ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ (DATA MINING) ΚΑΙ ΚΑΤΗΓΟΡΙΚΑ ΔΕΔΟΜΕΝΑ

Γεράσιμος Ε. Σταυλιώτης

Διπλωματική εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Ιούνιος 2008

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από την ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ'αριθμ. ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Κατέρη Μαρία (Επιβλέπουσα)
- Πολίτης Κωνσταντίνος
- Κοφίδης Ελευθέριος

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

DATA MINING AND CATEGORICAL DATA

By

Gerasimos E. Stavliotis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

June 2008

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τη κα. Κατέρη Μαρία, επ. Καθηγήτρια του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, για την καθοδήγηση και υποστήριξη κατά την υλοποίηση της παρούσας εργασίας, αλλά και για την πολύ καλή συνεργασία που είχαμε όλο αυτό το διάστημα.

Παράλληλα, θα ήθελα να εκφράσω τις ευχαριστίες μου στον κ. Πολίτη Κωνσταντίνο, επ. Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, καθώς και τον κ. Κοφίδη Ελευθέριο, Λέκτορα του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, για τη συμμετοχή τους στην τριμελή επιτροπή και για τις συμβουλές τους.

Τέλος, θέλω να ευχαριστήσω την οικογένειά μου για την στήριξη έως την ολοκλήρωση των σπουδών μου, καθώς και όσους συναδέλφους και φίλους με βοήθησαν τους τελευταίους μήνες.

ΣΤΑΥΛΙΩΤΗΣ ΓΕΡΑΣΙΜΟΣ

Ν. ΗΡΑΚΛΕΙΟ

Ιούνιος 2008

Περίληψη

Στις μέρες μας, η τεχνολογία μας επιτρέπει να συγκεντρώνουμε και να αποθηκεύουμε απεριόριστη πληροφορία σε σχετικό λογισμικό. Μια από τις πιο προκλητικές εργασίες της εποχής μας είναι η ανακάλυψη προτύπων, τάσεων και ανωμαλιών σε τεράστια σύνολα δεδομένων, καθώς και η σύνοψή τους μέσω απλών και εύχρηστων μοντέλων.

Είναι βέβαιο ότι ζούμε στην κοινωνία της πληροφορίας, όπου η μετατροπή των δεδομένων σε πληροφορία απαιτείται να οδηγεί στη μετατροπή της πληροφορίας σε γνώση. Η συνύπαρξη ετερόκλητων επιστημονικών πεδίων όπως της στατιστικής, της μηχανικής εκμάθησης, της θεωρίας της πληροφορίας και των υπολογιστικών διαδικασιών, έχει δημιουργήσει μια νέα επιστήμη με δυναμικά εργαλεία.

Η επιστήμη αυτή καλείται «Εξόρυξη Δεδομένων (ΕΔ)» (*Data Mining*) και είναι μέρος της διαδικασίας «Ανακάλυψης Γνώσης από Βάσεις Δεδομένων» (*Knowledge Discovery in Databases - KDD*). Τα εργαλεία της ΕΔ είναι οι αλγόριθμοί της, οι οποίοι επιχειρούν να βρουν χρήσιμα και κατανοητά πρότυπα στα δεδομένα.

Κύριος στόχος της Διπλωματικής Εργασίας μας είναι η συγκέντρωση βασικών αλγορίθμων και μεθόδων που επιλέγουν και καθαρίζουν δεδομένα, αναγνωρίζουν πρότυπα, βελτιστοποιούν ένα σύστημα διαχείρισης και συσταδοποιούν δεδομένα. Θα δώσουμε έμφαση σε αλγορίθμους που είναι κατάλληλοι για κατηγορικά δεδομένα. Επίσης, ενδιαφερόμαστε και για ένα τρίτο τύπο δεδομένων που καλείται «μικτά δεδομένα» και περιλαμβάνει αριθμητικά και κατηγορικά δεδομένα.

Εκτός από την καταγραφή των μεθόδων και εφαρμογών της ΕΔ και της KDD, θα εφαρμόσουμε τεχνικές συσταδοποίησης σε ένα κατηγορικό σύνολο δεδομένων, το οποίο περιλαμβάνει τους περσινούς δανειολήπτες της επιχειρηματικής μονάδας στεγαστικής πίστης μεγάλης τράπεζας. Η προσπάθειά μας έγκειται στην περιγραφή της ομάδας των πελατών που είναι πιθανό να αποπληρώσουν το στεγαστικό τους δάνειο και να απομακρύνουν το χαρτοφυλάκιό τους από τον τραπεζικό όμιλο.

Abstract

In our days, technology allows us to collect and store unlimited information in relevant software. One of the most challenging tasks of our era is discovering patterns, trends and anomalies in huge datasets, and summarizing them through simple and practical models.

It is sure that we live in the information community, where conversion of data into information must lead in conversion of information into knowledge. The coexistence of heterogeneous scientific fields, such as statistics, machine learning, information theory and computing has created a new science with powerful tools.

This science is called “Data Mining (DM)” and is part of the “Knowledge Discovery in Databases (KDD)” process. DM’s tools are its algorithms, which try to find useful and understandable patterns in data.

The main aim of our MSc Dissertation is the collection of basic algorithms and methods that select and clean data, recognize patterns, optimize a management system and cluster data. We will emphasize on algorithms that are suitable for categorical data. Also, we are interested in a third type of data, which is called “mixed data” and includes both numerical and categorical data.

Apart from the recording of DM and KDD methods and applications, we will perform data clustering techniques on a categorical dataset that contains last year’s loan-takers of a top-ranking bank’s mortgage lending business unit. Our effort lies in describing the group of clients that are probable to pay off their housing loan and remove their portfolio from the banking group.

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κατάλογος συντομογραφιών

ΕΔ	Εξόρυξη Δεδομένων (*Data Mining*)
KDD	*Knowledge Discovery in Databases* (ανακάλυψη γνώσης από βάσεις δεδομένων)

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Τι είναι η εξόρυξη δεδομένων;

Η σύγκλιση της προόδου υπολογιστικών συστημάτων και της εξέλιξης στην επικοινωνία έχει οδηγήσει στην δημιουργία μιας κοινωνίας ικανής να παρέχει διαρκώς νέες πληροφορίες. Το υλικό που συγκεντρώνεται καταγράφεται διαρκώς, με αποτέλεσμα τη δημιουργία τεράστιων βάσεων δεδομένων. Το γεγονός αυτό αποτελεί ένα σύγχρονο φαινόμενο, το οποίο παρατηρείται ως ανάγκη από τα απλούστερα ζητήματα της καθημερινής ζωής έως και τα πιο σύνθετα.

Ας σκεφτούμε, για παράδειγμα, ότι σε ένα σύστημα βάσεων δεδομένων μπορούν να καταγραφούν οι συναλλαγές που γίνονται σε ένα κατάστημα ή η χρήση πιστωτικών καρτών και δανείων από τους πελάτες μίας τράπεζας (συστήματα δοσοληψιών). Από την άλλη πλευρά, υπάρχουν και πολυπλοκότερα ζητήματα που χρειάζεται να οργανωθούν μέσω μιας βάσης δεδομένων, όπως θέματα ιατρικής φύσεως, φωτογραφίες από δορυφόρους, πειραματικά δεδομένα (διαδικασίες συσσώρευσης ψηφιακών αρχείων).

Το ζήτημα, λοιπόν, που προκύπτει είναι εάν υπάρχει τρόπος να διαχειριστούμε τις πολύ μεγάλες αυτές βάσεις δεδομένων που ανανεώνονται διαρκώς από τους χρήστες. Επίσης, θεωρείται αρκετά δύσκολη η άντληση του απαραίτητου υλικού από αυτές. Όλα αυτά τα θέματα προκάλεσαν το ενδιαφέρον και οδήγησαν στη διαδικασία της **Εξόρυξης Δεδομένων** (*Data Mining*). Πρόκειται για μία σειρά από τεχνικές που βασίζονται σε ανάπτυξη αλγορίθμων και είναι χρήσιμες σε πολλούς και ετερόκλητους κλάδους όπως οι: οικονομία, βιοστατιστική, δημαγραφία, μετεωρολογία και γεωλογία.

Αξίζει να αναφέρουμε ότι υπάρχουν αντικρουόμενες απόψεις γύρω από το ποιος θα μπορούσε να είναι ένας σαφής και περιεκτικός ορισμός για την Εξόρυξη Δεδομένων (ΕΔ). Ωστόσο, αναφέρουμε έναν ορισμό (Hand et al., 2001) που θεωρούμε κατάλληλο:

> «Εξόρυξη Δεδομένων (*Data Mining*) είναι η ανάλυση – συνήθως τεράστιων – παρατηρούμενων (*observational*) συνόλων δεδομένων, έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων».

Η δήλωση των σχέσεων και η σύνοψη των στοιχείων στην οποία αναφέρεται ο ορισμός αυτός, συχνά αναφέρεται ως **μοντέλο** (*model*) ή **πρότυπο** (*pattern*). Βασικοί στόχοι της ΕΔ είναι η **περιγραφή** και η **πρόβλεψη**. Δηλαδή, η αναγνώριση των προτύπων (*pattern recognition*) που επικρατούν σε ένα μεγάλο σύνολο δεδομένων και η δημιουργία προβλέψεων όσον αφορά τη μελλοντική αξία ή συμπεριφορά κάποιων μεταβλητών. Η αναγνώριση των προτύπων γίνεται μέσω γραμμικών εξισώσεων, κανόνων, διάκρισης σε συστάδες, απόδοσης γραφημάτων και δομών σε μορφή δέντρου, καθώς και επαναλαμβανόμενων προτύπων σε μορφή χρονοσειρών.

Στο δεύτερο κεφάλαιο της εργασίας αυτής, θα προσδιορίσουμε τις διαφορές μεταξύ περιγραφής και πρόβλεψης, καθώς και ανάμεσα στο μοντέλο και το πρότυπο. Όμως, ένα σημαντικό σημείο, στο οποίο πρέπει να σταθούμε, είναι ότι ο παραπάνω ορισμός αναφέρεται σε **παρατηρούμενα δεδομένα** (*observational data*) και όχι σε εμπειρικά ή πειραματικά (*experimental*), καθώς η ΕΔ στην ουσία ασχολείται με δεδομένα που έχουν ήδη συλλεχθεί για κάποιο σκοπό πέρα από την ανάλυση που θα γίνει μέσω των διαδικασιών της.

Το προηγούμενο σχόλιο δείχνει ότι ουσιαστικός σκοπός της ΕΔ δεν είναι η ανάπτυξη στρατηγικής ως προς τη συλλογή δεδομένων. Αυτός είναι ένας λόγος που η ΕΔ διαφέρει εν μέρει από τη Στατιστική αν και, όπως θα δούμε παρακάτω, οι δύο αυτοί κλάδοι σχετίζονται μεταξύ τους. Στην ουσία, η διαφορά της Στατιστικής είναι ότι εκεί συλλέγονται συχνά δεδομένα χρησιμοποιώντας αποτελεσματικές στρατηγικές, με σκοπό να απαντηθούν συγκεκριμένα ζητήματα. Για το λόγο αυτό, η ΕΔ αναφέρεται συχνά ως **«δευτερογενής»** ανάλυση δεδομένων.

1.2 Οι απαρχές της εξόρυξης δεδομένων

Στα πλαίσια της αναζήτησης περισσότερο αποτελεσματικών και δυναμικών εργαλείων διαχείρισης διαφορετικής φύσεως δεδομένων, ερευνητές από διάφορους επιστημονικούς κλάδους επιχείρησαν να ενώσουν τα αντικείμενα του ενδιαφέροντός τους. Η συνεργασία αυτή βρήκε πρόσφορο έδαφος στο πεδίο της ΕΔ, βασιζόμενη στην εφαρμογή μεθοδολογιών και αλγορίθμων που είχαν ήδη χρησιμοποιηθεί από τους ερευνητές.

Πιο συγκεκριμένα (Tan et al., 2005), η ΕΔ χρησιμοποιεί έννοιες όπως δειγματοληψία, εκτίμηση και έλεγχος υποθέσεων από τη Στατιστική, καθώς και εφαρμογές όπως αναζήτηση αλγορίθμων, τεχνικές δημιουργίας υποδειγμάτων (*modeling techniques*), θεωρίες τεχνητής νοημοσύνης (*artificial intelligence*), αναγνώρισης προτύπων και μηχανικής εκμάθησης

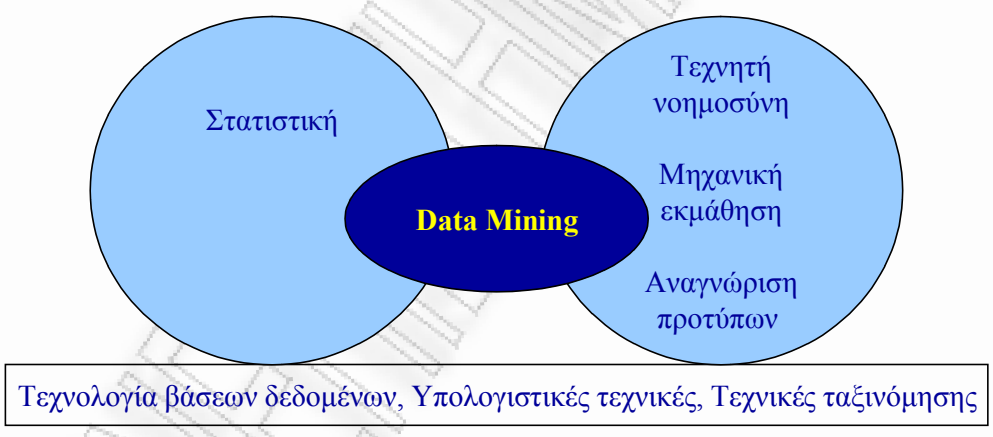
(*machine learning*). Άλλες έννοιες στις οποίες βασίζεται η ΕΔ είναι η θεωρία βελτιστοποίησης (*optimization*), εξελικτικός υπολογισμός (*evolutionary computing*), θεωρία της πληροφορίας (*information theory*), έλεγχος σημάτων (*signal processing*), νοερή απεικόνιση (*visualization*) και ανάκτηση πληροφορίας (*information retrieval*).

Επιπλέον, υπάρχουν αρκετοί άλλοι τομείς των επιστημών που στήριξαν την πρόοδο της ΕΔ. Για παράδειγμα, η τεχνολογία των βάσεων δεδομένων (*Database Technology*) δύναται να παρέχει βοήθεια μέσω τεχνικών αποτελεσματικής αποθήνευσης (*efficient storage*), ευρετηριοποίησης (*indexing*), διεξαγωγής ερωτημάτων (*processing*). Ακόμη, τεχνικές υψηλής απόδοσης από υπολογιστικής πλευράς (*Parallel Computing*) και σχετικές με την ταξινόμηση (*Distributed Computing*) παρέχουν βοήθεια σε σχέση με τη διαχείριση του μεγέθους και της συλλογής των τεράστιων συνόλων δεδομένων.

ΣΧΗΜΑ 1.1

Η εξόρυξη δεδομένων ως συνέπεια πολλών κλάδων

[Μετάφραση από: Tan et al., 2005]



Όπως έχουμε ήδη αναφέρει, η ουσία των εφαρμογών της ΕΔ είναι η αναγνώριση προτύπων. Η προσπάθεια αυτή έχει ξεκινήσει από πολύ παλιά, ενώ αρκετά ονόματα επιχειρήθηκαν κατά καιρούς να δοθούν στη διαδικασία αυτή. Όμως, ο τίτλος «εξόρυξη δεδομένων» που χρησιμοποιούμε και εμείς είναι αυτός που τελικά επικράτησε.

Η ΕΔ ανήκει στη γενικότερη μεθοδολογία της **ανακάλυψης της γνώσης από βάσεις δεδομένων** (*Knowledge Discovery in Databases – KDD*), με την οποία θα ασχοληθούμε εκτενέστερα στο επόμενο κεφάλαιο. Η ονομασία αυτή της KDD χρησιμοποιείται από το 1989

(πρώτο συνέδριο KDD) με στόχο να φανεί ότι η γνώση είναι το τελικό προϊόν μιας ανακάλυψης καθοδηγούμενης από τα δεδομένα (Piatetsky-Shapiro, 1991).

1.3 Ανακαλύπτοντας την «κρυμμένη γνώση»

Επεξεργαζόμενοι μια τεράστια βάση δεδομένων, είναι πολύ πιθανό να ανακαλύψουμε την ύπαρξη «κρυμμένης γνώσης». Δηλαδή, μπορεί να εντοπίσουμε συσχετίσεις, αλληλοεξάρτηση ή ομαδοποιήσεις μεταξύ των δεδομένων, πράγματα τα οποία μπορεί να μην είναι άμεσα εμφανή. Το είδος αυτό της γνώσης θεωρείται ότι δεν είναι εκ των προτέρων διαθέσιμο, αλλά μπορεί να αποδειχθεί πολύ χρήσιμο.

Υπό αυτές τις συνθήκες, κρίνεται απαραίτητη η «μη επιβλεπόμενη» ανάκτηση γνώσης, που υποστηρίζεται από την εφαρμογή αλγορίθμων. Στόχος είναι η ανακάλυψη της κρυμμένης γνώσης. Αυτήν την ανάγκη έρχεται να καλύψει η ΕΔ, η οποία αποτελεί τον πυρήνα της διαδικασίας ανακάλυψης της γνώσης από βάσεις δεδομένων (KDD).

Η διαδικασία KDD αναφέρεται στη διεργασία εξόρυξης γνώσης από μεγάλες αποθήκες δεδομένων. Ο όρος ΕΔ χρησιμοποιείται ως συνώνυμο της KDD, αλλά αποτελεί αναφορά στις πραγματικές τεχνικές που χρησιμοποιούνται για την ανάλυση και εξαγωγή της γνώσης από διάφορα σύνολα δεδομένων. Για να είναι σαφής η διαφορά μεταξύ διαδικασίας και εργαλείων, ο όρος KDD χρησιμοποιείται για την περιγραφή ολόκληρης της διαδικασίας ανακάλυψης γνώσης από ένα σύνολο δεδομένων, ενώ ο όρος ΕΔ αναφέρεται στις τεχνικές που χρησιμοποιούνται για την ανακάλυψη της γνώσης.

