησης, η φυσική συσταδοποίηση των περιπτώσεων και οι κανόνες συσχέτισης που περιγράφουν σχέσεις μεταξύ χαρακτηριστικών.
- Τα τελικά βήματα είναι για την επικύρωση και για την εφαρμογή των μοτίβων που προέρχονται από την επαγωγική μάθηση.

Στις ενότητες που ακολουθούν θα περιγραφεί αρκετά από τα σημαντικότερα σημεία αυτής της διαδικασίας και θα εστιάσουμε ειδικά **στο πως μπορούν να χρησιμοποιηθούν οι μέθοδοι βελτιστοποίησης για τα διάφορα τμήματα της διαδικασίας.**

### 4.2.1. Προ-επεξεργασία δεδομένων και διερευνητικές εξόρυξης δεδομένων

Σε κάθε εφαρμογή εξόρυξης δεδομένων, η βάση δεδομένων που πρόκειται να εξορυχτεί μπορεί να περιέχει θόρυβο ή άσχετα δεδομένα, ορισμένα δεδομένα μπορεί να λείπουν και σε όλες σχεδόν τις περιπτώσεις, οι βάσεις δεδομένων είναι μεγάλες. Η προ-επεξεργασία δεδομένων απαντά σε κάθε ένα από αυτά τα ζητήματα και περιλαμβάνει προκαταρκτικές εργασίες όπως είναι ο καθαρισμός των δεδομένων, η ενοποίηση των δεδομένων, η μετατροπή δεδομένων και η μείωση των δεδομένων. Επίσης, εφαρμόζεται στα αρχικά στάδια της διαδικασίας διερεύνησης εξόρυξης δεδομένων που περιλαμβάνει την ανακάλυψη μοτίβων στα δεδομένα χρησιμοποιώντας συνοπτικά στατιστικά στοιχεία και την απεικόνιση. Η βελτιστοποίηση και άλλα εργαλεία έχουν σχέση τόσο με την προ-επεξεργασία δεδομένων όσο και με τις διερευνητικές εξόρυξη δεδομένων. Σε αυτήν την ενότητα παρουσιάζουμε την επιλογή χαρακτηριστικού και την οπτικοποίηση δεδομένων.

#### 4.2.1.1. Επιλογή χαρακτηριστικού

Η επιλογή χαρακτηριστικού αποτελεί ένα σημαντικό πρόβλημα στην εξόρυξη δεδομένων. Αυτό περιλαμβάνει μια διαδικασία για τον καθορισμό του ποια χαρακτηριστικά είναι συναφή με την έννοια ότι προβλέπουν ή εξηγούν τα δεδομένα, και αντιστρόφως ποια χαρακτηριστικά είναι περιττά ή παρέχουν ελάχιστες πληροφορίες.

Κάνοντας την επιλογή χαρακτηριστικού πριν την εφαρμογή του αλγόριθμου μάθησης έχουμε πολλά οφέλη. Με την εξάλειψη πολλών χαρακτηριστικών γίνεται ευκολότερο η εκπαίδευση σε άλλες μεθόδους μάθησης, δηλαδή μειώνεται ο υπολογιστικός χρόνος της επαγωγής. Επίσης, το μοντέλο που προκύπτει μπορεί να είναι απλούστερο, το οποίο καθιστά πιο εύκολη την ερμηνεία και, συνεπώς, πιο χρήσιμο στην πράξη.

Επίσης, συχνά τα απλά μοντέλα γενικεύουν καλύτερα όταν εφαρμόζονται για την πρόβλεψη. Έτσι, ένα μοντέλο που χρησιμοποιεί λιγότερα χαρακτηριστικά είναι πιθανό να σημειώσει υψηλότερη βαθμολογία και μπορεί να έχει ακόμη καλύτερη επίδοση στην ακρίβεια. Τέλος, ποια χαρακτηριστικά θα πρέπει να διατηρηθούν, δηλαδή ο προσδιορισμός των χαρακτηριστικών που έχουν σχέση με τη λήψη αποφάσεων, συχνά παρέχουν πολύτιμες διαρθρωτικές πληροφορίες και επομένως είναι σημαντικές από μόνες τους. Η βιβλιογραφία σχετικά με την επιλογή χαρακτηριστικού είναι εκτεταμένη και οι πολλές μέθοδοι επιλογής χαρακτηριστικού βασίζονται στην εφαρμογή μιας προσέγγισης βελτιστοποίησης. Σύμφωνα με ένα πρόσφατο παράδειγμα, (Olafsson και Yang2004) το πρόβλημα επιλογής χαρακτηριστικού διατυπώνεται ως ένα απλό συνδυαστικό πρόβλημα βελτιστοποίησης με τις ακόλουθες μεταβλητές:

$$x_i = \{\frac{1 \text{ αν περιλαμβάνεται το χαρακτηριστικό i}}{0 \text{ αλλιώς}} \quad (7)$$

για i=1,2,3,4…m. Τότε το πρόβλημα βελτιστοποίησης είναι

$$\min f(x)$$

$$\text{s.t. } K_{min} \leq \sum_{i=1}^{m} xi \leq K_{max} \quad (8)$$

$$x_i \in_{\{0,1\}}$$

$$\text{όπου } x=(x_1,x_2,x_3,\ldots x_m) \text{ και } K_{min} \text{ και } K_{max}$$

όπου $x = (x_1, x_2,..., x_m)$ και $K_{min}$ και $K_{max}$ είναι ένας ορισμένος ελάχιστος και μέγιστος αριθμό χαρακτηριστικών που θα επιλεγούν. Ένα βασικό ζήτημα είναι η επιλογή της αντικειμενικής συνάρτησης και δεν υπάρχει μία μέθοδος για την αξιολόγηση της ποιότητας των χαρακτηριστικών που λειτουργούν καλύτερα για όλα τα προβλήματα εξόρυξης δεδομένων. Μερικές μέθοδοι αξιολογούν την ποιότητα κάθε χαρακτηριστικού ξεχωριστά, δηλαδή,

$$f(x) = \sum_{i=1}^{m} fixi \quad (9)$$

ενώ άλλοι αξιολογούν την ποιότητα ολόκληρου του υποσύνολου, δηλαδή, f (x) = f (X), όπου X={i: xi=1} είναι το υποσύνολο των επιλεγμένων χαρακτηριστικών. Οι Olafsson και Yang(2004) χρησιμοποιούν τη μέθοδο των nested partitions των Shi και Olafsson(2000) για την επίλυση αυτού του προβλήματος με τη χρήση πολλαπλών αντικειμενικών συναρτήσεων και των δύο τύπων που περιγράφονται παραπάνω και δείχνουν ότι μια τέτοια προσέγγιση βελτιστοποίησης είναι πολύ αποτελεσματική. Οι Olafsson καιYang(2006) βελτιώνουν τα αποτελέσματα αναπτύσσοντας μιας προσαρμοστική έκδοση του αλγόριθμου, ο οποίος σε κάθε βήμα χρησιμοποιεί ένα μικρό τυχαίου υποσύνολο του συνόλου των περιπτώσεων. Αυτό είναι σημαντικό διότι η εξόρυξη δεδομένων ασχολείται συνήθως με πολύ μεγάλο αριθμό περιπτώσεων και η επεκτασιμότητα σε σχέση με τον αριθμό των περιπτώσεων αποτελεί ένα κρίσιμο ζήτημα. Άλλες μέθοδοι που βασίζονται στη βελτιστοποίηση έχουν χρησιμοποιηθεί για τέτοια προβλήματα περιλαμβάνουν γενετικούς αλγόριθμους(Yang και Honavar, 1998), την εξελικτική αναζήτηση(Kim et al., 2000), την προσομοιωμένη ανόπτηση (Debuse και Rayward-Smith, 1997), την λογική ανάλυση των δεδομένων (Boros et al., 2000), και μαθηματικός προγραμματισμός(Bradley et al., 1998).

### 4.2.1.2. Οπτικοποίηση δεδομένων

Όπως γίνεται με τις περισσότερες άλλες αναλύσεις των δεδομένων, η οπτικοποίηση των δεδομένων παίζει σημαντικό ρόλο στην εξόρυξη των δεδομένων. Αυτό αποτελεί ένα δύσκολο πρόβλημα δεδομένου ότι τα στοιχεία είναι συνήθως πολλά, δηλαδή, ο αριθμός m των χαρακτηριστικών είναι μεγάλος, ενώ τα δεδομένα μπορούν να οπτικοποιηθούν σε δύο ή τρεις διαστάσεις. Ενώ είναι δυνατό να οπτικοποιηθούν δύο ή τρία χαρακτηριστικά κάθε φορά, μια καλύτερη εναλλακτική λύση είναι να χαρτογραφηθούν τα δεδομένα σε δύο ή τρεις διαστάσεις με τέτοιο τρόπο που να διατηρούνται η δομή των σχέσεων (δηλαδή, οι αποστάσεις) μεταξύ των περιπτώσεων. Το πρόβλημα αυτό έχει παραδοσιακά διατυπωθεί ως ένα μη γραμμικό μαθηματικού προγραμματισμού πρόβλημα.