Ένας άλλος όρος που χρησιμοποιείται αντί της ΕΔ είναι «εξόρυξη γνώσης». Θεωρείται, όμως, ότι ο όρος αυτός δε δίνει έμφαση στην ανάλυση και εξαγωγή προτύπων. Ο όρος ΕΔ αντιπροσωπεύει καλύτερα τη διαδικασία εύρεσης δομών γνώσης που περιγράφουν με ακρίβεια σύνολα πρωτογενών δεδομένων. Οι δομές αυτές αναδεικνύουν κρυμμένη γνώση (συνάφειες / κανόνες) που δεν είναι άμεσα ορατή και εκμεταλλεύονται πιθανές κοινές ιδιότητες των πρωτογενών δεδομένων.

Τα τελευταία χρόνια, παρατηρείται μεγάλη ερευνητική και βιομηχανική δραστηριότητα στο χώρο της ΕΔ και γενικότερα της KDD. Ο κλάδος αυτός έχει μεγάλη εξελιξιμότητα, ενώ κάθε χρόνο γίνονται αρκετά διεθνή συνέδρια και εκδίδονται επιστημονικά περιοδικά που προσφέρουν ιδιαίτερα ενδιαφέρον υλικό.

1.4 Διαδικασία και απαιτήσεις

Οι διαδικασίες της ΕΔ αποσκοπούν στη δημιουργία μοντέλων συναρμολογήσεων ή την εξαγωγή προτύπων των υπό εξέταση δεδομένων. Γνωρίζοντας τις παραμέτρους του μοντέλου μέσα από τα δεδομένα που υπάρχουν ή τα πρότυπα που προσδιορίζονται, εφαρμόζουμε τους κατάλληλους αλγορίθμους ΕΔ.

Οι αλγόριθμοι αυτοί βασίζονται σε τομείς όπως στατιστική και μηχανική μάθηση, αλλά η διαφορά τους είναι ότι έχουν τέτοιο σχεδιασμό ώστε να υπάρχει εξελιξιμότητά τους σε σχέση με το μέγεθος του συνόλου δεδομένων που εισάγεται.

1.4.1 Βασικά στάδια της εξόρυξης δεδομένων

Σε αυτήν την υποενότητα αναφέρουμε ενδεικτικά σε τρία στάδια τη διαδικασία της ΕΔ, έτσι όπως περιγράφεται από τους Fayyad et al. (1996-a). Ο διαχωρισμός αυτός σε συγκεκριμένα στάδια θα μας βοηθήσει να κάνουμε μια αρχική προσέγγιση των διαδικασιών της ΕΔ, αλλά και να εμπεδώσουμε στη συνέχεια τις απαιτήσεις της.

Έτσι, τα τρία στάδια της ΕΔ είναι:

1. **Περιγραφή μοντέλου**

   Στο πρώτο στάδιο της ΕΔ, επιχειρούμε να δηλώσουμε τη **λειτουργία** του μοντέλου, δηλαδή να δηλώσουμε το στόχο μας, όπως για παράδειγμα την **ταξινόμηση** (*classification*), την **παλινδρόμηση** (*regression*) ή τη **συσταδοποίηση** (*clustering*). Αυτοί οι εναλλακτικοί τύποι ΕΔ θα σχολιαστούν εκτενέστερα στο επόμενο κεφάλαιο, καθώς κάθε ένας από αυτούς εξυπηρετεί διαφορετικό σκοπό.

   Επίσης, μας ενδιαφέρει η **παραστατική μορφή** του μοντέλου, δηλαδή η απεικόνισή του έτσι ώστε να ταιριάζει με την απεικόνιση των δεδομένων και να είναι δυνατό να ερμηνευθεί. Χαρακτηριστικά παραδείγματα μοντέλων είναι τα δέντρα απόφασης, τα γραφικά μοντέλα, τα νευρωνικά δίκτυα και τα μοντέλα – συστήματα που βασίζονται σε παραδείγματα ή πιθανότητες (δίκτυα Bayes).

2. **Αξιολόγηση μοντέλου**

   Ύστερα από τη δημιουργία του μοντέλου, οφείλουμε να εξετάσουμε κατά πόσο ταιριάζει με τις συνθήκες της KDD. Προχωράμε, δηλαδή, στην αξιολόγηση του μοντέλου, ώστε να

κρίνουμε την εγκυρότητα των προτύπων και την ακρίβεια και χρησιμότητα του μοντέλου. Υπάρχουν διάφορα κριτήρια αξιολόγησης, όπως αυτό της μέγιστης πιθανότητας.

3. Αλγόριθμος αναζήτησης

Στόχος του σταδίου αυτού είναι η σύγκριση μοντέλων και παραμέτρων, δοθέντων του συνόλου δεδομένων, της οικογένειας μοντέλων και του κριτηρίου αξιολόγησης. Οι βασικότεροι τύποι αλγορίθμων αναζήτησης είναι αυτοί που αναζητούν **παραμέτρους βελτιστοποίησης** ενός κριτηρίου αξιολόγησης και αυτοί που αναζητούν **μοντέλα αντιπροσώπευσης** των δεδομένων.

1.4.2 Απαιτήσεις της εξόρυξης δεδομένων

Για να έχουμε ένα ολοκληρωμένο αποτέλεσμα από μια διαδικασία ΕΔ, πρέπει αρχικά να ελέγξουμε τα χαρακτηριστικά που αναμένουμε να έχει το σύστημα ΕΔ, καθώς και τις απαιτήσεις για την εφαρμογή των τεχνικών.

Με βάση τους Chen et al. (1996), τα κυριότερα ζητήματα που οφείλουμε κάθε φορά να λαμβάνουμε υπόψη είναι:

i) Χειρισμός διαφορετικών τύπων δεδομένων

Είναι ξεκάθαρο ότι ένα σύστημα ΕΔ πρέπει να μπορεί να εφαρμόζεται σε διαφορετικούς τύπους δεδομένων, καθώς συχνά χρησιμοποιούνται διαφορετικοί τύποι και ΒΔ σε διαφορετικές εφαρμογές. Επίσης, παρατηρείται συχνά η ύπαρξη συγγενών (*relational*) βάσεων δεδομένων. Επομένως, πρέπει ένα σύστημα ΕΔ να είναι σε θέση να υποστηρίζει τεχνικές για αποδοτική και αποτελεσματική ανάλυση συγγενικών δεδομένων.

Τέλος, ένα τέτοιο σύστημα θα έπρεπε να λειτουργεί ανεξάρτητα από τύπους δεδομένων, καθώς πολλά σύγχρονα συστήματα βάσεων δεδομένων περιέχουν σύνθετους τύπους δεδομένων (δομές δεδομένων και σύνθετα αντικείμενα, υπερκείμενο και στοιχεία πολυμέσων, χωροχρονικά στοιχεία κ.λπ.).

Η ποικιλία των τύπων δεδομένων και οι διαφορετικοί στόχοι της ΕΔ κάνουν πιο απίθανη την ύπαρξη ενός συστήματος ΕΔ που να μπορεί να χειριστεί όλα αυτά τα είδη δεδομένων. Καλό θα ήταν να διαμορφωθούν εξειδικευμένα συστήματα για εξόρυξη γνώσης πάνω σε συγκεκριμένους τύπους δεδομένων όπως βάσεις δεδομένων πολυμέσων, συστήματα που

ασχολούνται αποκλειστικά με την εξόρυξη γνώσης από σχεσιακές βάσεις δεδομένων, χωροχρονικές βάσεις δεδομένων, κ.λπ.

ii) Απόδοση και εξελιξιμότητα των αλγορίθμων ΕΔ

Για να έχουμε αποτελεσματική εξόρυξη γνώσης από μεγάλα σύνολα δεδομένων, πρέπει να έχουμε αλγορίθμους κατάλληλα προσαρμοσμένους σε αυτά. Επομένως, ο χρόνος εκτέλεσης των αλγορίθμων πρέπει να είναι αποδεκτός και αναμενόμενος για μεγάλες βάσεις δεδομένων.

Να σημειώσουμε εδώ ότι αλγόριθμοι με εκθετική ή πολυωνυμική πολυπλοκότητα δεν θεωρούνται πρακτικοί στη χρήση.

iii) Χρησιμότητα, βεβαιότητα, εκφραστικότητα των αποτελεσμάτων της ΕΔ

Η εξορυγμένη γνώση πρέπει να παρουσιάζει με ακριβή τρόπο τα περιεχόμενα των βάσεων δεδομένων και να είναι χρήσιμη για συγκεκριμένες εφαρμογές. Η ακρίβεια των αποτελεσμάτων θα μπορούσε να εκφραστεί μέσω κάποιων μέτρων βεβαιότητας, προσεγγιστικά ή ποσοτικά.

Εξαιρέσεις όπως θόρυβος και outliers πρέπει να αντιμετωπιστούν από τα συστήματα ΕΔ. Το γεγονός αυτό δίνει το κίνητρο για μια συστηματική μελέτη της ποιότητας της εξορυγμένης γνώσης, κατασκευάζοντας στατιστικά ή αναλυτικά μοντέλα, μοντέλα προσομοίωσης, καθώς και τα εργαλεία αυτών.

iv) Εκφράσεις διαφορετικού τύπου για τα αποτελέσματα

Όπως μπορούμε να φανταστούμε, από μεγάλα σύνολα δεδομένων μπορούν να προκύψουν διαφορετικοί τύποι γνώσεων. Επίσης, θα ήταν πολύ χρήσιμο να μπορούμε να ελέγξουμε τη γνώμη από διαφορετικές απόψεις και να την εκφράσουμε σε διάφορες μορφές.

Θεωρείται ότι θα ήταν πολύ καλό να μπορούν να εκφραστούν τα ερωτήματα της ΕΔ και η εξορυμένη γνώση σε γλώσσες υψηλού επιπέδου ή μέσω γραφικών διεπαφών των χρηστών. Έτσι, η ΕΔ θα μπορούσε να είναι εφαρμόσιμη και από μη ειδικούς και η εξορυμένη γνώση θα χρησιμοποιούταν άμεσα από όλους.

Τέλος, απαιτείται το σύστημα να υιοθετήσει εκφραστικές τεχνικές αναπαράστασης της γνώσης, έτσι ώστε να επιτευχθεί η αποτελεσματική παρουσίαση της γνώσης.

v) **Διαλογική ανακάλυψη γνώσης στα πολλαπλά εννοιολογικά επίπεδα**

Είναι δύσκολο να προβλεφθεί αυτό που θα μπορούσε να ανακαλυφθεί επακριβώς από μια βάση δεδομένων. Γι'αυτό, θα μπορούσε να καθοριστεί μια σειρά ερωτήσεων της ΕΔ προκειμένου να διαμορφωθεί η εστίαση στα δεδομένα, να δημιουργηθεί ένα λεπτομερέστερο επίπεδο ΕΔ και να παρατηρηθούν τα αποτελέσματα της ΕΔ σε πολλαπλά επίπεδα και από διαφορετικές πτυχές. Όλα αυτά μπορούν να επιτευχθούν μέσω της διαλογικής ανακάλυψης της γνώσης.

vi) **Εξόρυξη πληροφορίας από διαφορετικές πηγές δεδομένων**

Σε σχέση με τη σύνδεση των διάφορων πηγών δεδομένων, υπάρχει προβάδισμα της ευρέως διαθέσιμης σύνδεσης υπολογιστών σε τοπικό και ευρύτερο δίκτυο, συμπεριλαμβανομένου του διαδικτύου. Αυτό οδηγεί στη δημιουργία μεγάλων κατανεμημένων και ετερογενών βάσεων δεδομένων.

Επιπλέον, το τεράστιο μέγεθος των βάσεων δεδομένων, η υψηλή κατανομή των δεδομένων και η υπολογιστική πολυπλοκότητα ορισμένων μεθόδων ΕΔ οδηγούν στην ανάπτυξη παράλληλων και κατανεμημένων αλγορίθμων ΕΔ.

vii) **Προστασία ιδιωτικότητας και ασφάλεια δεδομένων**

Η προστασία και αποκλειστικότητα των δεδομένων απειλείται στην περίπτωση που αυτά μπορούν να παρατηρηθούν από πολλές διαφορετικές σκοπιές. Είναι σημαντικό να μελετήσουμε πότε μπορεί να οδηγηθούμε σε μια εισβολή στην ιδιωτικότητα μέσω της KDD και τι μέτρα ασφαλείας μπορούν να αναπτυχθούν για να εμποδιστεί η αποκάλυψη των ευαίσθητων πληροφοριών.

Να σημειώσουμε ότι μερικές από τις απαιτήσεις που αναφέραμε παραπάνω μπορεί να φέρουν αντικρουόμενους στόχους. Για παράδειγμα, ο στόχος της προστασίας της ασφάλειας δεδομένων μπορεί να αντικρούει στην απαίτηση για διαλογική εξόρυξη πολυεπίπεδης γνώσης από διαφορετικές σκοπιές.

Η παρουσίαση των απαιτήσεων αυτών γίνεται στα πλαίσια του ενδιαφέροντός μας για την ανάπτυξη αποτελεσματικών και εξελίξιμων αλγορίθμων. Για το λόγο αυτό, έγιναν συγκεκριμένες ομαδοποιήσεις των απαιτήσεων ώστε να γίνει μια γενική απεικόνιση.

1.5 Ταξινόμηση συστημάτων και μεθόδων

Τα τελευταία έτη έχει γίνει μεγάλη πρόοδος στην έρευνα και ανάπτυξη της ΕΔ. Οι μέθοδοι και τα συστήματα ΕΔ που έχουν αναπτυχθεί μπορούν να κατηγοριοποιηθούν με διάφορα κριτήρια. Για παράδειγμα, μπορεί να γίνει κατηγοριοποίηση των μεθόδων με βάση τους τύπους βάσεων δεδομένων που θα χρησιμοποιηθούν, τους τύπους γνώσης που θα εξαχθούν και τις τεχνικές που θα εφαρμοστούν.

Ένας ενδιαφέρων τρόπος ταξινόμησης των συστημάτων και μεθόδων που μας αποσχολούν θα ήταν μέσω της διαμόρφωσης συγκεκριμένων ερωτημάτων. Τα ερωτήματα αυτά θα μπορούσαν να αποτελούν τα κριτήρια με βάση τα οποία θα γινόταν ο διαχωρισμός.

Σύμφωνα με τους Chen et al. (1996), υπάρχουν τρία κριτήρια (ερωτήματα) στα οποία στηριζόμαστε ώστε να ταξινομήσουμε τα συστήματα ΕΔ. Τα κριτήρια αυτά είναι:

- **Τι είδους βάση δεδομένων χρησιμοποιείται;**
  Ένα σύστημα ΕΔ θα μπορούσε να ταξινομηθεί σύμφωνα με τα είδη ΒΔ στις οποίες εφαρμόζεται η ΕΔ. Για παράδειγμα, ένα σύστημα που χρησιμοποιείται για την εξαγωγή γνώσης από σχεσιακά δεδομένα καλείται σχεσιακό σύστημα γνώσης. Εάν εξάγει τη γνώση από αντικειμενοστραφείς ΒΔ καλείται αντικειμενοστραφές σύστημα ΕΔ. Γενικά, ένα σύστημα ΕΔ θα μπορούσε να ταξινομηθεί βασισμένο στους διάφορους τύπους συστημάτων βάσεων δεδομένων, όπως τα σχεσικά ή αντικειμενοστραφή συστήματα βάσεων δεδομένων, οι χωροχρονικές βάσεις, τα συστήματα βάσεων δεδομένων πολυμέσων κ.λπ.

- **Τι είδους γνώση εξάγεται;**
  Από ένα σύστημα ΕΔ μπορούν να ανακαλυφθούν διάφοροι τύποι γνώσης, όπως **κανόνες συνάφειας** (*association rules*), **συσταδοποίηση** (*clustering*), **κανόνες ταξινόμησης** (*classification rules*), **διαχωριστικοί κανόνες** (*discriminant rules*), θεωρία εξέλιξης (*evolution*) και **ανάλυση απόκλισης** (*deviation analysis*).
  Επιπλέον, ένα σύστημα ΕΔ θα μπορούσε να ταξινομηθεί σύμφωνα με το επίπεδο γενίκευσης της εξαγόμενης γνώσης, η οποία θα μπορούσε να είναι γενική, πρώτου επιπέδου ή πολυεπίπεδη γνώση.

- **Ποιο είδος τεχνικών χρησιμοποιείται;**

  Τα συστήματα ΕΔ θα μπορούσαν να ταξινομηθούν και ανάλογα με τις χρησιμοποιούμενες τεχνικές εξόρυξης δεδομένων. Για παράδειγμα, θα μπορούσαν να ταξινομηθούν σε αυτόνομα συστήματα, συστήματα προσανατολισμένα στα δεδομένα, συστήματα οδηγούμενα από ερωταποκρίσεις καθώς και διαλογικά συστήματα.

  Επίσης, σύμφωνα με την χρησιμοποιούμενη προσέγγιση, τα συστήματα ΕΔ θα μπορούσαν να ταξινομηθούν σε συστήματα γενικής εξόρυξης, εξόρυξης βασισμένης στα πρότυπα, στη στατιστική ή τα μαθηματικά κ.λπ.

1.6 Αντικείμενο εργασίας

Στα πλαίσια της εργασίας αυτής, θα ασχοληθούμε με μεθόδους και αλγορίθμους της ΕΔ που αφορούν στα κατηγορικά ή μικτά δεδομένα. Ως κατηγορικά εννοούμε τα δεδομένα των οποίων οι μεταβλητές είναι δίτιμες ή προέρχονται από διακριτή κατανομή, καθώς επίσης και δεδομένα των οποίων οι μεταβλητές θεωρούνται ή εμφανίζονται ως ονοματικές (*nominal*), διατάξιμες (*ordinal*) ή διαστηματικές (*interval*). Στα μικτά δεδομένα περιλαμβάνονται και συνεχείς μεταβλητές. Ένα σύνηθες φαινόμενο είναι να μετασχηματίζονται συνεχείς μεταβλητές, έτσι ώστε να αντιμετωπιστούν ως κατηγορικές, για να έχουμε πιο συγκροτημένη ανάλυση.