Ως εναλλακτική λύση στην παραδοσιακή διατύπωση, ο Abbiw-Jackson et al. (2006) παρείχε πρόσφατα την ακόλουθη διατύπωση δευτεροβάθμιας εξίσωσης. Δεδομένου των περιπτώσεων n σε $R^m$ και μια μήτρα $D^{old} \in R^n x R^n$, η μέτρηση της απόστασης μεταξύ εκείνων των περιπτώσεων, βρίσκεις τη βέλτιστη κατανομή εκείνων των περιπτώσεων σε ένα πλέγμα N σε $R^q$, q = 2, 3. Οι μεταβλητές απόφασης δίνονται από:

$$x_{ik} = \begin{cases} 1 & \text{αν περιλαμβάνεται στο σημείο πλέγματος } k \in N, \text{η περίπτωση } i \\ 0 & \text{αλλιώς} \end{cases} \quad (10)$$

Με αυτή την κατανομή, υπάρχει μια νέα απόσταση μήτρας $D^{new} \in R^q x R^q$ με το q = 2 ή 3 και το μαθηματικό πρόγραμμα μπορεί να γραφτεί ως εξής:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \in N}^{n} \sum_{l \in N}^{n} F(Dold, Dnew) XikXjl$$

$$\text{υπόκειται } \sum_{\kappa \in N} X\iota\kappa = 1, \forall \iota$$

$$x_{ik} \in \{0, 1\} \quad (11)$$

όπου F είναι συνάρτηση της απόκλισης μεταξύ των διαφορών μεταξύ των περιπτώσεων στον αρχικό χώρο και το νέο q-διάστατο χώρο. Οποιαδήποτε μέθοδος μπορεί να χρησιμοποιηθεί για την λύση, αλλά ο Abbiw-Jackson et al. (2006) προτείνουν μια ευρετική τοπικής αναζήτησης που λαμβάνει υπόψη την αντικειμενική συνάρτηση και συγκρίνει τα αποτελέσματα, με τις παραδοσιακές διατυπώσεις μη γραμμικού μαθηματικού προγραμματισμού. Καταλήγουν στο συμπέρασμα ότι παρέχει παρόμοια αποτελέσματα και τείνει να αποδίδει καλύτερα σε μεγάλα προβλήματα.

## 4.2.2. Ταξινόμηση

Μόλις τα δεδομένα υποστούν προ-επεξεργασία εφαρμόζεται ένας αλγόριθμος και ένα από τα πιο κοινά καθήκοντα στην εξόρυξη δεδομένων είναι η ταξινόμηση. Εδώ υπάρχει ένα συγκεκριμένο χαρακτηριστικό που ονομάζεται χαρακτηριστικό τάξης και μπορεί να πάρει ένα συγκεκριμένο αριθμό τιμών και ο στόχος είναι να προκαλέσει ένα μοντέλο που μπορεί να χρησιμοποιηθεί για να ξεχωρίσει τα νέα δεδομένα σε κατηγορίες σύμφωνα με αυτές τις τιμές. Η επαγωγή βασίζεται σε ένα σύνολο κατάρτισης, όπου κάθε περίπτωση επισημαίνεται σύμφωνα με την αξία των χαρακτηριστικών της τάξης. Ο στόχος της ταξινόμησης είναι να αναλύσει πρώτα τα δεδομένα της κατάρτισης και να αναπτύξει μια

ακριβή περιγραφή ή ένα πρότυπο για κάθε τάξη χρησιμοποιώντας τα διαθέσιμα χαρακτηριστικά στα δεδομένα.

Οι περιγραφές της κατηγορίας χρησιμοποιούνται στη συνέχεια για την ταξινόμηση μελλοντικών ανεξάρτητων δεδομένων ή για την ανάπτυξη μιας καλύτερης περιγραφής για την κάθε τάξη. Πολλές μέθοδοι έχουν μελετηθεί για την ταξινόμηση, που περιλαμβάνουν το decision tree induction, τις μηχανές διανυσμάτων υποστήριξης, τα νευρωνικά δίκτυα και δίκτυα Bayesian (Fayyad et al, 1996, Weiss και Kulikowski, 1991). Η **βελτιστοποίηση** έχει σχέση με πολλές μεθόδους ταξινόμησης και τις μηχανές διανυσμάτων υποστήριξης. Στην ενότητα αυτή έχουμε επικεντρωθεί σε τρεις επιπλέον δημοφιλείς προσεγγίσεις ταξινόμησης, δηλαδή στο decision tree induction, τα δίκτυα Bayesian και τα νευρωνικά δίκτυα.

### 4.2.2.1. Δέντρα αποφάσεων (Decision trees)

Μία από τις πιο δημοφιλείς τεχνικές για την ταξινόμηση είναι η μετατόπιση από τα πάνω προς τα κάτω των δέντρων αποφάσεων. Ένας από τους βασικούς λόγους της δημοτικότητάς τους φαίνεται να είναι η διαφάνεια τους, και ως εκ τούτου υπάρχει κάποιο σχετικό πλεονέκτημα από την άποψη της ερμηνείας του. Ένα άλλο πλεονέκτημα είναι η άμεση διαθεσιμότητα των ισχυρών εφαρμογών όπως είναι η CART ( Breiman et al., 1984) και η C4.5 (Quinlan, 1993).Οι περισσότεροι αλγόριθμοι decision tree induction δημιουργούν ένα δέντρο από πάνω προς τα κάτω επιλέγοντας τα χαρακτηριστικά, ένα προς ένα κάθε φορά και διαιρώντας τα δεδομένα σύμφωνα με τις τιμές αυτών των χαρακτηριστικών. Το πιο σημαντικό χαρακτηριστικό έχει επιλεγεί ως ο κορυφαίος κόμβος που διαχωρίζεται, και ούτω καθεξής.

Για παράδειγμα, στην C4.5 τα χαρακτηριστικά επιλέγονται για να μεγιστοποιήσουν το κέρδος που προκύπτει από την πληροφορία σε κάποιο κόμβο (Quinlan, 1993).Αποτελεί ένα μέτρο εντροπίας που αποσκοπεί στην αύξηση της μέσης καθαρότητας της κατηγορίας των υποσυνόλων που προκύπτουν. Αλγόριθμοι όπως ο C4.5 και ο CART είναι υπολογιστικά αποτελεσματικοί και έχουν αποδειχθεί πολύ επιτυχής στην πράξη. Ωστόσο, το γεγονός ότι περιορίζονται στην κατασκευή διαχωριστικών επιπέδων παράλληλα στον άξονα περιορίζει την αποτελεσματικότητάς τους σε εφαρμογές όπου κάποιος συνδυασμός των χαρακτηριστικών γνωρισμάτων της κατηγορίας είναι εξαιρετικά προβλέψιμος(Lee και Olafsson, 2006).

Οι τεχνικές μαθηματικής βελτιστοποίησης έχουν εφαρμοστεί άμεσα στη βέλτιστη κατασκευή ορίων αποφάσεων σε decision tree induction. Ειδικότερα, ο Bennett(1992) εισήγαγε μια επέκταση των τεχνικών γραμμικού προγραμματισμού στην κατασκευή των δέντρων αποφάσεων, αν και η διατύπωση αυτή περιορίζεται σε προβλήματα δύο κατηγοριών. Σε πρόσφατη εργασία, ο Street(2005) παρουσίασε ένα νέο αλγόριθμο για decision tree induction που βασίζεται σε μη γραμμικό προγραμματισμό. Ο αλγόριθμος, ονομάζεται Oblique Category SEParation (OC-SEP)και δείχνει βελτιωμένη γενίκευση σε αρκετά σύνολα δεδομένων του πραγματικού κόσμου.

Ένας από τους περιορισμούς των περισσότερων αλγόριθμων των δέντρων αποφάσεων είναι πως γνωρίζουμε ότι είναι ασταθείς. Αυτό ισχύει ιδιαίτερα όταν ασχολούμαστε με ένα μεγάλο σύνολο όπου μπορεί να μην είναι πρακτικό να έχουμε πρόσβαση σε όλα τα δεδομένα ταυτόχρονα και κατασκευάσουμε ένα δέντρο απόφασης (Fu et

al., 2003a). Για την αύξηση της ερμηνείας, είναι αναγκαίο να μειωθούν τα μεγέθη των δέντρων και αυτό μπορεί να καταστήσει τις διαδικασίες λιγότερο σταθερές.

Η εξεύρεση του βέλτιστου δέντρου απόφασης μπορεί να αντιμετωπιστεί ως συνδυαστικό πρόβλημα βελτιστοποίησης, τότε ένα NP-πλήρες πρόβλημα και η ευρετική πρέπει να εφαρμοστούν. Ο Kennedy et al. (1997) πρώτοι ανέπτυξαν ένα γενετικό αλγόριθμο για τη βελτιστοποίηση των δέντρων. Στην προσέγγισή τους, ένα δυαδικό δέντρο εκπροσωπείται από έναν αριθμό υποδέντρων (subtrees), κάθε ένα έχει ένα κόμβο και δύο διακλαδώσεις. Σε πιο πρόσφατες εργασίες, ο Fu et al. (2003α, β, 2006) χρησιμοποίησαν επίσης γενετικούς αλγόριθμους. Η μέθοδός τους χρησιμοποιεί τον αλγόριθμο C4.5 για να δημιουργήσει K δένδρα ως τον αρχικό πληθυσμό, και στη συνέχεια ανταλλάσει τα υποδέντρα (subtrees) μεταξύ των δένδρων (crossover) ή εντός του ίδιου δέντρου (μετάλλαξη).

Στο τέλος μιας γενιάς, πραγματοποιούνται λογικοί έλεγχοι και κλάδεμα για βελτιώσουν το δέντρο αποφάσεων. Δείχνουν ότι το δέντρο που προκύπτει αποδίδει καλύτερα από τον αλγόριθμο C4.5 και ο χρόνος υπολογισμού αυξάνει μόνο γραμμικά καθώς αυξάνει το μέγεθος του συνδυασμού κατάρτισης και βαθμολόγησης. Επιπλέον, δημιουργώντας κάθε δέντρο απαιτείται μόνο ένα μικρό ποσοστό των δεδομένων για την παραγωγή δέντρων αποφάσεων υψηλής ποιότητας. Όλες οι παραπάνω προσεγγίσεις χρησιμοποιούν κάποια συνάρτηση από την ακρίβεια του δέντρου για την κατάλληλη λειτουργία του γενετικού αλγόριθμου. Ειδικότερα, ο Fu et al. (2003α) κάνουν άμεση χρήση της μέσης ακρίβειας της ταξινόμησης (average classification accuracy), ενώ ο Fu et al. (2003b) χρησιμοποιούν μια κατανομή για την ακρίβεια που τους επιτρέπει να δικαιολογούν την ανοχή κινδύνων για το χρήστη. Αυτό επεκτείνεται περαιτέρω από το Fu et al. (2006) όπου η ταξινόμηση μοντελοποιείται χρησιμοποιώντας μια συνάρτηση απώλειας που γίνεται η συνάρτηση καταλληλότητας του γενετικού αλγορίθμου.

Τέλος, σε άλλες συναφείς εργασίες ο Dhar et al. (2000) χρησιμοποιούν μία προσαρμοσμένη μέθοδο resampling , όπου αντί να χρησιμοποιείται ένα πλήρες δέντρο αποφάσεων , όπως η χρωμοσωμική μονάδα, ένα χρωμόσωμα είναι απλά ο κανόνας, δηλαδή, κάθε πλήρη διαδρομή από τη ρίζα κόμβο του δένδρου σε έναν κόμβο-φύλλο. Κατά τη χρήση γενετικού αλγορίθμου για τη βελτιστοποίηση των τριών, συνήθως δεν υπάρχει μέθοδος για τον κατάλληλο έλεγχο της αύξησης του δέντρου, διότι ο γενετικός αλγόριθμος δεν αξιολογεί το μέγεθος του δέντρου. Έτσι λοιπόν, κατά τη διαδικασία αναζήτησης, το δέντρο μπορεί να γίνει υπερβολικά βαθύ και σύνθετο ή μπορεί να διευθετηθεί με ένα πολύ απλό δέντρο. Για να αντιμετωπιστεί αυτό, ο Niimi και ο Tazaki (2000) συνδύασαν το γενετικό προγραμματισμό με κανόνες συσχέτισης των αλγόριθμων για την κατασκευή δέντρου απόφασης. Σε αυτήν την προσέγγιση, οι κανόνες που δημιουργήθηκαν από τον Apriori αλγόριθμο αναζήτησης των κανόνων συσχέτισης (Apriori association rule discovery algorithm) (Agrawal et al., 1993), λαμβάνονται ως τα αρχικά δέντρα αποφάσεων για ένα μεταγενέστερο αλγόριθμο γενετικού προγραμματισμού.

Μια άλλη προσέγγιση για τη βελτίωση της βελτιστοποίησης του δέντρου απόφασης είναι να βελτιωθεί η συνάρτηση καταλληλότητας που χρησιμοποιείται από το γενετικό αλγόριθμο.

Οι παραδοσιακές συναρτήσεις καταλληλότητας χρησιμοποιούν τη μέση ακρίβεια για να μετρήσουν την απόδοση. Ο Fu et al. (2003b) έχουν διερευνήσει τη χρήση των διαφόρων εκατοστημορίων της κατανομής της ακρίβειας της ταξινόμησης στη θέση του μέσου όρου και ανέπτυξαν ένα γενετικό αλγόριθμο που ταυτόχρονα ελέγχει δύο κριτήρια καταλληλότητας. Οι Tanigawa και Zhao (2000) περιλαμβάνουν το μέγεθος των δένδρων στις συνάρτηση καταλληλότητας ώστε να ελέγξουν την αύξηση των δέντρων. Επίσης, η χρησιμοποίηση της μιας συνάρτησης καταλληλότητας με βάση το J-Measure, το οποίο

καθορίζει το περιεχόμενο των πληροφοριών ενός δέντρου, μπορεί να δώσει ένα κριτήριο προτίμησης για να βρείτε το δέντρο αποφάσεων που ταξινομεί μια σειρά περιπτώσεων με τον καλύτερο τρόπο (Folino et al., 2001).

### 4.2.2.2. Δίκτυα Bayesian

Η δημοφιλής μέθοδος Bayes είναι μια άλλη μια απλή αλλά αποτελεσματική μέθοδος ταξινόμησης. Η μέθοδος αυτή μαθαίνει την πιθανότητα του κάθε χαρακτηριστικού λαμβάνοντας υπόψη την ετικέτα της κατηγορίας από τα δεδομένα της εκπαίδευσης. Η ταξινόμηση τότε γίνεται με την εφαρμογή του κανόνα Bayes για να υπολογιστεί η πιθανότητας μιας τιμής της κατηγορίας λόγω της ιδιαίτερης περίπτωσης και της πρόβλεψης της τιμής της κατηγορίας με τη υψηλότερη πιθανότητα. Σε γενικές γραμμές αυτό θα απαιτούσε τον υπολογισμό των οριακών πιθανοτήτων του κάθε συνδυασμού των χαρακτηριστικών, η οποία δεν είναι εφικτή, ιδιαίτερα όταν ο αριθμός των χαρακτηριστικών είναι μεγάλος και μπορεί να υπάρχουν λίγες ή καθόλου παρατηρήσεις (περιπτώσεις) για ορισμένους συνδυασμούς των χαρακτηριστικών.

Ως εκ τούτου, γίνεται μια ισχυρή ανεξάρτητη παραδοχή όπου, δηλαδή, όλα τα χαρακτηριστικά γίνονται αποδεκτά δεδομένης της αξίας του χαρακτηριστικού της κατηγορίας. Δεδομένου της υπόθεσης αυτής, μόνο οι οριακές πιθανότητες του κάθε χαρακτηριστικού, λαμβάνοντας υπόψη την κατηγορία που θα πρέπει να υπολογιστούν. Ωστόσο, η υπόθεση αυτή είναι σαφώς μη ρεαλιστική και τα δίκτυα Bayesian ρητά μοντελοποιούν τις εξαρτήσεις μεταξύ των χαρακτηριστικών.

Ένα δίκτυο Bayesian αποτελεί είναι ένα κατευθυνόμενο άκυκλο γράφημα G που μοντελοποιεί σχέσεις πιθανοτήτων μεταξύ ενός συνόλου τυχαίων μεταβλητών U = {X1,. . . , Xm}, όπου κάθε μεταβλητή σε U έχει συγκεκριμένες τιμές ή αξίες (Jensen,1996).

Το m δηλώνει τον αριθμό των χαρακτηριστικών. Κάθε κόμβος στο γράφημα αντιπροσωπεύει μια τυχαία μεταβλητή, ενώ οι άκρες συλλαμβάνουν τις άμεσες εξαρτήσεις μεταξύ των μεταβλητών. Το δίκτυο κωδικοποιεί τις υποτιθέμενες σχέσεις ανεξαρτησίας ότι κάθε κόμβος είναι ανεξάρτητος από τους μη απογόνους αν ληφθούν υπόψη οι γονείς του (Castillo et al, 1997,Pernkopf, 2005). Υπάρχουν δύο βασικά θέματα που σχετίζονται με την βελτιστοποίηση κατά τη χρήση των δικτύων Bayesian. Πρώτον, όταν ορισμένοι από τους κόμβους του δικτύου δεν είναι παρατηρήσιμοι, δηλαδή, δεν υπάρχουν στοιχεία για τις τιμές των χαρακτηριστικών που ανταποκρίνονται σε αυτούς τους κόμβους, τότε για να βρεθούν οι πιθανότερες τιμές των πιθανοτήτων μπορεί να διατυπωθεί ως ένα μη γραμμικό μαθηματικό πρόγραμμα.

Στην πράξη, αυτό είναι συνήθως αντιμετωπίζεται με μια απλή προσέγγιση. Το δεύτερο πρόβλημα βελτιστοποίησης εμφανίζεται όταν η δομή του δικτύου είναι άγνωστη και μπορεί να διατυπωθεί ως ένα συνδυαστικό πρόβλημα βελτιστοποίησης. Το πρόβλημα της εκμάθησης της δομής ενός δικτύου Bayesian μπορεί να δηλώνεται ανεπίσημα ως εξής. Αν λάβουμε υπόψη ένα σύνολο κατάρτισης A = {u1, u2,. . . , un} του n περιπτώσεων του U μπορούμε να βρούμε ένα δίκτυο που ταιριάζει καλύτερα στο A.Η κοινή προσέγγιση στο πρόβλημα αυτό είναι η εισαγωγή μιας αντικειμενικής συνάρτησης που αξιολογεί κάθε δίκτυο σε σχέση με τα δεδομένα εκπαίδευσης και στη συνέχεια αναζητά το καλύτερο δίκτυο, σύμφωνα με αυτή τη συνάρτηση (Friedman et al., 1997). Το κλειδί για τις προκλήσεις της βελτιστοποίησης είναι η επιλογή μιας αντικειμενική συνάρτησης και του καθορισμού του

τρόπου αναζήτησης για το καλύτερο δίκτυο. Οι δύο κύριες αντικειμενικές συναρτήσεις που χρησιμοποιούνται συνήθως να μάθει κάποιος τα δίκτυα Bayesian είναι τα Bayesian scoring function(Cooper και Herskovits, 1992,Heckermanetal., 1995), καθώς και μια συνάρτηση που βασίζεται στην αρχή της ελάχιστης περιγραφής (minimal description length MDL) ((LamandBacchus,1994; Suzuki, 1993).Οποιαδήποτε μεταευρετική μπορεί να εφαρμοστεί για την επίλυση του προβλήματος. Για παράδειγμα, ο Larranaga et al. (1996)έχουν δουλέψει με τους γενετικούς αλγόριθμοι για την εκμάθηση των δικτύων Bayesian.