Οι τεχνικές και οι αλγόριθμοι της ΕΔ που αφορούν στα κατηγορικά δεδομένα είναι λίγες σχετικά με τη χρησιμότητα αυτού του τύπου δεδομένων. Υπάρχουν πολλά παραδείγματα αναλύσεων όπου όλες οι μεταβλητές της βάσης δεδομένων θεωρούνται κατηγορικές. Παραδείγματος χάριν, ας σκεφτούμε μια βάση δεδομένων όπου καταγράφονται οι αιτήσεις δανείων από μια τράπεζα. Εκτενέστερα σχόλια για όλα αυτά θα γίνουν στα επόμενα κεφάλαια.

Στόχος μας είναι η συγκέντρωση του υπάρχοντος υλικού της ΕΔ από πλευράς αλγορίθμων αναγνώρισης προτύπων, προ-επεξεργασίας δεδομένων και συσταδοποίησης, καθώς και η παρουσίαση του σχετικού διαθέσιμου λογισμικού. Η πραγματοποίηση χαρακτηριστικών εφαρμογών κρίνεται απαραίτητη, ώστε να γίνει περισσότερο κατανοητή η χρησιμότητα των αλγορίθμων.

Έτσι, ξεκινάμε με την παρουσίαση της διαδικασίας ανακάλυψης της γνώσης στο κεφάλαιο που ακολουθεί. Επίσης, γίνεται διαχωρισμός μεταξύ των μεθόδων ΕΔ, καθώς αποτελούν μέρος της αλυσίδας ανακάλυψης γνώσης. Στο τρίτο κεφάλαιο, γίνεται σύγκριση μεταξύ ΕΔ

και Στατιστικής, ενώ επιχειρούμε να συνδέσουμε την εφαρμογή της ΕΔ με τα κατηγορικά δεδομένα.

Αντικείμενο του τέταρτου κεφαλαίου είναι η παρουσίαση των μεθόδων προ-επεξεργασίας δεδομένων, ένα σημαντικό στάδιο πριν από αυτό της ΕΔ στην αλυσίδα ανακάλυψης της γνώσης. Το πέμπτο κεφάλαιο αφορά στη συσταδοποίηση, η οποία αποτελεί έναν από τους τρόπους χειρισμού μιας μεγάλης βάσης δεδομένων, όταν θέλουμε να περιγράψουμε τις σχέσεις μεταξύ των αντικειμένων (*objects*) της.

Η συσταδοποίηση είναι μια από τις μεθόδους ΕΔ που αναφέρθηκαν στο δεύτερο κεφάλαιο. Ο λόγος που επικεντρωθήκαμε σε αυτή τη μέθοδο ΕΔ στο πέμπτο κεφάλαιο είναι ότι αποτελεί ένα ιδιαίτερα σημαντικό και εξελισσόμενο μέρος της ΕΔ, στην περίπτωση που θέλουμε να περιγράψουμε τα δεδομένα και να ανακαλύψουμε πρότυπα (*patterns*). Η αναφορά μας στον τομέα αυτό γίνεται στα πλαίσια αναζήτησης τεχνικών συσταδοποίησης κατηγορικών δεδομένων. Το έκτο κεφάλαιο είναι ο πυρήνας της εργασίας μας, αφού εκεί καταγράφουμε τους αλγορίθμους συσταδοποίησης για κατηγορικά και μικτά δεδομένα.

Κλείνοντας, στο έβδομο κεφάλαιο παρουσιάζεται το υπάρχον λογισμικό ΕΔ και οι εφαρμογές που πραγματοποιήσαμε στα σχετικά προγράμματα και πλατφόρμες. Το πρακτικό κομμάτι αφορά σε εφαρμογές πάνω σε κατηγορικά δεδομένα. Πρόκειται για την ανάπτυξη ενός μοντέλου ΕΔ με στόχο τη διακράτηση των πελατών της Στεγαστικής Πίστης μεγάλης τράπεζας. Οι τεχνικές συσταδοποίησης βοηθούν στα πλαίσια εύρεσης των ομάδων πελατών που είναι πιθανό να αποχωρήσουν από την Τράπεζα. Έτσι, μπορούμε να βρούμε ποια είναι τα χαρακτηριστικά αυτών των ατόμων και ποια η πιθανότητα να αποχωρήσουν.

Η εφαρμογή που παρουσιάζεται στο τελευταίο κεφάλαιο έχει γίνει σε πραγματικά δεδομένα, στα πλαίσια συνεργασίας με τραπεζικό όμιλο, στην επιχειρηματική μονάδα στεγαστικής πίστης. Το πραγματικό σύνολο δεδομένων περιείχε 120.000 εγκεκριμένες αιτήσεις για στεγαστικό δάνειο και αναζητούσαμε τις ομάδες ατόμων που αναμένεται να αποπληρώσουν το στεγαστικό τους δάνειο και να αποχωρήσουν από την τράπεζα, μεταφέροντας το χαρτοφυλάκιό τους. Όμως, για τις ανάγκες της εργασίας αυτής και για λόγους διαφύλαξης προσωπικών δεδομένων, παραθέτουμε τα αποτελέσματα από ένα υποθετικό αρχείο μεγέθους 3.000 παρατηρήσεων.

ΚΕΦΑΛΑΙΟ 2

Η ανακάλυψη της γνώσης

2.1 Η διαδικασία της ανακάλυψης γνώσης από βάσεις δεδομένων

Η ανακάλυψη της γνώσης από βάσεις δεδομένων (*Knowledge Discovery in Databases – KDD*) είναι μία αυτοματοποιημένη διαδικασία, μέσω της οποίας γίνεται προσπάθεια διερευνητικής ανάλυσης και μοντελοποίησης τεράστιων αποθηκών δεδομένων. Πρόκειται για μια συγκροτημένη μεθοδολογία αναγνώρισης έγκυρων και πρωτότυπων προτύπων μέσα από πολύ μεγάλους και περίπλοκους πίνακες δεδομένων.

Στόχος της όλης εφαρμογής είναι τα πρότυπα που θα προκύψουν να είναι χρήσιμα και κατανοητά. Ένας γενικός ορισμός της KDD ερμηνεύει με σαφήνεια τον όρο αυτό (Fayyad et al., 1996-a) είναι:

> «KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα».

Στις ακόλουθες υποενότητες, θα επιχειρήσουμε να αναλύσουμε τον ορισμό αυτό, στηριζόμενοι στις λέξεις-κλειδιά μιας μικρής φράσης, η οποία όμως περιέχει αρκετές πληροφορίες και άλλες έννοιες. Επίσης, θα δώσουμε το πλαίσιο εφαρμογών και θα περιγράψουμε τα βήματα της διαδικασίας KDD, τα οποία θα χωρίσουμε και σε σχετικούς τομείς.

2.1.1 Ανάλυση ορισμού

Με βάση τους δημιουργούς του παραπάνω ορισμού, παραθέτουμε την ανάλυση των βασικών εννοιών που αναφέρονται σε αυτόν (Fayyad et al., 1996-a):

- **Δεδομένα**

Περιγράφουν **οντότητες** ή **συσχετίσεις** του πραγματικού κόσμου. Για παράδειγμα, ένα σύνολο εγγραφών που αναφέρονται στην έκδοση δανείων μιας τράπεζας και περιλαμβάνονται ιδιότητες όπως εισόδημα, οικογενειακή κατάσταση, άληκτο κεφάλαιο. Σύμφωνα με το λεξικό Merriam-Webster's Dictionary, η λέξη «δεδομένα» (*data*) σημαίνει:

> «Πραγματικές πληροφορίες που προκύπτουν από καταγραφή, μέτρηση ή στατιστική και αποτελούν τη βάση υπολογισμών ή επιχειρηματολογίας».

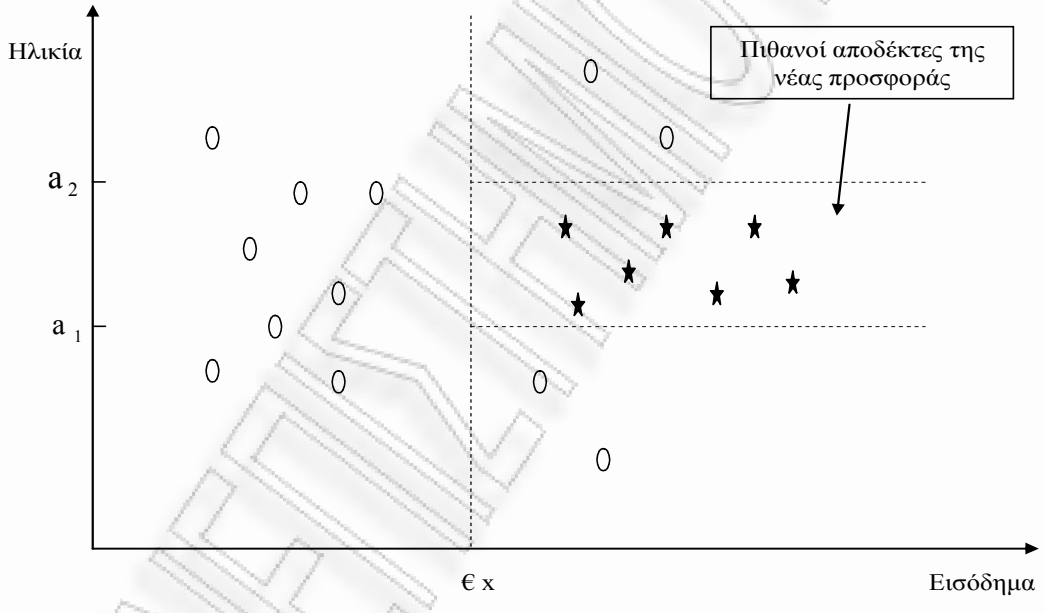
- **Πρότυπο**

Είναι μια έκφραση E σε μια γλώσσα L η οποία περιγράφει ένα υποσύνολο δεδομένων $F_E \subseteq F$ εκμεταλλευόμενο κοινές ιδιότητες των δεδομένων του. Σε αυτή την περίπτωση το pattern θεωρείται υποσύνολο του F και αφαίρεση του F.

Για παράδειγμα, έστω ο κανόνας: «Εάν οι δανειολήπτες έχουν εισόδημα > € x $\wedge$ age$[a_1, a_2]$, δηλαδή εισόδημα πάνω από μια τιμή x ευρώ και ηλικία μεταξύ του διαστήματος $[a_1, a_2]$, τότε ανταποκρίνονται στη νέα προσφορά συμπληρωματικού δανείου». Το Σχήμα που ακολουθεί απεικονίζει γραφικό αυτό το σχέδιο.

ΣΧΗΜΑ 2.1

Γραφική απεικόνιση ενός προτύπου



- **Διαδικασία**

Είναι μια διαδικασία πολλών βημάτων, η οποία περιλαμβάνει την **προ-επεξεργασία** των δεδομένων, την **αναζήτηση** των προτύπων και την **αξιολόγηση** της εξαγόμενης γνώσης.

- **Εγκυρότητα**

Ένα από τα βασικά προβλήματα και αντικείμενο έρευνας στην ΕΔ. Το εξαγόμενο πρότυπο θα πρέπει να είναι **συνεπές** σε νέα δεδομένα με κάποιον βαθμό βεβαιότητας.

- **Πιθανά χρήσιμο**

Η εξαγωγή των προτύπων θα πρέπει να ακολουθείται από μερικές χρήσιμες διεργασίες όπως η **αξιολόγησή** τους από κάποιες συναρτήσεις χρησιμότητας και η **διατήρηση** όσο το δυνατόν περισσότερης γνώσης από τα αρχικά δεδομένα.

Η διατηρηθείσα γνώση μπορεί να συμβάλλει στη λήψη αποφάσεων. Για παράδειγμα, εάν θεωρήσουμε τη βάση δεδομένων που καταγράφει τα δάνεια, θα ήταν χρήσιμο να υπάρχει μια ένδειξη της αναμενόμενης αύξησης στα κέρδη ή ένας κανόνας απόφασης όπως: «**Εάν** έσοδα < € x , **τότε** ο πελάτης δε μπορεί να πάρει δάνειο».

- **Τελικά κατανοητό**

Στόχος της Εξόρυξης Δεδομένων (ΕΔ) είναι να προσδιοριστούν τα πρότυπα και να είναι κατανοητά και από τους μη ειδικούς, ώστε να οδηγηθούν σε σημαντικά συμπεράσματα και αποφάσεις.

Είναι γεγονός ότι, όπως συνέβη με την ΕΔ, προτάθηκαν αρκετές ονομασίες και για τη διαδικασία ανακάλυψης γνώσης, όπως εξαγωγή γνώσης (*knowledge extraction*), ανακάλυψη πληροφορίας (*information discovery*) ή μη επιβλέπουσα αναγνώριση προτύπων (*unsupervised pattern recognition*). Επίσης, πολλές φορές, υπάρχει μια σύγχηση στη χρήση των όρων KDD και ΕΔ.

Για να ξεκαθαρίσουμε τις δύο έννοιες, αναφέρουμε το σχολιασμό που γίνεται από τους Fayyad et al. (1996-c), με βάση τους οποίους η KDD είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων στα δεδομένα, ενώ η ΕΔ είναι η χρήση αλγορίθμων με στόχο την εξαγωγή πληροφοριών και προτύπων που παράγονται από τη διαδικασία KDD.

Στην υποενότητα που ακολουθεί, δίνουμε το πλαίσιο εφαρμογών της KDD αλλά και της ΕΔ, έχοντας ως σκοπό να ανακαλύψουμε τον τρόπο συνεργασίας αυτών των δύο εννοιών.

2.1.2 Χρησιμότητα και εφαρμογές στον πραγματικό κόσμο

Μερικές από τις εφαρμογές της ΕΔ, στα πλαίσια ανακάλυψης της γνώσης, (Bramer, 2007) είναι:

- ✓ Ανάλυση οργανικών συνθέσεων (*analysis of organic compounds*)
- ✓ Αυτόματη αφαίρεση (*automatic abstracting*)
- ✓ Προσδιορισμός απειλών στον κλάδο των πιστώσεων (*fraud detection*)

- ✓ Πρόβλεψη κατανάλωσης ενέργειας
- ✓ Οικονομική πρόβλεψη
- ✓ Ιατρική διάγνωση
- ✓ Πρόβλεψη τηλεθέασης
- ✓ Σχεδιασμός παραγωγής
- ✓ Εκτίμηση ακινήτων
- ✓ Πώληση προς συγκεκριμένους «στόχους» (*Targeted marketing*)
- ✓ Ανάλυση κινδύνου από τοξικά (*toxic hazard analysis*)
- ✓ Βελτιστοποίηση παροχής θερμότητας στα φυτά (*thermal power plant optimization*)
- ✓ Πρόβλεψη καιρού

Αυτές ήταν μερικές από τις ομάδες εφαρμογών της ΕΔ, ενώ μπορούμε να ανακαλύψουμε και πολλές άλλες, αν αναλογιστούμε καθημερινά πρακτικά ζητήματα. Στόχος αυτής της υποενότητας είναι να συνειδητοποιήσουμε ότι η ΕΔ αποτελεί το **εργαλείο** της KDD. Για παράδειγμα, οι Fayyad et al. (1996-b) αναφέρουν ως εφαρμογές της KDD στον χώρο των επιχειρήσεων τις δραστηριότητες σε:

- ✓ Marketing
- ✓ Επενδύσεις
- ✓ Προσδιορισμό απειλών (*fraud detection*)
- ✓ Βιομηχανική παραγωγή
- ✓ Τηλεπικοινωνίες
- ✓ Καθαρισμό δεδομένων (*data cleaning*)

Προφανώς, η δράση της KDD σε αυτούς τους τομείς γίνεται μέσω της ΕΔ. Η ανάπτυξη της KDD είναι συνεχής και οφείλεται στη συνεργασία των ερευνητικών πεδίων που αναφέραμε στο πρώτο κεφάλαιο ως «απαρχές της ΕΔ». Το ερώτημα λοιπόν είναι σε τι διαφέρει η KDD από την αναγνώρση προτύπων, τη μηχανική εκμάθηση ή τα άλλα πεδία που συνεισφέρουν στην ΕΔ;

Η απάντηση δίνεται από τους Fayyad et al. (1996-b). Τα ανωτέρω πεδία προσφέρουν κάποιες από τις μεθόδους ΕΔ που χρησιμοποιούνται στο στάδιο της ΕΔ, εάν θεωρηθεί ως ένα μέρος της διαδικασίας KDD. Πράγματι, η διαδικασία KDD θεωρείται μια σειρά βημάτων, μερικά από τα οποία αποτελούν το μέρος της ΕΔ. Μάλιστα, το μέρος αυτό θεωρείται από τα

εύκολα κομμάτια της διαδικασίας KDD, ενώ το κομμάτι της προεπεξεργασίας των δεδομένων (καθαρισμός, μετασχηματισμός κ.λπ.), θεωρείται πιο περίπλοκο. Στην ακόλουθη υποενότητα παρουσιάζουμε αναλυτικά τα βήματα της διαδικασίας KDD.

Στην ουσία, στόχος της KDD είναι η ανακάλυψη γνώσης από τα δεδομένα, περιλαμβάνοντας όμως και διαδικασίες ελέγχου του τρόπου αποθήκευσης ή πρόσβασης των δεδομένων, όπως και της δυνατότητας επέκτασης των αλγορίθμων σε πολύ μεγάλα σύνολα δεδομένων, χωρίς να μειωθεί η αποτελεσματικότητά τους.

Επίσης, η KDD ενδιαφέρεται για τον τρόπο απεικόνισης των αποτελεσμάτων, αλλά και για τη γενικότερη συνεργασία με το ανθρώπινο στοιχείο, ώστε να συμβάλλει κατά τον καλύτερο δυνατό τρόπο. Άρα, οι εφαρμογές της KDD δεν έχουν να κάνουν μόνο με τη δράση και συμπεριφορά των αλγορίθμων. Αυτό αποτελεί καθαρά κομμάτι της ΕΔ. Βέβαια, όταν έχουμε να ασχοληθούμε με στοιχεία από τον πραγματικό κόσμο, όλες αυτές οι ιδιότητες ενώνονται, με τελικό στόχο την διάκριση χρήσιμων προτύπων από τα δεδομένα.

Οι διαδικασίες της KDD διευκολύνονται κατά πολύ μέσω των τεχνολογιών των βάσεων δεδομένων, καθώς και πεδίων όπως η **αποθήκευση δεδομένων** (*data warehousing*). Οι τεχνολογίες της αποθήκευσης δεδομένων εξυπηρετούν την KDD σε δύο πολύ σημαντικά σημεία: τον **καθαρισμό** και την **πρόσβαση** των δεδομένων (Fayyad et al., 1996-b). Επίσης, μια διάσημη προσέγγιση για την ανάλυση αποθηκών δεδομένων (*data warehouses*) είναι η **OLAP** (*online analytical processing*).

Τα εγαλεία OLAP συμβάλλουν στην πολυμεταβλητή ανάλυση δεδομένων και στοχεύουν στην απλοποίηση και την υποστήριξη της αλληλεπίδραστικής ανάλυσης δεδομένων (*interactive data analysis*). Βέβαια, η KDD και τα εργαλεία της είναι ένα βήμα πέρα από ότι υποστηρίζεται από τα περισσότερα τυπικά συστήματα βάσεων δεδομένων (Fayyad et al., 1996-b / Maimon and Rokach, 2005).

Περισσότερα για τα εγαλεία OLAP και τις άλλες έννοιες που συμβάλλουν στην ΕΔ και κατά συνέπεια στην KDD δίνονται στο βιβλίο της Dunham (2003). Ενδεικτικά αναφέρουμε τα συστήματα **DBMS** (*database management system*), τις διαδικασίες της **ασαφούς λογικής** (*fuzzy logic*), την **μοντελοποίηση διαστάσεων** (*dimensional modeling*), την **ευρετηριοποίηση** (*indexing*) κ.λπ.

2.1.3 Βήματα της διαδικασίας KDD

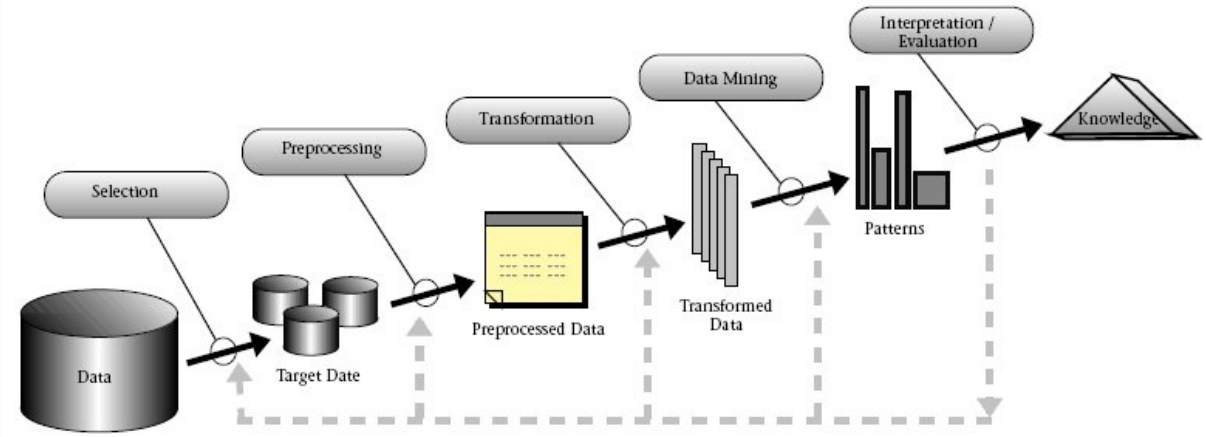
Η διαδικασία της KDD είναι μια διαλογική και επαναληπτική διαδικασία. Δηλαδή, μπορεί να απαιτηθεί η επιστροφή σε ένα προηγούμενο βήμα, όπως φαίνεται και στο σχήμα 2.2. Ως πρώτο βήμα θεωρούμε τον εντοπισμό των στόχων της KDD, ενώ στο τέλος αξιοποιούμε την ανακαλυφθείσα γνώση.

Στο Σχήμα 2.2, που ακολουθεί, παρατηρούμε τα βήματα της διαδικασίας KDD, υπό μορφή αλυσίδας, έτσι όπως θεωρούνται από τους Fayyad et al. (1996-b).

ΣΧΗΜΑ 2.2

Η διαδικασία ανακάλυψης της γνώσης

[Πηγή: Fayyad et al., 1996-b]



Με βάση τη σχετική βιβλιογραφία (Maimon and Rokach, 2005 / Fayyad et al., 1996-a), θα διαχωρίσουμε τη διαδικασία KDD σε εννέα βήματα, τα οποία συνοπτικά είναι τα ακόλουθα:

1ο ΒΗΜΑ: Ανάπτυξη και κατανόηση της περιοχής της εφαρμογής

Σε αυτό το προκαταρκτικό στάδιο γίνεται προετοιμασία για την κατανόηση του πλαισίου δράσης. Πρέπει να γίνει σαφές, δηλαδή, ποιες αποφάσεις θα ληφθούν σχετικά με μετασχηματισμούς, αλγορίθμους, αναπαράσταση κ.λπ.

Το Βήμα αυτό βοηθά στην κατανόηση των στόχων από τον τελικό χρήστη, καθώς και στην εύρεση του περιβάλλοντος όπου θα δράσει η διαδικασία ανακάλυψης της γνώσης. Στα πλαίσια αυτά περιλαμβάνεται και η προγενέστερη γνώση του υπό εξέταση τομέα. Είναι πιθανό να απαιτηθεί επανάληψη αυτού του Βήματος στην πορεία.

2ο ΒΗΜΑ: Επιλογή και δημιουργία ενός κατάλληλου συνόλου δεδομένων

Έχοντας ορίσει τους στόχους, θα έπρεπε να έχουν προσδιοριστεί και τα δεδομένα που θα χρησιμοποιηθούν. Το Βήμα αυτό περιλαμβάνει τον εντοπισμό των δεδομένων που είναι διαθέσιμα, την απόκτηση επιπρόσθετων αναγκαίων δεδομένων και την ενσωμάτωση όλων αυτών σε ένα σύνολο δεδομένων το οποίο θα περιλαμβάνει τα χαρακτηριστικά (*attributes*) που θα ληφθούν υπόψη.

Το Βήμα αυτό είναι πολύ σημαντικό, καθώς η ΕΔ μαθαίνει και ανακαλύπτει από τα δεδομένα που έχει εκείνη τη στιγμή στη διάθεσή της. Σε αυτή τη βάση κατασκευάζονται και τα μοντέλα. Είναι πιθανό, όμως, να προκύψουν προβλήματα στην περίπτωση όπου λείπουν χαρακτηριστικά από κάποιες παρατηρήσεις, καθώς μπορεί να δημιουργηθούν σφάλματα στη μελέτη. Άρα, χρειάζεται η μέγιστη δυνατή συλλογή χαρακτηριστικών.

Από την άλλη πλευρά, όμως, αυτή η ανάγκη ανεβάζει το κόστος διεξαγωγής της ανάλυσης. Για το λόγο αυτό, η διαδικασία της KDD αναλαμβάνει να αξιοποιήσει αρχικά το βέλτιστο διαθέσιμο σύνολο δεδομένων και στη συνέχεια επεκτείνεται και παρατηρεί τα αποτελέσματα στα πλαίσια της ανακάλυψης γνώσης και μοντελοποίησης.

3ο ΒΗΜΑ: Προ-επεξεργασία και καθαρισμός δεδομένων

Ένα πολύ σημαντικό σημείο που μας αποσχολεί είναι η αξιοπιστία των δεδομένων, η οποία μελετάται μέσα από αυτό το απαραίτητο Βήμα της διαδικασίας. Στα πλαίσια της αναζήτησης ενός αξιόπιστου συνόλου δεδομένων, οφείλουμε να πραγματοποιήσουμε καθαρισμό δεδομένων (*data cleaning*).

Με τη χρήση του όρου αυτού εννοούμε τη διαχείριση ελλειπουσών τιμών (*missing values*) και την απομάκρυνση θορύβου (*noise*) ή έκτροπων παρατηρήσεων (*outliers*). Οι διαδικασίες καθαρισμού των δεδομένων μπορούν να επιτευχθούν μέσω σύνθετων στατιστικών μεθόδων ή χρησιμοποιώντας έναν αλγόριθμο ΕΔ.

4ο ΒΗΜΑ: Μετασχηματισμός δεδομένων

Μέσω αυτού του βήματος, τα δεδομένα μετασχηματίζονται ή παγιώνονται σε μορφές κατάλληλες για εξόρυξη. Για το σκοπό αυτό εφαρμόζωνται μέθοδοι μείωσης διαστάσεων (επιλογή χαρακτηριστικού, εξαγωγή και καταγραφή δείγματος) και μετασχηματισμού χαρακτηριστικών (διακριτοποίηση συνεχών μεταβλητών, λειτουργικός μετασχηματισμός).

Αποτέλεσμα των εφαρμογών αυτών είναι η μείωση του αριθμού των υπό εξέταση μεταβλητών ή η εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές.

5ο ΒΗΜΑ: Επιλογή της κατάλληλης μεθόδου εξόρυξης δεδομένων

Ύστερα από όσα Βήματα έχουμε εκτελέσει, είμαστε σε θέση να αποφασίσουμε ποιον τύπο ΕΔ θα χρησιμοποιήσουμε (ταξινόμηση, παλινδρόμηση, συσταδοποίηση). Αυτή η επιλογή βασίζεται περισσότερο στους στόχους της KDD, αλλά και στα Βήματα που έχουν ήδη προηγηθεί.

Όπως έχουμε ήδη αναφέρει και θα σχολιάσουμε και παρακάτω, οι δύο βασικοί στόχοι της ΕΔ είναι η περιγραφή και η πρόβλεψη. Οι τεχνικές ΕΔ βασίζονται στην πλειοψηφία τους στην επαγωγική εκμάθηση (*inductive learning*), όπου κατασκευάζεται ένα σαφές ή εννοούμενο μοντέλο μέσω γενίκευσης ενός επαρκούς αριθμού εκπαιδευτικών παραδειγμάτων (*training examples*).

Βασική προϋπόθεση είναι ότι αυτό το μοντέλο εκπαίδευσης (*trained model*) θα μπορεί να εφαρμοστεί σε μελλοντικές περιπτώσεις. Επίσης, η στρατηγική αυτή λαμβάνει υπόψη την περίπτωση μετα-εκμάθησης (*meta-learning*) για το συγκεκριμένο σύνολο των διαθέσιμων δεδομένων.

6ο ΒΗΜΑ: Επιλογή αλγορίθμου εξόρυξης δεδομένων

Έχοντας ορίσει τη στρατηγική, μπορούμε να επιλέξουμε τον τρόπο επίτευξης του στόχου. Στο στάδιο αυτό εφαρμόζονται ευφυείς μέθοδοι με σκοπό την αναζήτηση ενδιαφέροντων προτύπων γνώσης. Για παράδειγμα, ένας έλεγχος ακρίβειας θα ήταν καλύτερα να γίνει μέσω νευρωνικών δικτύων, ενώ για την κατανόηση της δομής (*understandability*) θα επιλέγονταν τα δέντρα αποφάσεων.

Τα πρότυπα που αναζητώνται θα μπορούσαν να είναι μιας συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου αντιπροσωπεύσεων, όπως κανόνες ταξινόμησης, δέντρα, παλινδρόμηση, συσταδοποίηση κ.λπ. Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα Βήματα.

7ο ΒΗΜΑ: Εκτέλεση αλγορίθμου

Η κάλυψη των προηγούμενων προϋποθέσεων οδηγεί στο επιθυμητό σημείο όπου θα εκτελέσουμε τον επιλεγόμενο αλγόριθμο. Είναι πιθανή η επανάληψη του αλγορίθμου αυτού για αρκετές φορές μέχρι να προκύψει ικανοποιητικό αποτέλεσμα.

8ο ΒΗΜΑ: Αξιολόγηση

Σε αυτό το στάδιο γίνεται εκτίμηση και ερμηνεία των εξορυχθέντων προτύπων (κανόνες, αξιοπιστία κ.λπ.), λαμβάνοντας υπόψη τους στόχους που είχαν τεθεί στο πρώτο Βήμα. Επίσης, παρατηρούμε την επίδραση των Βημάτων 2, 3 και 4 (προεπεξεργασία δεδομένων) στον αλγόριθμο ΕΔ που έχει επιλεγεί μέσα από τα Βήματα 5, 6 και 7 (εξόρυξη δεδομένων). Για παράδειγμα, μπορεί να κριθεί αναγκαία η προσθήκη χαρακτηριστικών (μεταβλητών) στο βήμα 4, ώστε να επαναληφθεί η εφαρμογή της αλυσίδας KDD από εκεί.

Το Βήμα της αξιολόγησης επικεντρώνεται στην κρίση εάν το προκύπτων μοντέλο είναι κατανοητό και χρήσιμο, καθώς και στην επιλογή των πιο ενδιαφέροντων εξαγόμενων προτύπων. Επιπλέον, στο Βήμα αυτό, τεκμηριώνεται η ανακαλυφθείσα γνώση και είναι πλέον διαθέσιμη για περαιτέρω χρήση.

9ο ΒΗΜΑ: Παρουσίαση και χρήση της ανακαλυφθείσας γνώσης

Στο τελευταίο Βήμα, η εξορυγμένη γνώση ενσωματώνεται στο σύστημα για περαιτέρω δράση (πραγματοποίηση αλλαγών στο σύστημα, μέτρηση επιδράσεων). Η επιτυχία αυτού του Βήματος αποδεικνύει την αποτελεσματικότητα χρήσης της αλυσίδας KDD.

Επιπλέον, μέσα από αυτό το Βήμα γίνεται έλεγχος για επίλυση τυχών συγκρούσεων με προηγούμενη εξορυγμένη γνώση. Είναι πιθανό να αλλάξουν ορισμένες δομές δεδομένων, καθώς κάποιες μεταβλητές μπορεί να μην είναι πλέον διαθέσιμες. Επίσης, μπορεί να αλλάξει η περιοχή δράσης των δεδομένων, καθώς μπορεί να προκύψει για μια μεταβλητή μια τιμή η οποία να μην είχε υποτεθεί πριν.

Όπως παρατηρούμε από την απεικόνηση και την καταγραφή των Βημάτων της KDD, τα Βήματα 2, 3 και 4 ορίζουν τη διαδικασία **προεπεξεργασίας των δεδομένων** (*data preprocessing*). Η διαδικασία αυτή, η οποία αποτελεί ένα απαραίτητο στάδιο πριν την ΕΔ, σχολιάζεται στο τέταρτο κεφάλαιο.

Επίσης, τα Βήματα 5, 6 και 7 αποτελούν στην ουσία τον τομέα της ΕΔ, που μας απασχολεί σε αυτή την εργασία. Είναι σαφές, όμως, ότι πρέπει να γίνει ταξινόμηση μεταξύ των μεθόδων της ΕΔ, ώστε να είμαστε σε θέση να κατανοούμε τη διαδικασία και τις εκάστοτε ανάγκες των βημάτων.

Στην ενότητα που ακολουθεί γίνεται η διάκριση μεταξύ των μεθόδων της ΕΔ, ανάλογα με το ποιο είναι το αποτέλεσμα που θέλουμε να αποκτήσουμε από την ανάλυση των δεδομένων μιας μεγάλης βάσης δεδομένων. Ο διαχωρισμός αυτός είναι πιο εμπεριστατωμένος σε σχέση με την αρχική διάκριση μεταξύ των διαδικασιών της ΕΔ, που έγινε στο πρώτο κεφάλαιο. Στόχος της κατηγοριοποίησης που ακολουθεί είναι να κατανοήσουμε τις ουσιαστικές διαφορές μεταξύ των μεθόδων ΕΔ

2.2 Διαχωρισμός των μεθόδων εξόρυξης δεδομένων

Έχει γίνει ήδη κατανοητό ότι υπάρχουν αρκετές μέθοδοι ΕΔ, οι οποίες χρησιμοποιούνται για διαφορετικούς σκοπούς και μπορούν να καλύψουν άλλους στόχους. Η ποικιλία των μεθόδων είναι τόσο μεγάλη που καθιστά αναγκαία την ταξινόμησή τους σε σχετικές ομάδες, ανάλογα με τους στόχους που μπορεί να καλύψει κάθε ομάδα.

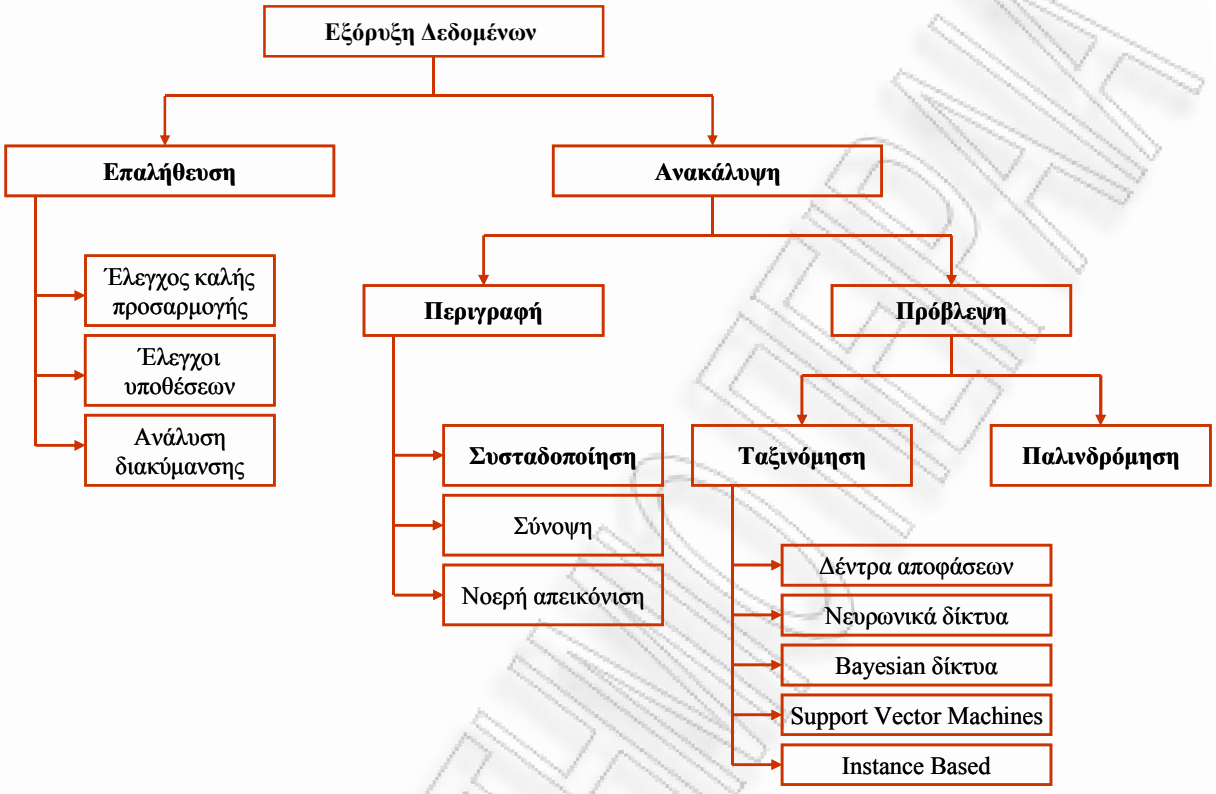
Σε αυτή την ενότητα, θα χωρίσουμε τις μεθόδους της ΕΔ σε τέσσερα βασικά μέρη. Αρχικά, όμως, πρέπει να αναφέρουμε ότι υπάρχουν δύο βασικοί τύποι ΕΔ: η **επαλήθευση** και η **ανακάλυψη**. Στον πρώτο τύπο, γίνεται επαλήθευση των υποθέσεων του χρήστη από το σύστημα. Ενώ, στο δεύτερο τύπο, το σύστημα βρίσκει νέους κανόνες και πρότυπα μέσα από αυτόνομες διαδικασίες.