### 4.2.2.3. Νευρωνικά δίκτυα

Μια άλλη δημοφιλής προσέγγιση για την ταξινόμηση είναι τα νευρωνικά δίκτυα. Τα νευρωνικά δίκτυα έχουν μελετηθεί εκτενώς στη βιβλιογραφία και μια εξαιρετική επανεξέταση της χρήσης των νευρωνικών δικτύων για την ταξινόμηση έχει δοθεί από Zhang(2000). Η επαγωγική μάθηση των νευρωνικών δικτύων από τα δεδομένα αναφέρεται ως κατάρτιση αυτού του δικτύου (training this network), και η πιο δημοφιλής μέθοδος κατάρτισης είναι η διάδοση του (Rumelhart και McClelland, 1986). Είναι γνωστό ότι η διάδοση μπορεί να θεωρηθεί ως διαδικασία βελτιστοποίησης και αφού έχει μελετηθεί λεπτομερώς αλλού μπορούμε εν συντομία να εξετάσουμε την πρωταρχική του σχέση με τη βελτιστοποίηση.

Ένα νευρωνικό δίκτυο αποτελείται από τουλάχιστον τρία στρώματα κόμβων. Το στρώμα εισόδου αποτελείται από ένα κόμβο για ανεξάρτητο χαρακτηριστικό. Το στρώμα εξόδου αποτελείται από κόμβο ή κόμβους για το χαρακτηριστικό ή τα χαρακτηριστικά της κατηγορίας και συνδέοντας αυτά τα στρώματα αποτελεί ένα ή περισσότερα ενδιάμεσα στρώματα των κόμβων που μετατρέπουν την είσοδο σε μια έξοδο.

Όταν συνδεθούν, αυτά τα στρώματα των κόμβων συνθέτουν το δίκτυο που ονομάζουμε νευρωνικό δίκτυο. Η εκπαίδευση στο νευρωνικό δίκτυο συνίσταται στον προσδιορισμό των παραμέτρων για αυτό το δίκτυο. Συγκεκριμένα, κάθε τόξο που συνδέει τους κόμβους στο δίκτυο αυτό έχει ειδικό βάρος και οι τιμές των βαρών καθορίζουν τον τρόπο με τον οποίοη είσοδος μετατρέπεται σε έξοδο. Οι περισσότεροι μέθοδοι κατάρτισης στα νευρωνικά δίκτυα, συμπεριλαμβανομένου και της διάδοσης, είναι εκ φύσεως μια διαδικασία βελτιστοποίησης.

Τα δεδομένα εκπαίδευσης αποτελούνται από τιμές για ορισμένα χαρακτηριστικά των εισροών (input layer) μαζί με το χαρακτηριστικό της κατηγορίας (output layer), το οποίο συνήθως αναφέρεται ως target value του δικτύου. Η διαδικασία βελτιστοποίησης επιδιώκει να καθορίσει τα βάρη του τόξου, προκειμένου να μειώσει το σφάλμα μεταξύ της πραγματικής εκροής και του στόχος εκροής (Ripley,1996) Αφού τα βάρη του δικτύου είναι συνεχείς μεταβλητές και η σχέση μεταξύ των εισροών και των εκροών είναι μη γραμμική, αυτό αποτελεί ένα μη γραμμικό μόνιμο πρόβλημα βελτιστοποίησης ή ένα μη γραμμικό πρόβλημα προγραμματισμού (non-linear programming problem NLP). Οποιοσδήποτε κατάλληλος NLP αλγόριθμος θα μπορούσε συνεπώς να εφαρμόζεται για την εκπαίδευση σε ένα νευρωνικό δίκτυο, αλλά στην πράξη μια απλή προσέγγιση εφαρμόζεται πιο συχνά (Ripley, 1996). Αυτό δεν εξασφαλίζει ότι η συνολική βέλτιστη λύση έχει βρεθεί, αλλά μάλλον τελειώνει στο πρώτο τοπικό βέλτιστο που αντιμετώπισε. Ωστόσο, λόγω του μεγέθους των προβλημάτων, η ταχύτητα του αλγορίθμου βελτιστοποίησης είναι συνήθως επιτακτική. Αυτή η ενότητα περιγράφει το πως η βελτιστοποίηση παίζει ένα σημαντικό ρόλο στην ταξινόμηση. Το πρόβλημα ταξινόμησης μπορεί να διατυπωθεί ως ένα πρόβλημα μαθηματικού

προγραμματισμού, και διαδραματίζει επίσης σημαντικό ρόλο, σε σχέση με άλλες μεθόδους ταξινόμησης.

Αυτό μπορεί να είναι τόσο για τη βελτιστοποίηση της εκροής άλλων αλγορίθμων ταξινόμησης, όπως η βελτιστοποίηση των δέντρων απόφασης, καθώς και για τη βελτιστοποίηση των παραμέτρων που χρησιμοποιούνται από άλλους αλγορίθμους ταξινόμησης, όπως είναι η εξεύρεση της βέλτιστης δομής ενός δικτύου Bayesian. Παρά τη σημαντική έρευνα που έχει γίνει στον τομέα αυτό εξακολουθούν να υπάρχουν πολλά άλυτα προβλήματα που πρέπει να αντιμετωπιστούν.

### 4.2.3. Συσταδοποίηση

Όταν τα δεδομένα είναι χωρίς τίτλο και κάθε περίπτωση δεν έχει μια δεδομένη ετικέτα κατηγορίας, η μάθηση είναι χωρίς επίβλεψη (unsupervised). Αν εξακολουθούμε να θέλουμε να προσδιορίσουμε ποιες περιπτώσεις ανήκουν μαζί, δηλαδή, σχηματίζουν φυσικές συστάδες περιπτώσεων, μπορεί να εφαρμοστεί ένας αλγόριθμος συσταδοποίησης (Jain et al, 199, Kaufman and Rousseeuw,1990) Τέτοιοι αλγόριθμοι μπορούν να χωριστούν σε δύο κατηγορίες: ιεραρχικής συσταδοποίησης και διαμερισματοποιημένη συσταδοποίηση (partitional clustering). Στην ιεραρχική συσταδοποίηση όλες οι περιπτώσεις οργανώνονται σε μια ιεραρχία που περιγράφει το βαθμό της ομοιότητας μεταξύ αυτών των περιπτώσεων. Μια τέτοια αντιπροσώπευση μπορεί να παρέχει πολλές πληροφορίες και έχουν προταθεί πολλοί αλγόριθμοι.

Η διαμερισματοποιημένη συσταδοποίηση, από την άλλη πλευρά, απλά δημιουργεί ένα διαμέρισμα των δεδομένων, όπου κάθε περίπτωση εμπίπτει σε μια συστάδα. Έτσι λοιπόν, παράγονται λιγότερες πληροφορίες, αλλά βελτιώνεται η ικανότητα να ασχοληθεί με μεγάλο αριθμό περιπτώσεων. Όπως φαίνεται να έχει επισημάνει, πρώτος ο Vinod (1969), το πρόβλημα της διαμερισματοποιημένης συσταδοποίησης μπορεί να διατυπωθεί ως ένα πρόβλημα βελτιστοποίησης. Τα βασικά θέματα είναι: πώς να καθοριστούν οι μεταβλητές απόφασης και πώς να καθοριστούν οι αντικειμενικές συναρτήσεις,

καμία από τις οποίες δεν μπορεί να έχει γενική εφαρμογή. Στη συσταδοποίηση, οι πιο κοινοί στόχοι είναι η ελαχιστοποίηση της διαφοράς των περιπτώσεων σε κάθε συστάδα (συμπαγή), η μεγιστοποίηση της διαφοράς μεταξύ περιπτώσεων σε διαφορετικές συστάδες (διαχωρισμός) ή κάποιος συνδυασμός των δύο αυτών μέτρων. Ωστόσο, άλλα μέτρα μπορούν επίσης να παρουσιάζουν ενδιαφέρον και να εκτιμούν σωστά την ποιότητα μιας συστάδας (Estevill-Castro, 2002; Grabmeier and Rudolph,2002; Osei-Bryson, 2005). Μια λεπτομερής συζήτηση αυτού του θέματος είναι εκτός του πεδίου εφαρμογής της εργασίας και στο μεγαλύτερο μέρος της εργασίας όπου γίνεται εφαρμογή της βελτιστοποίησης στην συσταδοποίηση επικεντρώνεται η προσοχή σε κάποια παραλλαγή του συμπαγούς. Εκτός από το ζήτημα της επιλογής του κατάλληλου τρόπου μέτρησης της ποιότητας της συσταδοποίησης ως αντικειμενική συνάρτηση δεν υπάρχει γενική συμφωνία για τον τρόπο με τον οποίο θα καθορίζεται η συσταδοποίηση. Ένα δημοφιλής τρόπος για να ορίσουμε την συσταδοποίηση των δεδομένων είναι να αφήσουμε κάθε συστάδα να καθορίζεται από το κεντρικό σημείο του $C_J \in R^m$ και στη συνέχεια να εκχωρήσει όλες τις περιπτώσεις προς το πλησιέστερο κέντρο. Έτσι, η συσταδοποίηση ορίζεται από μία m x k μήτρα C = (C1, C2,..., Ck).