Οι μέθοδοι ανακάλυψης είναι αυτές που εντοπίζουν αυτόματα πρότυπα στα δεδομένα. Το στάδιο αυτό χωρίζεται στην **περιγραφή** και την **πρόβλεψη** (Maimon and Rokach, 2005). Αυτές είναι οι δύο από τις ομάδες που θα δημιουργήσουμε στη συνέχεια. Όμως, οι μέθοδοι επαλήθευσης ασχολούνται με την εκτίμηση μιας υπόθεσης που προτείνεται από μια εξωτερική πηγή. Αυτές οι μέθοδοι περιλαμβάνουν περισσότερο παραδοσιακές μεθόδους από τη Στατιστική (όπως έλεγχος καλής προσαρμογής, έλεγχος υποθέσεων, ανάλυση διακύμανσης) και σχετίζονται λιγότερο με την ΕΔ απ'ότι οι μέθοδοι ανακάλυψης. Με τη σχέση ΕΔ και Στατιστικής θα ασχοληθούμε στο επόμενο κεφάλαιο.

Στο Σχήμα 2.3, είναι εμφανής η ταξινόμηση μεταξύ των εφορμογών της ΕΔ, ενώ στη συνέχεια δίνουμε τις τέσσερις ομάδες εργασιών της ΕΔ (*data mining tasks*), σύμφωνα με τους Hand et al. (2001), καθώς και τους Tan et al. (2005).

ΣΧΗΜΑ 2.3

Ταξινόμηση της εξόρυξης δεδομένων



2.2.1 Περιγραφική μοντελοποίηση (*Descreptive modeling*)

Ο στόχος ενός μοντέλου περιγραφής είναι να γίνει περιγραφή όλου του συνόλου δεδομένων ή της διαδικασίας που παράγει τα δεδομένα. Ας σκεφτούμε, ως εφαρμογή, περιγραφές που περιλαμβάνουν μοντέλα για την κατανομή πιθανότητας των δεδομένων (εκτίμηση πυκνότητας – *density estimation*), τη διαμέριση ενός χώρου n διαστάσεων σε ομάδες (ανάλυση κατά συστάδες και διαμερισμός – *cluster analysis and segmentation*) ή την περιγραφή των σχέσεων μεταξύ μεταβλητών (εξαρτημένα μοντέλα – *dependency modeling*).

Οι μέθοδοι περιγραφής έχουν ως στόχο την ερμηνεία των δεδομένων και επικεντρώνονται στην κατανόηση του τρόπου που σχετίζονται τα δεδομένα. Αυτό, για παράδειγμα, γίνεται μέσω της νοερής απεικόνισης (*visualization*) ή της σύνοψης (*summarization*), οι οποίες θα μπορούσαμε να πούμε ότι αποτελούν μέρος της Διερευνητικής Ανάλυσης Δεδομένων (*Exploratory Data Analysis – EDA*). Η σημαντικότερη εφαρμογή των περιγραφικών μοντέλων είναι η ανάλυση συστάδων που αναφέραμε και προηγουμένως.

Η **συσταδοποίηση** (*clustering*) επιχειρεί να βρει ομάδες παρατηρήσεων που είναι κοντά μεταξύ τους ως προς τα χαρακτηριστικά που περιλαμβάνουν. Οι μέθοδοι περιγραφής και ειδικά η συσταδοποίηση είναι πολύ χρήσιμες σε πελατοκεντρικά επαγγέλματα που βασίζονται στο CRM (*Customer Relationship Management*), καθώς έτσι μπορούν να εντοπιστούν ομάδες πελατών που αναμένεται να έχουν όμοια συμπεριφορά. Επίσης, μπορούν να βρεθούν οι περιοχές του ωκεανού που έχουν σημαντική επίδραση στο κλίμα της Γης ή να γίνει συμπίεση δεδομένων.

Περισσότερα σχόλια για τη συσταδοποίηση και αλγόριθμοι που χειρίζονται κατηγορικά δεδομένα δίνονται σε επόμενο σχετικό κεφάλαιο. Να σημειώσουμε μόνο, προς το παρόν, ότι ο αριθμός συστάδων που επιλέγεται ως βέλτιστος υπόκειται στην κρίση του ερευνητή, πράγμα το οποίο αντιτίθεται στη θεωρία της Ανάλυσης κατά Συστάδες (*Cluster Analysis*), όπου στόχος είναι η ανακάλυψη «φυσικών» ομάδων στα δεδομένα (όπως για παράδειγμα σε επιστημονικές βάσεις δεδομένων).

2.2.2 Μοντελοποίηση πρόβλεψης (*Predictive modeling*)

Η κατασκευή ενός μοντέλου πρόβλεψης στοχεύει στη δυνατότητα πρόγνωσης της τιμής μιας μεταβλητής (απόκριση) μέσα από τις τιμές άλλων μεταβλητών (επεξηγηματικές) που είναι γνωστές. Εάν η μεταβλητή απόκρισης είναι (ή μπορεί να θεωρηθεί) κατηγορική, τότε είμαστε σε θέση να εφαρμόσουμε μια **μέθοδο ταξινόμησης** (*classification*). Ένα παράδειγμα είναι η πρόβλεψη αγοράς ενός προϊόντος: ναι ή όχι (δίτιμη μεταβλητή). Όμως, αν έχουμε συνεχή απόκριση, τότε προχωράμε σε **παλινδρόμηση** (*regression*). Μια ενδεικτική εφαρμογή είναι η πρόγνωση της μελλοντικής τιμής ενός αποθέματος (εάν αφήσουμε ως συνεχή μεταβλητή την τιμή και δεν την ομαδοποιήσουμε).

Όπως και αν δράσουμε, όμως, κοινός στόχος των δύο εφαρμογών είναι η δημιουργία ενός μοντέλου που ελαχιστοποιεί το σφάλμα στην προβλεφθείσα και τις πραγματικές τιμές. Ο όρος «πρόβλεψη», εδώ, χρησιμοποιείται γενικά ως έννοια και δε θεωρείται απαραίτητα αυστηρή συνέχεια στο χρόνο. Για παράδειγμα, μπορεί να θέλουμε να προβλέψουμε εάν ένας πελάτης τράπεζας θα προπληρώσει το δάνειό του σε συγκεκριμένο μελλοντικό χρονικό διάστημα, αλλά μπορεί και να μας ενδιαφέρει ο προσδιορισμός της διάγνωσης για έναν ασθενή.

Η εξέλιξη της στατιστικής και της μηχανικής μάθησης και η σχετική βιβλιογραφία έχουν δώσει αρκετές μεθόδους πρόβλεψης και μεγάλη πρόοδο στη θεωρία και τη βαθύτερη

κατανόηση. Το στοιχείο για να διακρίνουμε τις διαδικασίες πρόβλεψης από αυτές της περιγραφής είναι ότι αντικειμενικός σκοπός της πρόβλεψης είναι μια συγκεκριμένη μεταβλητή, πράγμα που δε συμβαίνει στην περιγραφή. Εάν έχουμε πολυμεταβλητή απόκριση και επιθυμούμε να πραγματοποιήσουμε μεθόδους ταξινόμησης ή παλινδρόμηση, τότε μπορούμε να αναζητήσουμε τεχνικές boosting (Lutz and Bühlmann, 2006).

2.2.3. Ανάλυση συνάφειας (*Association analysis*)

Η ανάλυση αυτού του τύπου χρησιμοποιείται ώστε να ανακαλυφθούν πρότυπα που περιγράφουν χαρακτηριστικά συνάφειας μεταξύ των δεδομένων. Τα πρότυπα αυτά απεικονίζονται συνήθως στα πλαίσια κανόνων συνεπαγωγής (*implication rules*) ή υποομάδων των χαρακτηριστικών.

Η χαρακτηριστικότερη εφαρμογή και η αιτία από την οποία ξεκίνησαν οι κανόνες συνάφειας (*association rules*) είναι η ανάλυση του «καλαθιού αγοράς» (*market basket analysis*). Ας σκεφτούμε τις υπεραγορές (*super-markets*), όπου οι καταναλωτές γεμίζουν το καλάθι τους με τα προϊόντα της επιλογής τους. Ο όγκος της πληροφορίας που μπορεί να συλλεχθεί είναι τεράστιος. Οι κανόνες συνάφειας αξιοποιούν αυτή την πληροφορία. Για παράδειγμα, ένας κανόνας μπορεί να είναι: «Οι πελάτες που αγοράζουν ψωμί του τοστ, αγοράζουν παράλληλα και αλλαντικά σε ποσοστό 70%».

Στον κανόνα που δώσαμε έχουμε ένα αίτιο (αγορά ψωμιού για τοστ), το οποίο συνδέεται με ένα αποτέλεσμα (αγορά αλλαντικών). Επίσης, δίνεται και εκτίμηση για το πόσο πιθανό να συμβεί αυτή η σχέση αιτίας – αιτιατού. Οι κανόνες συνάφειας καλούνται και κανόνες «if – then».

Άλλες εφαρμογές τους πραγματοποιούνται στην προώθηση προϊόντων, στην τοποθέτηση προϊόντων στα ράφια καταστημάτων, στη διαχείριση αποθεμάτων κ.λπ. Ενδιαφέροντα σχόλια για την ανάλυση συνάφειας γίνονται από τους Hand et al. (2001), Dunham (2003), αλλά και τους Han και Kamber (2001).

2.2.4 Ανίχνευση παρεκτροπών (*Anomaly detection*)

Σε αυτή την ομάδα μεθόδων ανήκουν εργασίες εντοπισμού παρατηρήσεων των οποίων τα χαρακτηριστικά διαφέρουν σημαντικά από αυτά του υπόλοιπου συνόλου δεδομένων. Τέτοιες παρατηρήσεις καλούνται παρέκτροπες (*anomalies*) ή outliers. Μια σχετική μέθοδος είναι η ανίχνευση αλλαγών και αποκλίσεων (*change and deviation detection*).

Ο σκοπός ενός αλγορίθμου ανίχνευσης παρεκτροπών είναι η ανακάλυψη πραγματικών ανωμαλιών / παρεκτροπών και η αποφυγή λανθασμένου χαρακτηρισμού ενός φυσιολογικού αντικειμένου ως έκτροπο. Δηλαδή, επιθυμούμε ανίχνευση υψηλού επιπέδου όσον αφορά τις τυχούσες ανωμαλίες, διατηρώντας όμως χαμηλά ποσοστά λανθασμένης προειδοποίησης.

Ως εφαρμογή της ανίχνευσης παρεκτροπών μπορούμε να αναφέρουμε τον προσδιορισμό απειλής (*fraud detection*) στην έγκριση δανείων ή πιστωτικών καρτών από μια τράπεζα. Αυτό είναι ένα ζήτημα υψίστης σημασίας για τέτοιους οργανισμούς. Επίσης, παραδείγματα αυτής της ομάδας αποτελούν οι εισβολές σε ένα δίκτυο (*network intrusions*), τα ασυνήθιστα καιρικά φαινόμενα ή οι διαταραχές του οικοσυστήματος.

2.3 Σύγκριση ταξινόμησης και συσταδοποίησης

Στην ενότητα αυτή, επικεντρώνουμε το ενδιαφέρον μας σε δύο μεθόδους που είναι από τις πιο χαρακτηριστικές στην ΕΔ και μας αφορούν στα πλαίσια της ενασχόλησής μας με τα κατηγορικά δεδομένα. Επιθυμώντας να διευκρινήσουμε τη διαφορά μεταξύ ταξινόμησης και συσταδοποίησης, πρέπει να σημειώσουμε ότι στην ταξινόμηση επιχειρείται η περιγραφή μιας λειτουργίας που αντιστοιχεί (ταξινομεί) ένα στοιχείο σε μια εκ των κατηγοριών οι οποίες είναι ήδη προκαθορισμένες.

Η **ταξινόμηση** χαρακτηρίζεται από ένα καλά ορισμένο σύνολο κατηγοριών, καθώς και ένα σύνολο προκαθορισμένων (*pre-classified*) παραδειγμάτων. Αντιθέτως, η **συσταδοποίηση** δε στηρίζεται σε προκαθορισμένες κατηγορίες ή παραδείγματα. Επιπλέον, στόχος μιας μεθόδου ταξινόμησης είναι αφενός η εκμάθηση (*learning*) και αφετέρου η κατηγοριοποίηση, δηλαδή η ταξινόμηση. Στην ουσία, δημιουργείται ένα μοντέλο που θα μπορεί να χρησιμοποιηθεί για την ταξινόμηση μελλοντικών δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη.

Πριν προχωρήσουμε στην επόμενη ενότητα, όπου δίνουμε μια χαρακτηριστική εφαρμογή και την εξηγούμε πάνω σε κάθε μέθοδο ΕΔ που έχουμε αναφέρει ως εδώ, αξίζει να αναφέρουμε ότι μια άλλη ορολογία για τις μεθόδους πρόβλεψης είναι **εποπτευόμενη εκμάθηση** (*supervised learning*). Αντιθέτως, οι μέθοδοι περιγραφής χαρακτηρίζονται ως **μη εποπτευόμενη εκμάθηση** (*unsupervised learning*), καθώς δεν υπάρχει προηγούμενη πληροφορία πάνω στην οποία μπορούμε να βασιστούμε για την περιγραφή των δεδομένων και η περιγραφή που κάνουμε στηρίζεται πάνω στη συγκεκριμένη βάση δεδομένων.

Ο όρος αυτός της μη εποπτευόμενης μάθησης, όμως, θεωρείται ότι ταιριάζει καλύτερα κυρίως στη συσταδοποίηση και όχι σε όλες τις μεθόδους περιγραφής. Ενώ, ο όρος της εποπτευόμενης μάθησης, αφορά τις μεθόδους πρόβλεψης γενικά. Ας περάσουμε, όμως σε μια θεωρητική εφαρμογή για να εμπεδώσουμε όσα ειπώθηκαν στις προηγούμενες ενότητες.

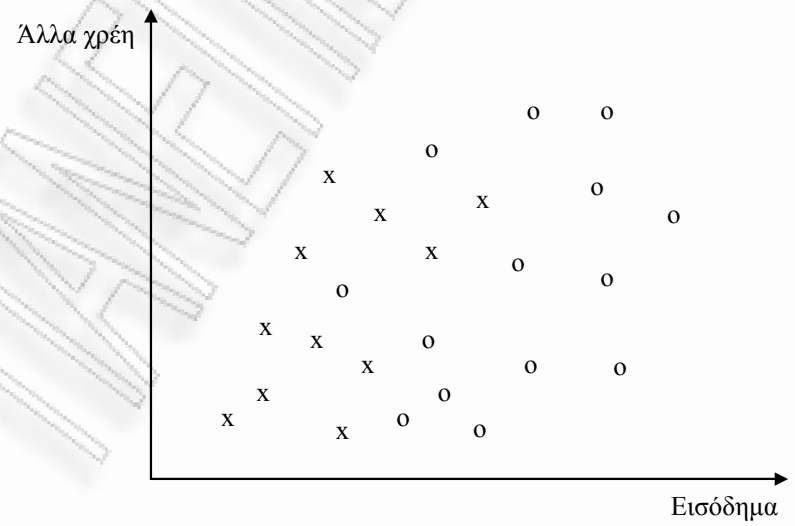
2.4 Εφαρμογή εμπέδωσης

Στα πλαίσια του ενδιαφέροντός μας για κατανόηση των βασικών τεχνικών της ΕΔ, θα επιχειρήσουμε να διευκρινίσουμε γραφικά τις βασικές διαφορές μέσω ενός παραδείγματος. Όπως, αναφέραμε, οι βασικότερες μέθοδοι είναι η ταξινόμηση και η παλινδρόμηση, εάν θέλουμε να προβούμε σε μοντελοποίηση πρόβλεψης, καθώς και η συσταδοποίηση, στην περίπτωση που θέλουμε να εφαρμόσουμε τεχνικές περιγραφικής μοντελοποίησης.