Αυτό για παράδειγμα γίνεται με το κλασικό και ακόμα δημοφιλές αλγόριθμο k-means (MacQueen, 1967). Ο αλγόριθμος κ-means είναι ένας απλός επαναληπτικός αλγόριθμος που προχωρεί ως εξής: Ξεκινώντας με τυχαία επιλεγμένες περιπτώσεις ως κέντρα, κάθε

περίπτωση είναι η πρώτη η οποία αποδίδεται στο πλησιέστερο κέντρο. Δεδομένων αυτών των αναθέσεων, τα κέντρα των συστάδων υπολογίζονται εκ νέου και κάθε περίπτωση και πάλι τοποθετούνται στο πλησιέστερο κέντρο. Αυτό επαναλαμβάνεται μέχρις ότου καμία περίπτωση να μην αλλάζει συστάδα μετά τον επαναϋπολογισμό των κέντρων ,δηλαδή, ο αλγόριθμος συγκλίνει σε ένα τοπικό βέλτιστο. Μεγάλο μέρος των εργασιών για την διαμόρφωση της βελτιστοποίησης χρησιμοποιεί την ιδέα του προσδιορισμού μιας συσταδοποίησης χρησιμοποιώντας ένα σταθερό αριθμό κέντρων. Αυτό ισχύει για τα πρώτα έργα του Vinod (1969), όπου ο συγγραφέας δίνει δύο ακέραιες διατυπώσεις του προβλήματος συσταδοποίησης. Για παράδειγμα, στην πρώτη διατύπωση, η μεταβλητή απόφασης ορίζεται ως ένας δείκτης της συστάδας στο οποίο αποδίδεται κάθε περίπτωση:

$$x_{ij} = \begin{cases} 1 & \text{αν η περίπτωση } i \text{ αποδίδεται στην συστάδα } j \\ 0 & \text{αλλιώς} \end{cases} \quad (12)$$

και ο στόχος είναι να ελαχιστοποιηθεί το συνολικό κόστος της εκχώρησης, όπου $w_{ij}$ είναι κάποιο κόστος της εκχώρησης της i περίπτωσης στο σύμπλεγμα j.

## 4.3. Εφαρμογές στη διαχείριση των ηλεκτρονικών υπηρεσιών

Πολλές από τις πιο σημαντικές εφαρμογές της εξόρυξης των δεδομένων βρίσκονται σε τομείς που σχετίζονται με τη διαχείριση των ηλεκτρονικών υπηρεσιών. Σε αυτές τις εφαρμογές, η αυτόματη συλλογή και αποθήκευση δεδομένων είναι συχνά φθηνή και απλή, παράγει μεγάλες βάσεις δεδομένων στις οποίες μπορεί να εφαρμοστεί η εξόρυξη δεδομένων. Σε αυτή την ενότητα εξετάζουμε δύο τέτοιες περιοχές εφαρμογών, πιο συγκεκριμένα την διαχείριση πελατειακών σχέσεων και την εξατομίκευση. Επιλέγουμε τις εφαρμογές αυτές λόγω της σημασίας τους και τη χρησιμότητας της εξόρυξης δεδομένων για την επίλυσή τους, καθώς και για τις σχετικά ανεξερεύνητες δυνατότητες που υπάρχουν για την ενσωμάτωση της τεχνολογίας βελτιστοποίησης που επιτρέπει και εξηγεί τα αποτελέσματα της εξόρυξης δεδομένων.

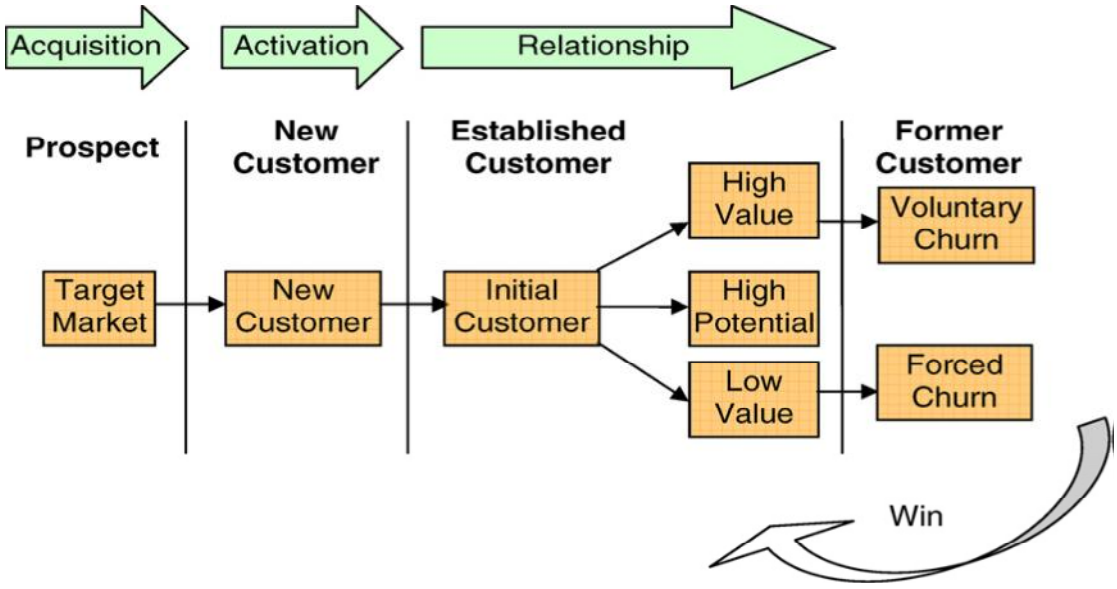
### 4.3.1. Διαχείριση πελατειακών σχέσεων

Οι σχέσεις μάρκετινγκ και διαχείρισης πελατειακών σχέσεων(CRM) σε γενικές γραμμές έχουν γίνει κεντρικά ζητήματα των επιχειρήσεων. Με πιο έντονο ανταγωνισμό σε πολλές ώριμες αγορές, οι εταιρείες έχουν αντιληφθεί ότι η ανάπτυξη των σχέσεων με περισσότερους επικερδείς πελάτες είναι ένας κρίσιμος παράγοντας για να μείνουν στην αγορά. Έτσι, τεχνικές διαχείρισης πελατειακών σχέσεων(CRM) έχουν αναπτυχθεί οι οποίες προσφέρουν νέες ευκαιρίες για τις επιχειρήσεις να λειτουργούν καλά σε μια αγορά. Η διαχείριση των πελατειακών σχέσεων(CRM) εστιάζει στον πελάτη και στις δυνατότητες για αύξηση των εσόδων, καθώς και με τον τρόπο αυτό, ενισχύει την ικανότητα μιας επιχείρησης να ανταγωνιστεί και να συγκρατήσει τους βασικούς της πελάτες. Η σχέση μεταξύ μιας επιχείρησης και των πελατών μπορεί να περιγραφεί ως εξής: Ένας πελάτης αγοράζει προϊόντα και υπηρεσίες, ενώ οι επιχειρήσεις αγοράζουν, πωλούν, παρέχουν και εξυπηρετούν τους πελάτες. Σε γενικές γραμμές, υπάρχουν τρεις τρόποι για τις επιχειρήσεις να αυξήσουν την αξία των πελατών:

- αύξηση της χρήσης (ή της αγοράς) των προϊόντων ή των υπηρεσιών που οι πελάτες έχουν ήδη
- πώληση στους πελάτες περισσοτέρων προϊόντων ή προϊόντα με μεγαλύτερη κερδοφορία
- διατήρηση πελατών για μεγάλο χρονικό διάστημα

Ένα αξιόλογος πελάτης δεν είναι συνήθως στατικός και η σχέση του εξελίσσεται και αλλάζει με το χρόνο. Έτσι, κατανόηση αυτής της σχέσης αποτελεί είναι ένα κρίσιμο μέρος της διαχείρισης πελατειακών σχέσεων (CRM). Αυτό μπορεί να επιτευχθεί με την ανάλυση του κύκλου ζωής του πελάτη ή τη διάρκεια της ζωής του πελάτη, η οποία αναφέρεται σε διάφορα στάδια της σχέσης μεταξύ του πελάτη και των επιχειρήσεων. Ένα χαρακτηριστικό του κύκλου ζωής του πελάτη φαίνεται στο σχήμα. 3.

Πρώτον, οι εκστρατείες εξαγοράς είναι εκστρατείες μάρκετινγκ που κατευθύνονται προς την αγορά και επιδιώκουν να επηρεάσουν τις προοπτικές και το ενδιαφέρον για το προϊόν ή την υπηρεσία μιας εταιρείας.




**Πίνακας 3. Απεικόνιση του κύκλου ζωής ενός πελάτη**

Εάν υπάρξει ανταπόκριση στην έρευνα της εταιρείας, τότε θα γίνουν ανταποκρινόμενοι. Οι ανταποκρινόμενοι γίνονται πελάτες, όταν η σχέση μεταξύ αυτών και των εταιρειών έχει τεκμηριωθεί. Για παράδειγμα, έχουν κάνει την αρχική αγορά ή η αίτηση τους για μια ορισμένη πιστωτική κάρτα έχει εγκριθεί. Σε αυτό το σημείο, οι εταιρείες θα αποκτήσουν έσοδα από τη χρήση που κάνουν οι πελάτες. Επιπλέον, η αξία των πελατών θα αυξηθεί, όχι μόνο από σταυροειδείς πωλήσεις(cross-selling)που ενθαρρύνουν τους πελάτες να αγοράζουν περισσότερα προϊόντα ή υπηρεσίες, αλλά και από το up-selling που ενθαρρύνει τους πελάτες να αναβαθμίσουν τα υφιστάμενα προϊόντα και τις υπηρεσίες.