Ας υποθέσουμε ότι έχουμε ένα απλό σύνολο δεδομένων, όπου καταγράφονται οι περιπτώσεις ανθρώπων που έχουν πάρει δάνειο από μια συγκεκριμένη τράπεζα την τελευταία περίοδο. Έστω ότι έχουμε 25 περιπτώσεις και ότι για κάθε περίπτωση (άτομο) καταγράφονται οι τιμές δύο μεταβλητών: εισόδημα και χρέη προς τρίτους (άλλα δάνεια, πιστωτικές κάρτες κ.λπ.). Ένα απλό διδιάστατο γράφημα, που απεικονίζει τις 25 περιπτώσεις, δίνεται στο Σχήμα 2.4:

ΣΧΗΜΑ 2.4

Απεικόνιση συνόλου δεδομένων



Παρατηρώντας τον οριζόντιο άξονα, βλέπουμε πως κατανέμονται τα άτομα με βάση το εισόδημα, ενώ στον κάθετο άξονα βλέπουμε το ύψος των χρεών κάθε ατόμου. Στο σύνολο δεδομένων υπάρχει και άλλη μια στήλη, με βάση την οποία τα άτομα ταξινομούνται σε δύο κλάσεις:

(1) Αυτά που έχουν καθυστερήσει να πληρώσουν το δάνειό τους (συμβ. x)
(2) Αυτά που διατηρούν το δάνειό τους σε καλή κατάσταση (συμβ. o)

Το δάνειο στο οποίο αναφερόμαστε στις δύο αυτές κλάσεις είναι το δάνειο που έχουν πάρει τα 23 άτομα από τη συγκεκριμένη τράπεζα και δε συμπεριλαμβάνεται στη μεταβλητή που καταγράφουμε τα χρέη του ατόμου προς τρίτους. Βέβαια, στην πραγματικότητα, θα είχαμε ένα διάγραμμα περισσότερων διαστάσεων.

Όπως έχουμε ήδη αναφέρει, οι δύο διαφορετικοί στόχοι που μπορεί να έχουμε κατά τη διενέργεια μιας μεθόδου ΕΔ είναι η περιγραφή και η πρόβλεψη. Στα πλαίσια της πρόβλεψης, θεωρείται σαφές ότι πρέπει να επιλέξουμε κάποιες μεταβλητές ή γενικά ένα πεδίο της βάσης δεδομένων, με βάση το οποίο θα επιχειρήσουμε να προβλέψουμε κάποια κατάσταση ή μελλοντικές τιμές των άλλων μεταβλητών που μας απασχολούν. Όμως, η περιγραφή επικεντρώνεται στην εύρεση σημαντικών και κατανοητών προτύπων, τα οποία θα περιγράφουν τα δεδομένα (Fayyad et al., 1996-b).

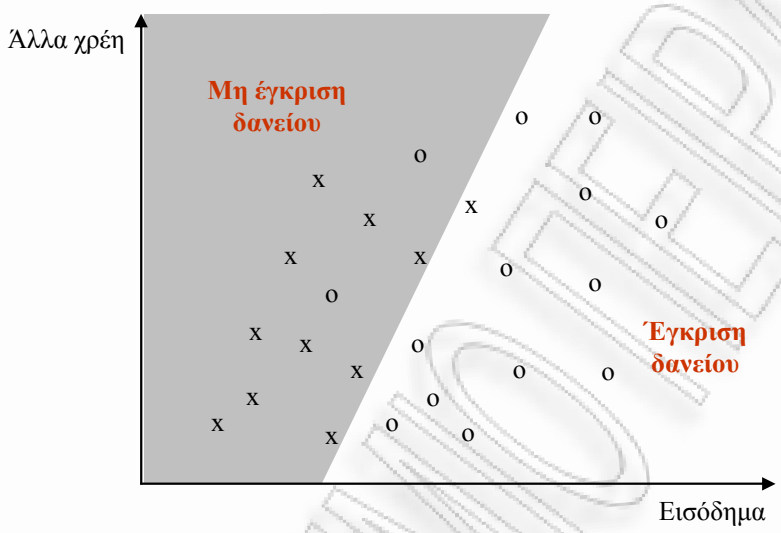
Τα όρια της διάκρισης μεταξύ περιγραφής και πρόβλεψης δεν είναι αρκετά ισχυρά. Δηλαδή, μια προβλεπτική μέθοδος μπορεί υπό άλλες συνθήκες να θεωρηθεί περιγραφική, ή και αντίθετα. Υπάρχουν διάφορες μέθοδοι περιγραφής, αλλά εμείς θα επικεντρωθούμε στη συσταδοποίηση. Επίσης, από την πρόβλεψη θα τονίσουμε μέσω σχημάτων τις διαφορές μεταξύ ταξινόμησης και παλινδρόμησης.

Μιλώντας για **ταξινόμηση**, εννοούμε την εκμάθηση (*learning*) μιας συνάρτησης, η οποία χαρτογραφεί (ταξινομεί) ένα αντικείμενο σε μια από τις υπάρχουσες προκαθορισμένες ομάδες (Weiss and Kulikowski, 1991 / Hand, 1981). Στο παράδειγμά μας, παρατηρώντας το Σχήμα 2.5, έχουμε μια διαμέριση των δεδομένων σε δύο περιοχές, ανάλογα με το εάν η τράπεζα θα συνεχίσει να τους παρέχει τις υπηρεσίες της διατηρώντας το δάνειο, ή εάν θα το

οδηγήσει σε οριστική καθυστέρηση. Αυτός ο διαχωρισμός σημαίνει ότι η τράπεζα ξεχωρίζει αυτούς τους 25 πελάτες σε «καλούς» και «κακούς» πελάτες.

ΣΧΗΜΑ 2.5

Αποτελέσματα ταξινόμησης



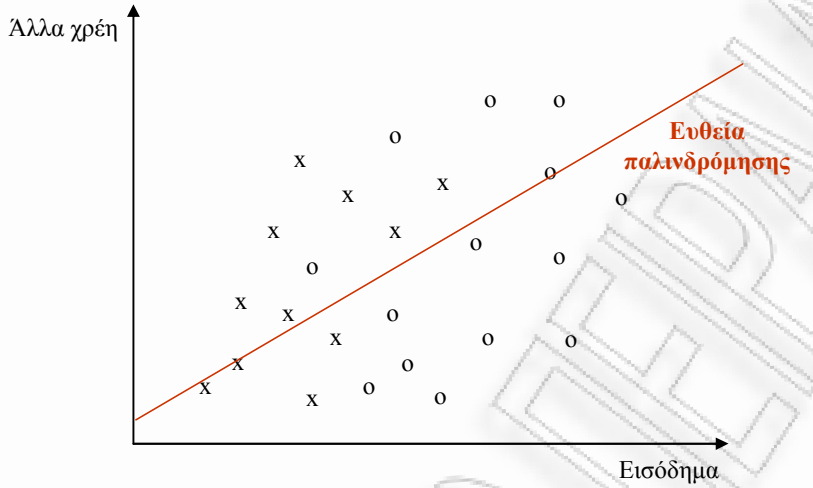
Με βάση αυτή την εφαρμογή ταξινόμησης, η τράπεζα θα μπορεί να κρίνει μελλοντικά ποιες αιτήσεις δανείων από νέους πελάτες θα εγκριθούν ή όχι. Να σημειώσουμε, όμως, ότι είναι δύσκολο να γίνει ένας «τέλειος» διαχωρισμός των ομάδων.

Στην περίπτωση που η μεταβλητή πρόβλεψης είναι συνεχής, τότε η συνάρτηση με βάση την οποία θα χαρτογραφηθούν τα δεδομένα κατασκευάζεται μέσω της **παλινδρόμησης**. Οι εφαρμογές που μπορούν να πραγματοποιηθούν μέσω αυτής της τεχνικής είναι πολλές. Για παράδειγμα, μπορεί να μας απασχολεί το συνολικό χρέος του πελάτη. Τότε, αυτό μπορεί να εκφραστεί ως γραμμική συνάρτηση του εισοδήματος.

Στο Σχήμα που ακολουθεί βλέπουμε πως θα μπορούσε να είναι η προσαρμογή της ευθείας παλινδρόμησης ώστε να μπορούμε να προβλέψουμε το συνολικό χρέος, συναρτήσει του εισοδήματος. Βέβαια, η προσαρμογή εδώ θεωρείται μη ικανοποιητική, καθώς έχουμε μικρό πλήθος παρατηρήσεων και υπάρχει ελαφριά συσχέτιση μεταξύ των δύο μεταβλητών.

ΣΧΗΜΑ 2.6

Εφαρμογή παλινδρόμησης

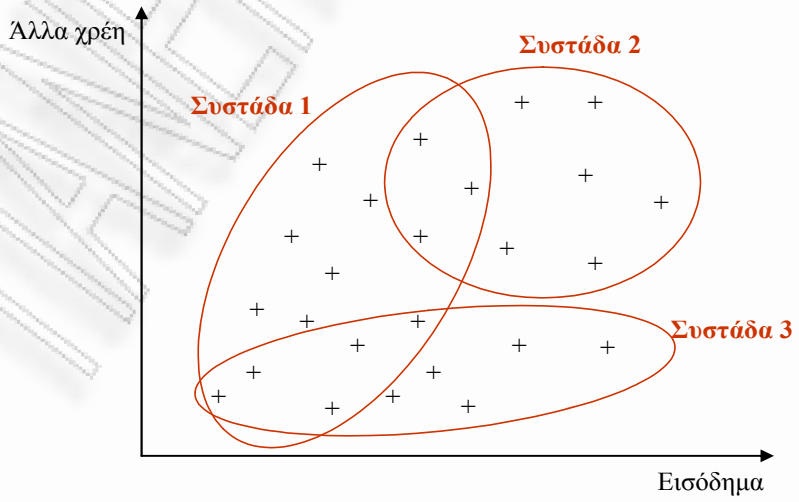


Κλείνοντας την ενότητα, θα αναφέρουμε μια κλασσική εργασία περιγραφής, που δεν είναι άλλη από τη **συσταδοποίηση**. Εάν επιχειρήσουμε να εφαρμόσουμε αυτή την τεχνική (Σχήμα 2.7), αυτό σημαίνει ότι θέλουμε μόνο να προσδιορίσουμε ένα πεπερασμένο σύνολο κατηγοριών ή συστάδων, ώστε να περιγράψουμε τα δεδομένα (Jain and Dubes, 1988).

Οι συστάδες που θα δημιουργηθούν μπορεί να είναι ξεχωριστές και συγκεκριμένες, ή να περιέχουν πιο πλούσια παρουσίαση, όπως για παράδειγμα μια συστάδα μέσα σε μια άλλη ή συστάδες που να επικαλύπτουν η μια την άλλη.

ΣΧΗΜΑ 2.7

Συσταδοποίηση



Στο παραπάνω Σχήμα, παρατηρούμε ότι υφίσταται το φαινόμενο της επικάλυψης συστάδων. Αυτό είναι ένα σύνηθες φαινόμενο σε πραγματικές αναλύσεις. Στο Σχήμα αυτό, έχουν δημιουργηθεί τρεις συστάδες, η καθεμία από τις οποίες έχει κάποια χαρακτηριστικά ή ετικέτες (*labels*). Οι συστάδες αυτές και οι ετικέτες τους, μπορεί να χρησιμοποιηθούν στη συνέχεια ως οι συγκεκριμένες ομάδες που ζητά η ταξινόμηση για να προχωρήσει σε καταμερισμό των ατόμων, προβλέποντας για παράδειγμα ποιοι πελάτες είναι πιο πιθανό να απομακρύνουν το χαρτοφυλάκιό τους από την τράπεζα.

Παρατηρώντας το Σχήμα 2.7, βλέπουμε ότι στη θέση των x και ο υπάρχουν +. Αυτό σημαίνει ότι στη συσταδοποίηση δε μας απασχολούν οι αρχικές ετικέτες των δεδομένων, δηλαδή αν έχουν καθυστερήσει να πληρώσουν το δάνειό τους ή όχι.

2.5 Τύποι δομής: μοντέλα και πρότυπα

Με βάση τους ορισμούς της ΕΔ στο πρώτο κεφάλαιο και της KDD παραπάνω, κρίνουμε απαραίτητο να κάνουμε ένα διαχωρισμό μεταξύ των όρων «**μοντέλο**» και «**πρότυπο**». Όπως έχουμε κατανοήσει, η ουσιαστική διαφορά είναι ο στόχος που έχουμε θέσει και ο τρόπος που θέλουμε να αξιοποιήσουμε και παρουσιάσουμε τα δεδομένα. Για παράδειγμα, ως στόχο μπορεί να έχουμε την περιγραφή των δεδομένων ή τη δημιουργία μιας πρόβλεψης για ένα θέμα που έχει τεθεί.

Αναλογιζόμενοι όλα αυτά θεωρούμε ότι, μέσα από τις διαδικασίες που θα πραγματοποιήσουμε, μπορούμε να οδηγηθούμε σε ένα **συνολικό** μοντέλο ή σε ένα **τοπικό** πρότυπο (Hand et al., 2001). Όταν μιλάμε για ένα τύπο δομής ως μοντέλο, ενοούμε μια συνολική σύνοψη ενός συνόλου δεδομένων. Με βάση το μοντέλο αυτό, εξετάζονται όλοι οι δυνατοί ισχυρισμοί. Για παράδειγμα, ας σκεφτούμε γεωμετρικά έναν πίνακα n γραμμών με διανύσματα διάστασης p. Έχουμε, δηλαδή, ένα χώρο p διαστάσεων. Ένα μοντέλο μπορεί να βγάλει ένα αποτέλεσμα (από έναν ισχυρισμό) για κάθε σημείο του χώρου, δηλαδή να προσδιορίσει ένα σημείο μιας συστάδας ή να προβλέψει την τιμή μιας άλλης μεταβλητής.

Ας θεωρήσουμε ένα απλό μοντέλο της μορφής $Y = aX + b$. Οι Y και X είναι οι μεταβλητές και οι a και b είναι οι παράμετροι του μοντέλου, οι οποίες προσδιορίζονται από την εκπόνηση των διεργασιών της ΕΔ. Στα πλαίσια των διεργασιών αυτών, η Y θεωρείται γραμμική συνάρτηση του X. Όμως, στη στατιστική, ένα μοντέλο λέγεται γραμμικό εάν αποτελεί

γραμμική συνάρτηση των παραμέτρων του. Μέσα από ένα μοντέλο όπως αυτό, μπορούν να απαντηθούν πολλά ερωτήματα.

Σε αντίθεση με τη συνολικότητα στην έκφραση των μοντέλων, εάν επιθυμούμε να επιλέξουμε ως τύπο δομής ένα πρότυπο, τότε σημαίνει ότι θέλουμε να απαντήσουμε κάποιους ισχυρισμούς που περιορίζονται σε συγκεκριμένες περιοχές του χώρου των p διαστάσεων. Ας σκεφτούμε έναν απλό ισχυρισμό όπως: «Εάν $X > x_1$, τότε $P(Y > y_1) = p_1$». Για τις μεταβλητές X και Y, εδώ, υπάρχουν συγκεκριμένοι περιορισμοί, οι οποίοι σχετίζονται με τη θεωρία πιθανοτήτων.

Κατά συνέπεια, διαπιστώνουμε ότι σε αντίθεση με ένα (συνολικό) μοντέλο, ένα (τοπικό) πρότυπο περιγράφει μια δομή που ασχολείται με ένα μικρό μέρος των δεδομένων ή του χώρου που ορίζουν. Προφανώς, εάν έχουμε μία βάση δεδομένων με πολλές καταγραφές ατόμων ή περιπτώσεων (*cases*), μόνο κάποιες από αυτές θα συμπεριφέρονται με έναν όμοιο τρόπο πάνω σε κάποιο θέμα. Το πρότυπο που θα προκύπτει, θα χαρακτηρίζει αυτή τη συγκεκριμένη ομάδα ατόμων.

Για παράδειγμα, εάν παρατηρούμε τις πωλήσεις προϊόντων μέσω διαδικτύου για μια συγκεκριμένη εταιρεία, μπορεί να προκύψει ότι μια ομάδα ανθρώπων που αγοράζει ένα προϊόν, μπορεί να αγοράζει και κάποιο άλλο με συγκεκριμένη πιθανότητα. Από την άλλη πλευρά, βέβαια, μπορεί να προκύψει ότι ένα μικρό σύνολο έκτροπων παρατηρήσεων διαφέρουν σε κάποιο ζήτημα σε σχέση με την πλειοψηφία. Αυτές οι παρατηρήσεις είναι και πάλι συγκεκριμένο μέρος του συνόλου και εκφράζονται μέσω ενός προτύπου.

Όπως είδαμε από τις απλοποιημένες εκφράσεις ενός μοντέλου και ενός προτύπου, υπάρχουν παράμετροι οι οποίες εκτιμώνται μέσω της ανάλυσης των διαθέσιμων δεδομένων και αφού έχουμε επιλέξει τον τύπο δομής που θέλουμε να εκφράσουμε. Για το μοντέλο οι παράμετροι ήταν τα a και b, ενώ για το πρότυπο είναι τα x_1, y_1 και p_1.

Είναι πλέον σαφές ότι ο διαχωρισμός μεταξύ μοντέλου και προτύπου είναι αρκετά σημαντικός. Η διάκριση αυτή, όμως, δε θεωρείται πάντοτε μια εύκολη διαδικασία. Οι διαδικασίες της Διερευνητικής Ανάλυσης Δεδομένων (*EDA*) και τα μοντέλα περιγραφής και πρόβλεψης που αναλύθηκαν στην ενότητα 2.2 έχουν σχέση με τη δημιουργία μιας δομής μοντέλου. Ενώ, η ανάλυση συνάφειας, οδηγεί στην έκφραση δομών προτύπου.

ΚΕΦΑΛΑΙΟ 3

Data Mining και Στατιστική

3.1 Εισαγωγή

Η λειτουργία της Εξόρυξης Δεδομένων (ΕΔ) αποτελεί τη διεπαφή μεταξύ της Επιστήμης Υπολογιστών και της Στατιστικής, καθώς χρησιμοποιεί στοιχεία από την εξελικτική πορεία των δύο αυτών τομέων. Στόχος μιας εφαρμογής της ΕΔ είναι η εξαγωγή ενδιαφέρουσας γνώσης από πολύ μεγάλες βάσεις δεδομένων (Glymour et al., 1997).

Η στατιστική επιστήμη παρουσιάζει ιδιαίτερη εξελιξιμότητα. Όμως, οι σύγχρονες ανάγκες διαχείρισης μεγάλου όγκου δεδομένων, απαιτούν τη συνεργασία με το διαθέσιμο λογισμικό. Αποτέλεσμα αυτού του συνδυασμού είναι μια νέα προσέγγιση κάθε ζητήματος όπου απαιτείται η ανάλυση δεδομένων.

Ο τομέας της ΕΔ ασχολείται με την εκμάθηση μέσα από δεδομένα, αλλά και την οργάνωση των δεδομένων και τη μετατροπή τους σε πληροφορία. Σε αυτό το πλαίσιο, οι στόχοι της ΕΔ και της στατιστικής είναι όμοιοι. Όμως, η ΕΔ αποσκοπεί σε αναδρομική ανάλυση των δεδομένων. Έτσι, παρατηρούμε ότι υπάρχουν ομοιότητες αλλά και διαφορές μεταξύ ΕΔ και Στατιστικής.

Οι άνθρωποι που ασχολούνται ενεργά με την ΕΔ ενδιαφέρονται περισσότερο για την κατανόηση των δεδομένων, παρά για την ακρίβεια ή την προβλεπτική τους ικανότητα. Οι υπό μελέτη εφαρμογές περιλαμβάνουν ένα τεράστιο πλήθος καταχωρημένων παρατηρήσεων, με πάρα πολλές καταγεγραμμένες μεταβλητές ανά παρατήρηση.

Για το λόγο αυτό, η ΕΔ στην πράξη επικεντρώνεται στην εξαγωγή προτύπων και όχι στη δημιουργία μοντέλων. Στην ουσία, αντί για την κατασκευή ενός συνολικού μοντέλου όπου περιλαμβάνονται όλες οι ενδιαφέρουσες μεταβλητές, προτιμώνται τα απλά και κατανοητά μοντέλα, υπό μορφή κανόνων, δέντρων, γραφημάτων κ.λπ.

Ένας αλγόριθμος ΕΔ έχει ως στόχο τη δήλωση ισχυρισμών για τοπικές εξαρτήσεις μεταξύ μεταβλητών. Έτσι, για να θεωρηθεί επιτυχημένος ένας αλγόριθμος, απαιτείται να έχει ιδιαίτερη υπολογιστική ικανότητα, χωρίς όμως να ξεχνάμε τη χρήση της Στατιστικής. Δηλαδή, η στατιστική επιστήμη αποτελεί τον πυρήνα όλων αυτών των διεργασιών. Στην

ενότητα που ακολουθεί παρουσιάζουμε τις βασικότερες έννοιες και θέματα της Στατιστικής που σχετίζονται με την έννοια της ΕΔ.

3.2 Απαιτούμενα θέματα από τη Στατιστική

Στην ενότητα αυτή, θα περιγράψουμε κάποιες από τις βασικές στατιστικές εφαρμογές που θεωρούμε ότι σχετίζονται με την ΕΔ. Πρόκειται για τις σημαντικότερες έννοιες της Στατιστικής που αξιοποιούνται σε μεγάλο βαθμό από την ΕΔ (Glymour et al., 1997). Για παράδειγμα, πολλές εφαρμογές της ΕΔ στηρίζονται σε συγκεκριμένες **κατανομές πιθανότητας**, καθώς και ιδιότητες των **τυχαίων μεταβλητών**. Επιπλέον, σημαντικό ρόλο παίζουν έννοιες όπως **ανεξαρτησία** τυχαίων μεταβλητών, υπό συνθήκη ανεξαρτησία και διάφορα μέτρα εξάρτησης, όπως είναι ο **συντελεστής συσχέτισης**.

Οι οικογένειες κατανομών διαθέτουν ιδιότητες με βάση τις οποίες μπορούν να προσδιορίσουν κάθε μέλος της οικογένειας δεδομένων, να κατασκευάσουν ένα **μοντέλο**, ή να προχωρήσουν σε **συμπερασματολογία** (βλ. Cox, 2006 / Casella and Berger, 2002). Όπως καταλαβαίνουμε, η γνώση των ιδιοτήτων κάθε οικογένειας κατανομών εξυπηρετεί στην ανάλυση δεδομένων, τη διεξαγωγή στατιστικών ελέγχων και τη διενέργεια κατάλληλης στατιστικής συμπερασματολογίας.

Στην περίπτωση που έχουμε ένα δείγμα αντί ολόκληρου του πληθυσμού, τότε προχωράμε την ανάλυσή μας κάνοντας κάποια **εκτίμηση** για τις τιμές των παραμέτρων. Συχνά, για να γίνει μια εκτίμηση, απαιτείται η εκπλήρωση ορισμένων προϋποθέσεων. Για παράδειγμα, ένας εκτιμητής πρέπει να είναι **συνεπής** (βλ. Cox and Hinkley, 1974) και να έχει τη μικρότερη δυνατή μεταβλητότητα. Επίσης, μια εκτίμηση πρέπει να παρουσιάζει τη **μικρότερη δυνατή αβεβαιότητα**.

Υπάρχουν αρκετές τεχνικές για την εκτίμηση της αβεβαιότητας. Οι βασικότερες από αυτές είναι η **επαναδειγματοληψία** (*resampling* – βλ. Efron and Tibshirani, 1993) και η **προσομοίωση** (βλ. Ross, 1997). Επιπλέον, ένας ακόμη στόχος της στατιστικής έρευνας είναι η μείωση των αναγκαίων υποθέσεων για να θεωρηθεί καλή μια εκτίμηση. Αυτό σημαίνει ότι επιθυμούμε να έχουμε έναν όσο γίνεται **εύρωστο** (*robust*) εκτιμητή (βλ. Huber, 1981).

Μια χρήσιμη τεχνική, εναλλακτική της κλασσικής προσέγγισης, είναι η **εκτίμηση κατά Bayes** (*Bayesian estimation* – βλ. Young and Smith, 2005). Με βάση αυτή τη μέθοδο μπορούμε, αντί της διενέργειας εκτίμησης μέσω ενός μοναδικού μοντέλου, να υποθέσουμε

περισσότερα «κατάλληλα» μοντέλα και να λάβουμε ως εκτίμηση το σταθμισμένο μέσο όρο των εκτιμήσεων που δίνονται από κάθε μοντέλο (βλ. Madigan and Raftery, 1994). Η εφαρμογή αυτή καλείται **Bayesian model averaging** και οδηγεί σε βελτίωση της προβλεπτικής ικανότητας (βλ. Bernardo and Smith, 1994).

Στην ΕΔ, τα μοντέλα που κατασκευάζονται είναι συνήθως αποτέλεσμα αυτοματοποιημένης αναζήτησης. Έτσι, ο υπολογισμός των πιθανών σφαλμάτων χρήζει ιδιαίτερης σημασίας. Αυτό μπορεί να απαιτεί **Monte Carlo ανάλυση** (βλ. Gentle, 2002). Παρατηρούμε ότι, αρχικά, οι αναλυτές δεδομένων αναγκάζονταν να αποφύγουν την ανάλυση περίπλοκων Μπεϋζιανών μοντέλων και τον υπολογισμό σύνθετων πιθανοφανειών. Αυτή η επιλογή οφειλόταν στη δυσχέρεια διενέργειας υπολογισμών.

Όμως, η πρόοδος των μεθόδων Monte Carlo και ειδικότερα μια ομάδα μεθόδων προσομοίωσης, οι **Μαρκοβιανές αλυσίδες Monte Carlo** (βλ. Gamerman and Lopez, 2006), οδήγησαν στην απελευθέρωση των υπολογιστικών διαδικασιών. Περισσότερες πληροφορίες δίνονται από το βιβλίο των Gilks et al. (1996).

Πέρα από αυτά οφείλουμε να επικεντρωθούμε σε δύο τομείς της στατιστικής επιστήμης, στοιχεία από τους οποίους χρησιμοποιούνται πολύ συχνά ώστε να γίνει έλεγχος περιπτώσεων και να ληφθούν σημαντικές αποφάσεις. Οι τομείς αυτοί είναι:

Έλεγχος υποθέσεων (*Hypothesis testing*)

Μιλώντας για στατιστικό έλεγχο υπόθεσης, θεωρούμε ότι έχουμε μια μηδενική και μια εναλλακτική υπόθεση. Υπάρχουν πολλές εκφράσεις ελέγχου υποθέσεων για παραμέτρους του πληθυσμού ή ενός δείγματος. Η μηδενική (H_0) και η εναλλακτική (H_1) υπόθεση καλύπτουν συνολικά όλες τις πιθανές τιμές της υπό ερώτηση παραμέτρου ή παραμέτρων (Aczel, 1989).

Η απόφασή μας λαμβάνεται με βάση την τιμή μιας **στατιστικής συνάρτησης**, σύμφωνα με την οποία οδηγούμαστε σε **αποδοχή** ή **απόρριψη** της μηδενικής υπόθεσης, εφαρμόζοντας ένα **κανόνα απόφασης**. Για τους κανόνες απόφασης γίνονται σχόλια στο τέλος της ενότητας, ενώ αναλυτικές πληροφορίες για την εκτίμηση παραμέτρων και τον έλεγχο υποθέσεων δίνονται στο βιβλίο του Koch (1999).

Εφόσον έχουμε ευρεία χρήση κάποιων στατιστικών ελέγχων, οφείλουμε να δηλώνουμε και κάποιους σημαντικούς περιορισμούς τους. Θεωρώντας τον έλεγχο υποθέσεων ως μια

μονόπλευρη μέθοδο εκτίμησης, τότε τον κρίνουμε ως ασυνεπή εκτός κι αν το επίπεδο σημαντικότητας α ενός ελέγχου μειώνεται κατάλληλα όσο μεγαλώνει το μέγεθος δείγματος (Glymour et al., 1997).

Γενικά, εάν συνδυάσουμε και ελέγξουμε από κοινού δύο ελέγχους υπόθεσης επιπέδου α, τότε δε θα οδηγηθούμε σε έλεγχο επιπέδου α. Η άποψη αυτή βασίζεται στη **διόρθωση Bonferroni** (*Bonferroni correction*), σύμφωνα με την οποία εάν θέλουμε να κάνουμε m ελέγχους υποθέσεων επιπέδου σημαντικότητας α ο καθένας, τότε το επίπεδο σημαντικότητας, έστω γ, για κάθε έλεγχο ορίζεται από τη σχέση:

$$\gamma = \frac{\alpha}{m}$$

Η διόρθωση Bonferroni αναπτύχθηκε από τον Ιταλό μαθηματικό Carlo Emilio Bonferroni και βασίζεται στην ανισότητα Bonferroni (*Bonferroni inequality* – βλ. Casella and Berger, 2002).

Μια εναλλακτική προσέγγιση της ταυτόχρονης συμπερασματολογίας δίνεται από τον Scheffé, η οποία καλείται και μέθοδος **S** (βλ. Scheffé, 1959). Επίσης, οι ειδικές περιπτώσεις όπου απαιτείται ταυτόχρονος έλεγχος πολλαπλών υποθέσεων και οι σχετικοί κανόνες καταγράφονται από τον Miller (1981).

Ένα σημαντικό συμπέρασμα που αφορά την ΕΔ είναι ότι το επίπεδο σημαντικότητας α ενός ελέγχου δεν έχει καμμία σχέση με την πιθανότητα σφάλματος σε μια διαδικασία αναζήτησης που περιλαμβάνει τον έλεγχο μιας σειράς υποθέσεων (Glymour et al., 1997).

Έτσι, σε διαδικασίες ΕΔ που χρησιμοποιούν μια σειρά από ελέγχους υποθέσεων, το επίπεδο εμπιστοσύνης α δεν θεωρείται γενικά μια εκτίμηση της πιθανότητας σφάλματος σε σχέση με το αποτέλεσμα της αναζήτησης. Οι ενασχολούμενοι με την ΕΔ πρέπει να προσέχουν ότι ενώ οι πιθανότητες σφάλματος των ελέγχων σχετίζονται με την «αλήθεια» των υποθέσεων, ο συνδυασμός τους είναι ένα λεπτό ζήτημα.

Αξιολόγηση μοντέλου (*Model scoring*)

Οι μαρτυρίες που παρέχονται από τα δεδομένα θα έπρεπε να μας οδηγούν στην επιλογή ή προτίμηση κάποιων μοντέλων ή υποθέσεων από άλλα. Ένα σκορ (*score*) είναι κάθε κανόνας που διατάσσει τα μοντέλα που κατασκευάζονται βάσει κάποιων δεδομένων και τα αντιστοιχεί

με βάση την προτίμηση ανά μοντέλο. Οι κανόνες αξιολόγησης θεωρούνται μια εναλλακτική των ελέγχων (Glymour et al., 1997).

Οι πιο γνωστοί κανόνες αξιολόγησης είναι τα **κριτήρια πληροφορίας** των **Akaike** (Akaike, 1974) και **Bayes** (Schwarz, 1978), αλλά και το **ελάχιστο μήκος περιγραφής** (*Minimum Description length - MDL*). Στη γενική περίπτωση, το κριτήριο του Akaike ή AIC (*Akaike information criterion*), δίνεται από τον τύπο:

$$\mathrm{AIC} = 2\log \mathrm{L} + 2\mathrm{q}$$

όπου q το πλήθος των παραμέτρων του μοντέλου και L η πιθανοφάνεια υπολογισμένη στον εκτιμητή μέγιστης πιθανοφάνειας του υπό εκτίμηση μοντέλου (Maimon and Rokach, 2005).

Το κριτήριο Bayes ή BIC (*Bayesian information criterion*) προσεγγίζει εκ των υστέρων πιθανότητες σε μεγάλα δείγματα, δεδομένου ότι έχει εκφραστεί μια εκ των προτέρων πιθανότητα. Αυτή η εκ των υστέρων πιθανότητα αποτελεί από μόνη της μια συνάρτηση αξιολόγησης. Ο τύπος του κριτηρίου αυτού είναι:

$$\mathrm{BIC} = -2 \cdot \log(\mathrm{L}) + \mathrm{q} \cdot \log(\mathrm{n})$$

όπου n το μέγεθος του δείγματος και q, L ομοίως με παραπάνω. Περισσότερες πληροφορίες σχετικά με τα κριτήρια πληροφορίας και τις διαδικασίες επιλογής μοντέλου (*model selection*) δίνονται από τους Burnham και Anderson (2002), αλλά και από τους Hand et al. (2001).

Η προσέγγιση με βάση το κριτήριο MDL δίνει ένα κριτήριο επιλογής που τυπικά φαίνεται όμοιο με το BIC, αλλά στηρίζεται σε καλύτερη διενέργεια κωδικοποίησης. Σύμφωνα με το κριτήριο αυτό, η καλύτερη υπόθεση για ένα δοθέν σύνολο δεδομένων είναι αυτή που οδηγεί στη μεγαλύτερη συμπίεση των δεδομένων. Ένα ενδιαφέρον βιβλίο που ασχολείται με το κριτήριο MDL και την πρακτική εφαρμογή του είναι αυτό των Grünwald και Rissanen (2007).

Συνεχίζοντας την περιήγησή μας στους τομείς και τις έννοιες της Στατιστικής που συνεισφέρουν στην ΕΔ, ας σκεφτούμε ότι πολλές φορές, ενδιαφερόμαστε να χρησιμοποιήσουμε ένα δείγμα ή μια βάση δεδομένων, έτσι ώστε να **προβλέψουμε** τα χαρακτηριστικά ενός νέου δείγματος. Για τη διενέργεια αυτής της πρόβλεψης, θεωρούμε ότι τα δύο δείγματα έχουν ληφθεί από την ίδια κατανομή πιθανότητας. Όπως και στην εκτίμηση, έτσι και στην **πρόβλεψη** (*prediction*), ενδιαφερόμααστε για τη μέτρηση της **αξιοπιστίας** και της **αβεβαιότητας**, λαμβάνοντας υπόψη τη διακύμανση του στοιχείου που οδηγεί στην πρόβλεψη.

Οι μέθοδοι πρόβλεψης υποθέτουν κάποια δομή στην κατανομή πιθανότητας. Στην ΕΔ, η δομή αυτή ορίζεται από τους ειδικούς ή λαμβάνεται αυτοματοποιημένα από τη βάση δεδομένων σε μορφή συμπεράσματος. Για παράδειγμα, η **παλινδρόμηση** υποθέτει ένα συγκεκριμένο λειτουργικό τύπο που σχετίζεται με τις μεταβλητές, καθώς διαχωρίζει τις μεταβλητές σε επεξηγηματικές (ερμηνευτικές) μεταβλητές και σε μεταβλητές απόκρισης, ενώ υποθέτει ότι υπάρχει γραμμική σχέση μεταξύ τους (βλ. Casella and Berger, 2002). Επίσης, μια δομή μπορεί να προσδιοριστεί στα πλαίσια περιορισμών, όπως ανεξαρτησία, υπό συνθήκη ανεξαρτησία, προϋποθέσεις βασισμένες σε συσχετίσεις κ.λπ.

Κλείνοντας αυτή την ενότητα, αναφέρουμε μια κρυμμένη δύναμη η οποία βρίσκεται πίσω από την ιστορική ανάπτυξη της στατιστικής επιστήμης. Αυτή είναι η εμπέδωση της σχέσης **αιτίας – αιτιατού**. Από την περίοδο δράσης των Bernoulli και Laplace, η έλλειψη αιτιολογικής σύνδεσης μεταξύ δύο μεταβλητών οδηγούσε στην θεώρησή τους ως ανεξάρτητες. Η ίδια ιδέα αποτελεί θεμελιώδες συστατικό για τη θεωρία του **πειραματικού σχεδιασμού**.

Το 1921, ο Wright εισήγαγε την ιδέα παρουσίασης **αιτιολογικών υποθέσεων** (*causal hypothesis*) με τη χρήση **κατευθυνόμενων γραφημάτων**. Στη συνέχεια, το 1982, οι Kiiveri και Speed συνδύασαν τα κατευθυνόμενα γραφήματα με μια γενικευμένη σύνδεση μεταξύ ανεξαρτησίας και έλλειψης αιτιολογικής σύνδεσης την οποία κάλεσαν **συνθήκη Markov**. Έτσι, οδηγηθήκαμε στα λεγόμενα **Μπεϋζιανά δίκτυα** (βλ. Jensen, 2001).

Οι ενασχολούμενοι με την ΕΔ οφείλουν να προσέχουν τα λάθη που προκαλούνται από συμπερασματολογίες βασισμένες σε αιτίες από ανεξέλεγκτα δείγματα. Τα προβλήματα μπορεί να δημιουργούνται από **μη καταγεγραμμένες αιτίες συσχετίσεων** μεταξύ καταγεγραμμένων μεταβλητών, ή από **μεροληψία** κατά την επιλογή δείγματος. Επίσης, η **έλλειψη δεδομένων** ή η **διαφορά δομών** μεταξύ των παρατηρήσεων αποτελεί πρόβλημα.