Από την άλλη, σε κάποιο σημείο, κάποιοι μόνιμοι πελάτες σταματούν να είναι πελάτες (φιλαράκι - chum). Υπάρχουν δύο διαφορετικοί τύποι πελατών.

Οι πρώτοι είναι εθελοντικοί, πράγμα που σημαίνει ότι οι καθιερωμένοι πελάτες επιλέγουν να σταματήσουν να είναι πελάτες. Ο άλλος τύπος είναι οι υποχρεωμένοι, δηλαδή οι

καθιερωμένοι πελάτες που δεν είναι πλέον καλοί πελάτες και η εταιρεία ακυρώνει τη σχέση τους. Ο κύριος σκοπός της διαχείρισης πελατειακών σχέσεων(CRM) είναι η μεγιστοποίηση των αξιών των πελατών σε ολόκληρο τον κύκλο ζωής. Με μεγάλο όγκο δεδομένων να παράγεται στη διαχείριση πελατειακών σχέσεων(CRM), η εξόρυξη δεδομένων διαδραματίζει ηγετικό ρόλο στην όλη διαχείριση πελατειακών σχέσεων(CRM) (Rud και Brohman, 2003,Shawetal,2001).

Σε εκστρατείες εξαγοράς, η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για το προφίλ ανθρώπων που ανταποκρίθηκαν σε προηγούμενες παρόμοιες εκστρατείες και τα προφίλ εξόρυξης δεδομένων είναι χρήσιμα για να βρεθούν τα καλύτερα τμήματα της πελατείας που η εταιρεία θα πρέπει να στοχεύσει (Adomavicius και Tuzhilin, 2003). Μια άλλη εφαρμογή είναι να κοιτάξουμε για προοπτικές που έχουν παρόμοια πρότυπα συμπεριφοράς για τους σημερινούς μόνιμους πελάτες.

Σε εκστρατείες ανταπόκρισης, η εξόρυξη δεδομένων μπορεί να εφαρμοστεί για να καθοριστεί σε ποιες προοπτικές θα υπάρξει ανταπόκριση και από αυτούς που θα ανταποκριθούν ποιοι θα καθιερωθούν ως πελάτες. Οι καθιερωμένοι πελάτες είναι επίσης μια σημαντική περιοχή για την εξόρυξη δεδομένων. Η αναγνώριση του είδους της συμπεριφοράς των πελατών από δεδομένα χρήσης του πελάτη και η πρόβλεψη για το ποιοι πελάτες θα ανταποκριθούν στις πωλήσεις πιθανό είναι πολύ σημαντικό για τις επιχειρήσεις (Chiang και Lin, 2000). Όσον αφορά τους πρώην πελάτες, η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για να αναλύσει τη φυγή τους (Chiang et al., 2003).

Η βελτιστοποίηση διαδραματίζει επίσης σημαντικό ρόλο στη διαχείριση πελατειακών σχέσεων CRM και ειδικότερα στον καθορισμό του τρόπου ανάπτυξης της προληπτικής στρατηγικής αλληλεπίδρασης με τους πελάτες(proactive customer interaction strategy) για την μεγιστοποίηση της διάρκειας ζωής των πελατών. Ο πελάτης είναι κερδοφόρος, εάν τα έσοδα από αυτόν τον υπερβαίνουν το κόστος της εταιρείας για την προσέλκυση, πώληση και εξυπηρέτηση πελάτη του.

Το πλεόνασμα αυτό ονομάζεται «αξία μιας ζωής» (customer lifetime value - LTV). Με άλλα λόγια, το LTV είναι η συνολική αξία που μπορεί να προκύψει, ενώ ο πελάτης είναι ακόμα ενεργός και είναι ένα από τα πιο σημαντικά μετρήσιμα στοιχεία στη διαχείριση πελατειακών(CRM). Μεγάλο μέρος της έρευνας έχει γίνει στον τομέα της μοντελοποίησης του LTV χρησιμοποιώντας τεχνικές επιχειρησιακής έρευνας (Schmittlein et al, 1987,Blattberg και Deighton, 1991,Dreze και Bonfrer, 2002, Chingetal, 2004).

Ακόμα και όταν το μοντέλο LTV μπορεί να διατυπωθεί, είναι δύσκολο να βρεθεί η βέλτιστη λύση, με την παρουσία τόσο μεγάλου όγκου δεδομένων. Ορισμένοι ερευνητές το έχουν αντιμετωπίσει με τη χρήση της εξόρυξης δεδομένων για να βρουν τις βέλτιστες παραμέτρους για το μοντέλο LTV. Για παράδειγμα, Rosset et al. (2002) διατυπώνουν το παρακάτω LTV μοντέλο:

$$\text{LTV} = \int_0^\infty S(t)v(t)D(t)dt$$

όπου v (t) περιγράφει αξία του πελάτη με την πάροδο του χρόνου, S (t) περιγράφει την πιθανότητα ότι ο πελάτης είναι ακόμα ενεργός τη χρονική στιγμή t, και D (t) είναι ένας παράγοντας αποδοκιμασίας . Η εξόρυξη δεδομένων χρησιμοποιείται για την εκτίμηση της αξία των εσόδων των μελλοντικών πελατών (μελλοντική) και την εκτίμηση του πότε μπορεί να φύγει ο πελάτης με την πάροδο του χρόνου. Αυτό το πρόβλημα είναι πρόβλημα είναι δύσκολο στην πράξη, λόγω του μεγάλου όγκου των δεδομένων που εμπλέκονται.

Ο Padmanabhan και oTuzhilin (2003) παρουσίασαν δύο κατευθύνσεις ώστε να μειωθεί η πολυπλοκότητα του προβλήματος βελτιστοποίησης LTV. Μία κατεύθυνση είναι να βρεθεί καλή ευρετική για τη βελτίωση των τιμών LTV και η άλλη στρατηγική είναι η βελτιστοποίηση ορισμένων απλούστερων α μέτρων επίδοσης που σχετίζονται με την τιμή LTV. Όσο για την τελευταία κατεύθυνση, ο συγγραφέας επισημαίνει ότι η εξόρυξη δεδομένων και ή βελτιστοποίηση μπορούν να συγχωνευθούν ώστε να δημιουργηθεί προφίλ των πελατών, τα οποία είναι ζωτικής σημασίας σε πολλές εφαρμογές διαχείρισης πελατειακών σχέσεων (CRM).

Η εξόρυξη δεδομένων χρησιμοποιήθηκε για πρώτη φορά για να ανακαλύψουν τις συνήθειες των πελατών και στη συνέχεια χρησιμοποιούνται κανόνες καθώς και η βελτιστοποίηση για την επιλογή ενός μικρού αριθμού των καλύτερων μοντέλων από το παρελθόν. Τέλος, σύμφωνα με το προφίλ του πελάτη, η εταιρεία μπορεί να επιτύχει τη στοχοθέτηση και να δαπανήσει χρήματα για τους πελάτες που είναι πιθανόν να ανταποκριθούν εντός του προϋπολογισμού τους.

Η εκστρατεία βελτιστοποίησης είναι ένα άλλο πρόβλημα όπου μπορεί να εφαρμοστεί ένας συνδυασμός της εξόρυξης δεδομένων και της επιχειρησιακής έρευνας. Στη διαδικασία βελτιστοποίησης μια εταιρεία θα πρέπει να καθορίσει ποιού είδους προσφορές θα πρέπει να πάνε στο ποιο τμήμα πελατών και μέσω ποιου καναλιού επικοινωνίας.

Ο Vercellis (2002) παρουσιάζει δύο στάδια της εκστρατείας μοντέλων βελτιστοποίησης τόσο με την τεχνολογία εξόρυξης δεδομένων όσο και με την στρατηγική βελτιστοποίησης. Στην πρώτο στάδιο, τα υπό-προβλήματα βελτιστοποίησης επιλύονται για κάθε εκστρατεία και οι πελάτες διαχωρίζονται με βάση τις βαθμολογίες τους. Στο δεύτερο στάδιο, ένα μικτό μοντέλο βελτιστοποίησης είναι σχεδιασμένο για να λύσει το συνολικό πρόβλημα βελτιστοποίησης με βάση την κατηγοριοποίηση των πελατών και των περιορισμένων διαθέσιμων πόρων. Όπως περιγράφεται παραπάνω, πολλά προβλήματα διαχείρισης πελατειακών σχέσεων (CRM) θα μπορούσαν να διατυπωθούν ως προβλήματα βελτιστοποίησης, αλλά υπάρχουν συνήθως πολύ μεγάλοι όγκοι των δεδομένων που καθιστούν δύσκολη την επίλυση του προβλήματος.

Ο συνδυασμός του προβλήματος βελτιστοποίησης με την εξόρυξη δεδομένων είναι σημαντικό στο πλαίσιο αυτό. Για παράδειγμα, η εξόρυξη δεδομένων μπορεί να χρησιμοποιείται για τον προσδιορισμό μοτίβων από δεδομένα των πελατών και στη συνέχεια αυτά τα σχέδια μπορούν να χρησιμοποιηθούν για τον προσδιορισμό περισσότερων σχετικών περιορισμών για τα μοντέλα βελτιστοποίησης. Επιπλέον, η εξόρυξη δεδομένων μπορεί να εφαρμοστεί για να μειωθεί ο χώρος αναζήτησης και να βελτιωθεί ο χρόνος υπολογισμού.

Έτσι, η διερεύνηση του πως θα συνδυαστεί η βελτιστοποίηση και η εξόρυξη δεδομένων για την αντιμετώπιση των προβλημάτων διαχείρισης των πελατειακών σχέσεων (CRM) αποτελεί μια πολλά υποσχόμενη περιοχή της έρευνας για την κοινότητα της επιχειρησιακής έρευνας.

### 4.3.2. Εξατομίκευση

Εξατομίκευση είναι η ικανότητα παροχής περιεχομένου και υπηρεσιών προσαρμοσμένες στα άτομα με βάση τη γνώση που αποκτούμε σχετικά με τις προτιμήσεις και τη συμπεριφορά τους. Η έρευνα της εξόρυξης δεδομένων που σχετίζεται με την εξατομίκευση επικεντρώνεται κυρίως σε συστήματα συστάσεων και συναφή θέματα, όπως

είναι το συλλογικό φιλτράρισμα (collaborative filtering), και συστήματα συστάσεων έχουν διερευνηθεί εντατικά από την κοινότητα εξόρυξης δεδομένων (Breese et al. 1998, Geyer-Schulz και Hahsler, 2002, Lieberman, 1997; Lin et al., 2000). Τα συστήματα αυτά μπορούν να κατηγοριοποιηθούν σε τρεις ομάδες: συστήματα με βάση το περιεχόμενο, εξόρυξη κοινωνικών δεδομένων και συνεργατικό φιλτράρισμα. Τα συστήματα συστάσεων με βάση το περιεχόμενο χρησιμοποιούν αποκλειστικά τις προτιμήσεις του χρήστη που λαμβάνει τη σύσταση (Hilletal. 1995).

Αυτές οι προτιμήσεις που αντλήθηκαν μέσω σιωπηρών ή ρητών σχολίων των χρηστών, συνήθως αντιπροσωπεύουν το προφίλ του χρήστη. Τα συστήματα συστάσεων που βασίζονται στην εξόρυξη κοινωνικών δεδομένων λαμβάνουν υπόψη πηγές δεδομένων που δημιουργούνται από ομάδες ανθρώπων, ως μέρος της καθημερινής τους δραστηριότητας και γίνεται εξόρυξη των δεδομένων αυτών για δυνητικά χρήσιμες πληροφορίες. Ωστόσο, η σύσταση των συστημάτων εξόρυξης των κοινωνικών δεδομένων δεν είναι συνήθως εξατομικευμένες αλλά εκπέμπουν προς το σύνολο της κοινότητας των χρηστών. Από την άλλη, η εξατομίκευση επιτυγχάνεται με το συνεργατικό φιλτράρισμα (Resnicketal., 1994; Shardanan και Maes, 1995; Goodetal.,1999) το οποίο ταιριάζει χρήστες με παρόμοια ενδιαφέροντα και χρησιμοποιεί τις προτιμήσεις αυτών των χρηστών για να κάνουν συστάσεις.

Όπως υποστήριξαν ο Adomavicius και ο Tuzhilin (2003), το πρόβλημα της σύστασης μπορεί να διατυπωθεί ως πρόβλημα βελτιστοποίησης που επιλέγει τα καλύτερα στοιχεία για να τα συστήσει σε έναν χρήστη.

Συγκεκριμένα, δίνεται ένα σύνολο U χρηστών και ένα σύνολο V των στοιχείων, η συνάρτηση f: Ux V→ R μπορεί να οριστεί για να καθορίσει πώς σε κάθε χρήστη u ∈U αρέσει κάθε είδος v ∈ V. Το πρόβλημα της σύστασης μπορεί να διατυπωθεί ως το παρακάτω πρόβλημα βελτιστοποίησης:

$$\max f(u\ v)$$

$$\text{υπόκειται σε } u \in U$$

$$v \in V$$

Η πρόκληση για αυτό το πρόβλημα είναι ότι η συνάρτηση μπορεί συνήθως να οριστεί μερικώς, δηλαδή, όλες οι καταχωρήσεις στον πίνακα {f(u, v)}u2U,v2V δεν έχουν γνωστές τιμές. Ως εκ τούτου, είναι αναγκαίο να διευκρινιστεί πως οι άγνωστες τιμές θα πρέπει να εκτιμηθούν από το σύνολο των προκαθορισμένων τιμών (Padmanabhan και Tuzhilin, 2003). Πολυάριθμες μέθοδοι έχουν αναπτυχθεί για την εκτίμηση αυτών των τιμών και ο Pazzani (1999) και οι Adomavicius και Tuzhilin (2003) περιγράφουν ορισμένες από αυτές τις μεθόδους. Μόλις το πρόβλημα βελτιστοποίησης καθοριστεί, η εξόρυξη δεδομένων μπορεί να συμβάλει στην επίλυσή τους με την εκμάθηση πρόσθετων περιορισμών με μεθόδους εξόρυξης δεδομένων. Για περισσότερες πληροφορίες σχετικά με την εξατομίκευση στην επιχειρησιακή έρευνα παραπέμπουμε στους Murthi και Sarkar (2003), και σημειώνουμε ότι αυτή η περιοχή παρουσιάζει πλήθος ευκαιριών για την κοινότητα της επιχειρησιακής έρευνας.

# 5. Συμπεράσματα

Στην παρούσα πτυχιακή εργασία παρουσιάσαμε την σύνδεση και την αλληλεπίδραση δύο επιστημονικών πεδίων, της επιχειρησιακής έρευνας και της εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων έχει αναπτυχθεί με πολύ γρήγορο ρυθμό τα τελευταία χρόνια, και έχει βασιστεί σε ιδέες και ανθρώπους από διάφορα επιστημονικά πεδία.

Όμως, δεν έχει αποκρυσταλλωθεί ακόμη μια ενιαία ταυτότητα της Εξόρυξης Δεδομένων, και ένας ενιαίος τρόπος αλληλεπίδρασης των επιμέρους επιστημονικών πεδίων.

Επομένως:

- Οι άνθρωποι της Στατιστικής συνεχίζουν να κάνουν Στατιστική.
- Οι άνθρωποι της Τεχνητής Νοημοσύνης συνεχίζουν να κάνουν Τεχνητή Νοημοσύνη.
- Οι άνθρωποι των ΒΔ ρίχνουν το βάρος στις μεγάλες συλλογές δεδομένων,

και ...

- **σπανίως οι ειδικοί του ενός πεδίου ασχολούνται/συνεργάζονται με ειδικούς ενός άλλου πεδίου.**

Στην εποχή της οικονομικής κρίσης την οποία βιώνουμε, όπου όλο και περισσότερες επιχειρήσεις παρουσιάζουν ζημιές και ελλείμματα, η ανταγωνιστικότητα μέσα αλλά και μεταξύ των επιχειρήσεων αυξάνεται και γίνεται πλέον όχι μόνο ζήτημα κέρδους αλλά και βιωσιμότητας για κάθε εταιρία (ζήτημα μείωσης ζημιών ίσως πλέον).

Έτσι λοιπόν, είναι σημαντικό να υπάρχει σωστή διαχείριση κάθε λειτουργίας μιας επιχείρησης αλλά και μεταξύ των λειτουργιών συνολικά για να είναι υγιείς και βιώσιμη μια επιχείρηση. Η χρήση λογισμικού αλλά και διαδικτύου είναι πλέον στις μέρες μας δεδομένη για κάθε σημαντική ή και έστω σοβαρή επιχείρηση. Η γρήγορη και αποτελεσματική λήψη αποφάσεων, οι βέλτιστες λύσεις, η μείωση κόστους-προβλημάτων ήταν και γίνονται όλο και πιο κρίσιμα ζητήματα που πρέπει να αντιμετωπιστούν.

Με την ταχύτητα σε όλους τους τομείς και το ρητό «*Ο χρόνος είναι χρήμα*» να παίζουν σημαντικότατο ρόλο στον κόσμο των επιχειρήσεων αλλά και της σύγχρονης κοινωνίας, όπως επίσης με την αύξηση του όγκου των δεδομένων αλλά και της πολυπλοκότητας διαφόρων λειτουργιών και διαδικασιών **η απλούστευση** μέσω διάφορων μεθόδων και τεχνικών βελτιστοποίησης είναι το κλειδί έτσι ώστε να υπάρχει σωστή διαχείριση έργων αλλά και δεδομένων και μια επιχείρηση να μπορέσει να αποκτήσει ανταγωνιστικό πλεονέκτημα ή έστω τώρα ποια να επιβιώσει.

Ειδικότερα, όπως φαίνεται στην εργασία, η κοινότητα της επιχειρησιακής έρευνας έχει συμβάλει σημαντικά τα τελευταία χρόνια στον ολοένα αναπτυσσόμενο τομέα της εξόρυξης δεδομένων.

Η συμβολή των μεθόδων βελτιστοποίησης στην εξόρυξη δεδομένων αγγίζει κάθε τομέα της διαδικασίας εξόρυξης δεδομένων, από την οπτικοποίηση των δεδομένων και την προ-επεξεργασία, μέχρι την επαγωγική μάθηση και την επιλογή του καλύτερου μοντέλο μετά την εκμάθηση.

Επιπλέον, η εξόρυξη δεδομένων μπορεί να είναι χρήσιμη σε πολλούς τομείς της επιχειρησιακής έρευνας και μπορούν να χρησιμοποιηθούν με συμπληρωματικό τρόπο ώστε

να βοηθήσουν στη βελτιστοποίηση της μεθόδου για τον εντοπισμό των ορίων και της μείωσης του χώρου αναζήτησης.

Παρά το γεγονός ότι υπάρχει ήδη μεγάλος όγκος εργασιών που καλύπτουν το σημείο τομής της επιχειρησιακής έρευνας και της εξόρυξης δεδομένων, θεωρούμε ότι η τρέχουσα εργασία είναι μόνο η αρχή.

Το ενδιαφέρον για την εξόρυξη δεδομένων συνεχίζει να αυξάνεται τόσο στα πανεπιστήμια όσο και στη βιομηχανία και στον επιχειρηματικό κόσμο, τα περισσότερα θέματα εξόρυξης δεδομένων στα οποία υπάρχει η δυνατότητα να χρησιμοποιηθούν μέθοδοι βελτιστοποίησης απαιτούν πολύ περισσότερη έρευνα.

Αυτό αντιμετωπίζεται επί του παρόντος, ως αυξανόμενο ενδιαφέρον για την εξόρυξη δεδομένων μέσα στην κοινότητα της επιχειρησιακής έρευνας και ελπίζουμε ότι αυτή η πτυχιακή εργασία θα βοηθήσει στην παραπέρα κινητοποίηση περισσότερων ερευνητών που θα συμβάλουν σε αυτό το συναρπαστικό τομέα.

# 6. Βιβλιογραφία


- Operations research and data miningSigurdurOlafsson , Xiaonan Li, Shuning Wu *European Journal of Operational Research Available online at* www.sciencedirect.com
- ΑΛΓΟΡΙΘΜΟΙ ΚΑΙ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΡΟΕΣ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ ΝΙΚΟΛΑΟΣ Χ. ΤΣΙΡΑΚΗΣ Μεταπτυχιακή εργασία
- ΑΛΓΟΡΙΘΜΟΙ ΣΥΝΔΥΑΣΤΙΚΗΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΜΕ ΕΜΦΑΣΗ ΣΕΜΕΤΑΕΥΡΕΤΙΚΕΣ ΤΕΧΝΙΚΕΣ Διδακτορική διατριβή ΧΡΗΣΤΟΣ Γ. ΓΚΟΓΚΟΣ
- Αλγόριθμος Simplex και ειδικές μέθοδοι επίλυσης προβλημάτων γραμμικού προγραμματισμού με χρήση Η/Υ
- Διπλωματική εργασία Μαναμσίδης Οδυσσέας Γραμμικός Προγραμματισμός Δικτυακός προγραμματισμός, Πολυκριτηριακή βελτιστοποίηση, Μη γραμμικές μέθοδοι βελτιστοποίησης – Εξελικτικοί και γενετικοί αλγόριθμοι
- Ανδρέας Ευστρατιάδης και Δημήτρης Κουτσογιάννης Τομέας Υδατικών Πόρων Εθνικό Μετσόβιο Πολυτεχνείο
- Επιχειρησιακή Έρευνα ΤΕΙ Χαλκίδας Σχολή Διοίκησης και Οικονομίας Τμήμα Διοίκησης Επιχειρήσεων Αλέξιος Πρελορέντζος, ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ
- Επιστήμη των Αποφάσεων, Διοικητική Επιστήμη, Δημήτρης Λέκκας Επίκουρος Καθηγητής Τμήμα Στατιστικής & Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών


## Βιβλία:


- Επιχειρησιακή έρευνα ΛΗΨΗ ΕΠΙΧΕΙΡΗΜΑΤΙΚΩΝ ΑΠΟΦΑΣΕΩΝ Παντελή Γ. Υψηλάντη Εκδόσεις ΕΛΛΗΝ
- Εξόρυξη Γνώσης από Βάσεις Δεδομένων. Μ. Βαζιργιάννης και Μ. Χαλκίδη, Μ. Η. Dunham, Data Mining, Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα. Επιμέλεια Ελληνικής Έκδοσης: Β. Βερύκιος και Γ. Θεοδωρίδης. Εκδόσεις
- SHAUM'S Επιχειρησιακή έρευνα Δεύτερη αμερικάνικη έκδοση εκδόσεις κλειδάριθμος
- Θέματα επιχειρηματικής Νοημοσύνης: Θεωρητική Θεμελίωση & Εφαρμογές Εκδόσεις Κωσταράκη Αθήνα 2003 Βουτσινά Β.
- Boutsina B, Antzoulatos G Alevizos P. A novel classification algorithm based on clustering In: 1st International Conference "From Scientific Computing to Computuational Engineering". Athens Greece 2004.
- Burges C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 1998
- M. H. Dunham, Data Mining – Introduction and Advanced topics

- Kaufman, Leonard and Rousseeuw, Peter J.,Finding groups in data: an introduction to cluster analysis,John Wiley & Sons,1990
- McQueen, J., Some methods for classification and analysis of multivariate observations. 5th Berkeley Symposium on mathematics, Statistics and Probability
- Abdullah S, Ahmadi S, Burke E and Dror M. (2006). Investigating Ahuja-Orlin's large neighborhood search approach for examination timetabling.OR Spectrum, Vol. 29
- Τεχνητή Νοημοσύνη *Μια σύγχρονη προσέγγιση* Stuart Russell & Peter Norvig Εκδόσεις Κλειδάριθμος
- Di Caspero L., Schaerf A. (2001). Tabu Search Techniques for Examination
- Timetabling. Burke and Erben (eds). PATAT 2000 LNCS 2079
- Τεχνητά Νευρωνικά Δίκτυα: Θεωρία και Εφαρμογές, Εκδόσεις Νέων Τεχνολογιών, Αθήνα 1996. Ρίζος Γ.


**Αναφορές 4ου Κεφαλαίου**


- Abbiw-Jackson, R., Golden, B., Raghavan, S., Wasil, E., 2006. A divide-and-conquer local search heuristic for data visualization. Computers and Operations Research
- Adomavicius, G., Tuzhilin, A., 2003. Recommendation technologies: Survey of current methods and possible extensions. Working paper, Stern School of Business, New York University, New York.
- Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data
- Bennett, K.P., Campbell, C., 2000. Support vector machines: Hype or hallelujah?. SIGKDD Explorations
- Berry, M., Linoff, G., 2000. Mastering Data Mining. Wiley, New York.
- Blattberg, R., Deighton, J., 1991. Interactive marketing: Exploiting the age of addressability. Sloan Management Review
- Castillo, E., Gutie´rrez, J.M., Hadi, A.S., 1997. Expert Systems and Probabilistic Network Models. Springer, Berlin.
- Chiang, I., Lin, T., 2000. Using rough sets to build-up web-based one to one customer services. IEEE Transactions.
- Ching, W., Wong, K., Altman, E., 2004. Customer lifetime value: Stochastic optimization approach. Journal of the Operational Research Society

- Cooper, G.F., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning

- Dhar, V., Chou, D., Provost, F., 2000. Discovering interesting patterns for investment decision making with GLOWER – a genetic learner overlaid with entropy reduction. Data Mining and Knowledge Discovery

- Dreze, X., Bonfrer, A., 2002. To pester or leave alone: Lifetime value maximization through optimal communication timing. Working paper, Marketing Department, University of California, Los Angeles, CA.

- Estevill-Castro, V., 2002. Why so many clustering algorithms – a position paper. SIGKDD Explorations
- Fayyad, U., Piatetsky-Shapiro, G., Smith, P., Uthurusamy, R.,1996. Advances in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA.

- Felici, G., Truemper, K., 2002. AMINSAT approach for learning in logic domains.INFORMSJournal on Computing

- Fu, Z., Golden, B., Lele, S., Raghavan, S., Wasil, E., 2003b.
Genetically engineered decision trees: Population diversity produces smarter trees. Operations Research
-
- Fu, Z., Golden, B., Lele, S., Raghavan, S., Wasil, E., 2006. Diversification for smarter trees. Computers and Operations Research.

- Glover, 1990. Improved linear programming models for discriminant analysis. Decision Sciences

- Glover, F., Laguna, M., 1997. Tabu Search. Kluwer Academic, Boston.

- Glover, F., Laguna, M., Marti, R., 2003. Scatter search. In: Tsutsui, Ghosh (Eds.), Theory and Applications of Evolutionary Computation: Recent Trends. Springer, Berlin, pp.

- Glover, F., Kochenberger, G.A., 2003. Handbook of Metaheuristics. Kluwer Academic Publishers, Boston, MA.

- Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York. Kennedy, H., Chinniah, C., Bradbeer, P., Morss, L., 1997. The construction and evaluation of decision trees: A comparison of evolutionary and concept learning methods. In: Come, D., Shapiro, J. (Eds.), Evolutionary Computing, Lecture Notes in Computer Science. Springer, Berlin,

- Kim, Y., Street, W.N., Menczer, F., 2000. Feature selection in nunsupervised learning via evolutionary search. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-00)

- Lam, W., Bacchus, F., 1994. Learning Bayesian belief networks. An approach based on the MDL principle. Computational Intelligence
- Lauritzen, S.L., 1995. The EM algorithm for graphical association models with missing data. Computational Statistics and Data Analysis Lee, J.-Y., Olafsson, S., 2006. Multiattribute decision trees and decision rules. In: Triantaphyllou, Felici (Eds.), Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques,

- Li, X., Olafsson, S., 2005. Discovering dispatching rules using data mining. Journal of Scheduling

- Mangasarian, O.L., 1965. Linear and nonlinear separation of patterns by linear programming. Operations Research

- Olafsson, S., Yang, J., 2004. Intelligent partitioning for feature selection. INFORMS Journal on Computing

- Rastogi, R., Shim, K., 1998. Mining optimized association rules for categorical and numeric attributes. In: Proceedings of International Conference of Data Engineering.
- Resende, M.G.C., Ribeiro, C.C., 2003