Η εμπέδωση της σχέσης αιτίας και αιτιατού οδηγεί στη λήψη αποφάσεων. Η **θεωρία της «λογικής» επιλογής** (*theory of rational choice*) προϋποθέτει ότι ο λήπτης απόφασης έχει στη διάθεσή του ένα συγκεκριμένο σύνολο εναλλακτικών κινήσεων και την πιθανότητα εμφάνισης όλων των άλλων εναλλακτικών. Τότε, ένας **κανόνας απόφασης** προσδιορίζει ποια από τις εναλλακτικές κινήσεις θα έπρεπε να πραγματοποιηθεί.

Με βάση τη στατιστική και οικονομική βιβλιογραφία, υπάρχουν πολλοί κανόνες απόφασης, όπως η **μεγιστοποίηση αναμενόμενης χρησιμότητας** (*maximizing expected*

utility) ή η **ελαχιστοποίηση μέγιστης απώλειας** (*minimizing maximum loss*). Στην ουσία, η λογική λήψη αποφάσεων και ο σχεδιασμός αποτελούν τους στόχους της ΕΔ. Πέρα από την παροχή τεχνικών και μεθόδων για ΕΔ, η θεωρία λογικής επιλογής παρέχει κανόνες για τη χρήση της πληροφορίας που αποκτάται από μια βάση δεδομένων.

Το πλαίσιο δράσης της **λογικής λήψης αποφάσεων** απαιτεί τη γνώση των **πιθανοτήτων**, αλλά και των αποτελεσμάτων που μπορεί να έχει κάθε εναλλακτική. Ένας από τους προταρχικούς στόχους της ΕΔ, αλλά και της στατιστικής συμπερασματολογίας γενικά, είναι η γνώση των αποτελεσμάτων των κινήσεων, καθώς αυτό θα σημαίνει γνώση μέρους της σχέσης αιτίας και αποτελέσματος.

Αυτές ήταν οι βασικότερες έννοιες που δανείζεται η ΕΔ από τη στατιστική επιστήμη. Βέβαια, πέρα από τις παραπάνω εφαρμογές, υπάρχουν και άλλες αρκετά χρήσιμες, όπως η **Ανάλυση Χρονοσειρών** (βλ. Hamilton, 1994 / Davidson and Mackinnon, 2004) και η **Μετα-ανάλυση** (βλ. Schulze, 2004).

3.3 Σύγκριση των δύο τομέων

Από τη στιγμή που οι ιδέες και μέθοδοι της Στατιστικής θεωρούνται τόσο σπουδαίες για την ΕΔ, είναι λογικό να αναρωτιόμαστε εάν τελικά υπάρχουν ουσιαστικές διαφορές μεταξύ των δύο τομέων. Όμως, θεωρείται ότι η ΕΔ δεν είναι απλά μια διερευνητική στατιστική εφαρμογή, αλλά πολύ περισσότερα.

Όπως αποδεικνύεται και από τα προηγούμενα, η ΕΔ δανείζεται πολλά πράγματα από τον κλάδο της στατιστικής επιστήμης. Όμως, εάν επιθυμούσαμε να κάνουμε σύγκριση μεταξύ των δύο τομέων, θα μπορούσαμε να πούμε ότι υπάρχουν διαφορετικά σημεία υπεροχής για κάθε τομέα. Στις υποενότητες που ακολουθούν, γίνονται σχετικά σχόλια ανά τομέα, με βάση τις απόψεις των Hand et al. (2001).

3.3.1 Σημεία υπεροχής της Εξόρυξης Δεδομένων

Η πιο θεμελιώδης διαφορά μεταξύ των κλασσικών στατιστικών εφαρμογών και της ΕΔ είναι το **μέγεθος του συνόλου δεδομένων**. Για ένα κλασσικό στατιστικό, ένα «τεράστιο» σύνολο δεδομένων θα μπορούσε να περιέχει μερικές εκατοντάδες χιλιάδες σημείων (παρατηρήσεων). Όμως, κάποιος που ασχολείται με την ΕΔ, είναι πιθανό να χειριστεί μερικά εκατομμύρια ή και δισεκατομμύρια σημείων!

Μια συνηθισμένη εργασία για την ΕΔ μπορεί να είναι η ενασχόληση με μια βάση δεδομένων μεγέθους αρκετών gigabyte ή ακόμη και terabyte. Ας σκεφτούμε για παράδειγμα των εταιρεία Mobil Oil η οποία, σύμφωνα με σχετική αναφορά, αποθήκευσε το 1993 πάνω από 100 terabyte δεδομένων σχετικών με τη διερεύνηση πετρελαίου. Επίσης, το σύστημα παρατήρησης της Γης από τη NASA (*NASA Earth Observing System*) σχεδιάστηκε ώστε να παράγει πολλαπλά gigabyte δεδομένων σε γραμμές (*row data*) ανά ώρα (Fayyad et al., 1996). Φυσικά, αυτά είναι μόνο δύο παραδείγματα από όσα μπορεί να πραγματοποιήσει η ΕΔ.

Όσον αφορά το μέγεθος του συνόλου δεδομένων, υπάρχουν και μεγαλύτερες δυσκολίες που προκύπτουν όταν υπάρχουν **πολλές μεταβλητές**. Ένας στατιστικός αντιμετωπίζει πρόβλημα όταν το πλήθος των κελιών αυξάνει εκθετικά ανά παρατήρηση, λόγω της αύξησης των μεταβλητών. Έτσι, ζητήματα όπως η ακριβής εύρεση εκτιμήσεων των πυκνοτήτων πιθανότητας σε πολυδιάστατους χώρους γίνονται πολύ δύσκολα. Σε ένα χώρο πολλών διαστάσεων, το «κοντινότερο» σημείο μπορεί να είναι πολύ μακριά.

Όπως καταλαβαίνουμε, η περίπτωση ύπαρξης πολλών μεταβλητών σε ένα σύνολο δεδομένων οδηγεί στην δημιουργία επιπρόσθετων περιορισμών στην αρχική επιλογή μοντέλου από έναν ειδικό. Στα πλαίσια της ΕΔ, ένα σύνολο δεδομένων δεν είναι αυτό που βλέπει απλά ένας στατιστικός, δηλαδή μερικές γραμμές που παρουσιάζουν τα αντικείμενα και κάποιες στήλες όπου δίνονται οι μεταβλητές. Για παράδειγμα, η επιλογή ενός τυχαίου δείγματος από ένα σύνολο δεδομένων μπορεί να μην είναι μια εύκολη υπόθεση, όπως θα θεωρούσε ένας στατιστικός. Στην ΕΔ, ένα αρχείο μπορεί να αποθηκεύεται ταυτόχρονα σε πολλές μηχανές, διαιρεμένο σε μέρη.

Τέλος, εκτός από το πρόβλημα κατά την αύξηση των μεταβλητών, δεν πρέπει να ξεχνάμε τα σύνολα δεδομένων που αναπτύσσονται διαρκώς. Ας σκεφτούμε, για παράδειγμα την καταγραφή εισερχομένων κλήσεων στο τμήμα τηλεφωνικής εξυπηρέτησης ή την καταγραφή κατανάλωσης ηλεκτρικού ρεύματος. Αρχεία σαν αυτά πολλαπλασιάζουν διαρκώς το μέγεθός τους και προκαλούν αλλαγές στη φύση του προβλήματος και την αναζήτηση λύσης. Στην περίπτωση αυτή, η ΕΔ μπορεί να βοηθήσει.

3.3.2 Σημεία υπεροχής της Στατιστικής

Στην προηγούμενη υποενότητα αναφέραμε τα προβλήματα που ανακύπτουν σε μια στατιστική εφαρμογή όταν το μέγεθος ενός συνόλου δεδομένων είναι αρκετά μεγάλο ή αυξάνει διαρκώς. Όμως, δε μπορούμε να παραβλέψουμε τις αδυναμίες της ΕΔ, που αποτελούν σημεία υπεροχής της Στατιστικής.

Όπως σχολιάσαμε και στο πρώτο κεφάλαιο, η ΕΔ είναι μια δευτερεύουσα διαδικασία ανάλυσης δεδομένων. Αυτό σημαίνει ότι τα προς ανάλυση δεδομένα είχαν συλλεχθεί αρχικά για κάποιο άλλο σκοπό. Το προσόν, λοιπόν, της Στατιστικής είναι ότι αποτελεί την **πρωτογενή ανάλυση** για τα δεδομένα. Δηλαδή, τα δεδομένα συλλέγονται ύστερα από τη διαμόρφωση συγκεκριμένων ερωτημάτων, ενώ στη συνέχεια αναλύονται ώστε να απαντηθούν τα ερωτήματα αυτά.

Στην πραγματικότητα, η Στατιστική συλλέγει δεδομένα με τη διενέργεια πιο έγκυρων μεθόδων, όπως του Πειραματικού Σχεδιασμού (*Experimental Design*). Το πρόβλημα που αντιμετωπίζει η ΕΔ είναι ότι όταν επιχειρείται η επίλυση ενός ζητήματος μέσω της ανάλυσης δεδομένων που δεν είχαν συλλεχθεί για το ζήτημα αυτό, τότε μπορεί να μην προκύψει το ιδανικό αποτέλεσμα. Μπορεί, δηλαδή, να μην υπάρξει το κατάλληλο ταίριασμα των δεδομένων στο συγκεκριμένο ζήτημα.

Πέρα από τον τρόπο συλλογής των δεδομένων, τα μεγάλα σύνολα δεδομένων που χειρίζεται η ΕΔ αντιμετωπίζουν και άλλα σοβαρά προβλήματα. Για παράδειγμα, ένα τεράστιο σύνολο δεδομένων μπορεί να περιέχει **ελλείπουσες τιμές** (*missing values*), **θόρυβο** (*noise*), ή **«φθαρμένα»** (*corrupted*) στοιχεία. Περισσότερα σχόλια πάνω σε αυτό το πρόβλημα που αντιμετωπίζει η ΕΔ γίνονται στο επόμενο κεφάλαιο. Βέβαια, τέτοια ζητήματα προκύπτουν και σε στατιστικές εφαρμογές, αλλά εκεί είναι πιο εύκολη η αντιμετώπιση των προβληματικών στοιχείων.

Κλείνοντας, αξίζει να αναφέρουμε ότι η ΕΔ συμβαδίζει με τις κλασσικές τεχνικές διερευνητικής ανάλυσης δεδομένων της Στατιστικής, αλλά είναι σε θέση να αντιμετωπίζει και άλλα ζητήματα. Ένα πολύ μεγάλο ή ασυνήθιστο (μη παραδοσιακό) σύνολο δεδομένων δε μπορεί να διαχειριστεί τόσο εύκολα από μια στατιστική εφαρμογή.

3.4 Πλαίσιο συνεργασίας

Απο τα μέσα της δεκαετίας του 1960, στα πλαίσια της απελευθερωμένης διερευνητικής ανάλυσης δεδομένων, γίνονταν προσπάθειες επίλυσης προβλημάτων σχετικών με τη μοντελοποίηση. Σύμφωνα με τον Tukey, έπρεπε να γίνει με πιο επιστημονικό τρόπο η προσέγγιση τέτοιων ζητημάτων. Τριάντα χρόνια μετά, η στατιστική κοινότητα υιοθέτησε την άποψη του και έκανε μεγάλη πρόοδο (Glymour et al., 1997).

Με βάση την εξέλιξη αυτή, είμαστε σε θέση να πούμε ότι η **αναζήτηση μοντέλου** (*model search*) είναι ένα κρίσιμο και αναπόφευκτο βήμα στη διαδικασία μοντελοποίησης. Επιπλέον, παρατηρήθηκε η επινόηση υπολογιστικών μεθόδων για την αναζήτηση συμπερασματικών διαδικασιών. Από τη σύγχρονη στατιστική, υπάρχουν τρία ζητήματα υψίστης σημασίας για τους ενασχολούμενους με την ΕΔ: η **σαφήνεια** στον ορισμό των στόχων, η χρήση μεθόδων που αποτελούν **αξιόπιστο** μέσο προς το στόχο, αλλά και η συναίσθηση της **αβεβαιότητας** των μοντέλων και των εκτιμήσεων.

Για να μπορούν να λυθούν αυτά τα ζητήματα, πρέπει ο χρήστης των μεθόδων ΕΔ να είναι σε θέση να εντοπίζει τις εκάστοτε προϋποθέσεις που μπορεί να απαιτούνται. Πέρα από αυτό, ο συγκερασμός ΕΔ και Στατιστικής μπορεί να οδηγήσει σε ενδιαφέροντα αποτελέσματα, αρκεί να συνηδειτοποιήσουμε ορισμένα ζητήματα που είναι πιθανό να λάβουν χώρα στην πράξη.

Ακολουθώντας τις εξελίξεις της εποχής, βλέπουμε την εξέλιξη στον ψηφιακό κόσμο και τα διαθέσιμα εργαλεία λογισμικού. Πλέον, η πρόσβαση σε ψηφιοποιημένα στοιχεία είναι πολύ εύκολη. Αυτό, όμως, δε σημαίνει ότι το ίδιο εύκολα μπορεί να γίνει και μια στατιστική ανάλυση. Το φαινόμενο αυτό είναι μια πραγματικότητα στην ΕΔ. Βέβαια, σύμφωνα με τους Glymour et al. (1997), η στατιστική είναι ένα απαραίτητο αλλά όχι επαρκές συστατικό για την πρακτική της ΕΔ.

Με βάση τις απόψεις των ενασχολούμενων με την ΕΔ, οι στατιστικές έννοιες δεν αφορούν κανέναν παρά μόνο τους ερευνητές, καθώς δεν αντικατοπτρίζονται στην πραγματικότητα. Οι εφαρμογές στον πραγματικό κόσμο θεωρείται ότι δεν συμβαδίζουν πάντα με τις προϋποθέσεις της Στατιστικής. Αυτή είναι και η αιτία που επιχειρούν να συνδυαστούν οι σύγχρονες στατιστικές εφαρμογές με την μηχανική πρακτική. Οι σύγχρονοι μηχανικοί ερευνητές απαιτείται να γνωρίζουν στοιχεία στατιστικής όπως θεωρία πιθανοτήτων κ.λπ. Το προσόν τους είναι η εφαρμογή των εργαλείων τους πάνω στις στατιστικές μεθόδους.

Όπως μπορούμε να αντιληφθούμε, ο συνδυασμός αυτός μπορεί να επιφέρει αξιόλογα αποτελέσματα, τα οποία μάλιστα θα καλύπτουν και τις ανάγκες της σύγχρονης εποχής. Ήδη έχουμε δει αξιοσημείωτη πρόοδο στην οικονομική μοντελοποίηση (*financial modeling*), την αναγνώριση φωνής και την επιδημιολογία. Ακόμη, η στατιστική προσέγγιση στην ανάλυση δεδομένων έδωσε ζωή στη Βιοστατιστική, έναν αναπτυσσόμενο τομέα των επιστημών υγείας.

Λαμβάνοντας υπόψη όλα τα προηγούμενα, εκφράζουμε την άποψη ότι η ΕΔ είναι ικανή να προσφέρει πολλά στο σύχρονο τρόπο ζωής, χωρίς όμως να ξεχνάμε την προσφορά της στατιστικής επιστήμης. Η ουσία της συμβίωσης ΕΔ και Στατιστικής είναι ότι οι ενασχολούμενοι με την ΕΔ χρειάζεται να εμπεδώσουν τους κανόνες της Στατιστικής, ενώ οι στατιστικοί πρέπει να κατανοήσουν τη φύση των σημαντικών προβλημάτων που καλείται να αντιμετωπίσει η ΕΔ. Άλλωστε, η Στατιστική έχει ασκήσει την επιρροή της σε πολλούς τομείς στο παρελθόν και σίγουρα μπορεί να προσφέρει ακόμη περισσότερα.

3.5 Εφαρμογή της Εξόρυξης Δεδομένων σε Κατηγορικά Δεδομένα

Ολοκληρώνοντας το πρώτο κεφάλαιο της εργασίας αυτής, αναφέραμε ότι στόχος μας είναι να συγκεντρώσουμε τις μεθοδολογίες και αλγορίθμους που επιχειρούν να εφαρμόσουν τις τεχνικές του ΕΔ σε κατηγορικά ή μικτά σύνολα δεδομένων. Στα πλαίσια αυτής της αναζήτησης, θεωρούμε ότι πρέπει να παρουσιάσουμε συνοπτικά την έννοια των κατηγορικών δεδομένων. Να σημειώσουμε επίσης ότι ως μικτά σύνολα δεδομένων εννοούμε αυτά που περιλαμβάνουν συνεχείς και κατηγορικές μεταβλητές.

Μια κατηγορική μεταβλητή αποτελείται από ένα σύνολο κατηγοριών οι οποίες δεν επικαλύπτουν η μία την άλλη (βλ. Kateri, 2008). Για το λόγο αυτό, τα κατηγορικά δεδομένα θεωρούνται ότι είναι κάποιες **μετρήσεις** (*counts*), δηλαδή η καταγραφή των συχνοτήτων εμφάνισης κάθε κατηγορίας της υπό μελέτη μεταβλητής. Η κλίμακα μέτρησης μιας κατηγορικής μεταβλητής μπορεί να είναι **διατάξιμη** (*ordinal*) ή **ονοματική** (*nominal*).

Ως διατάξιμες εννοούμε τις μεταβλητές που περιλαμβάνουν κατηγορίες οι οποίες θεωρείται ότι έχουν μια φυσική διάταξη. Για παράδειγμα, μεταβλητές όπως «κοινωνική τάξη» ή «επίπεδο μόρφωσης» είναι διατάξιμες, καθώς μπορεί να έχουν τρία επίπεδα, όπως «χαμηλό», «μέτριο» και «υψηλό». Αντιθέτως, οι ονοματικές μεταβλητές περιλαμβάνουν κατηγορίες οι οποίες δε μπορούν να διαταχθούν σε καμμία περίπτωση. Παραδείγματα τέτοιων μεταβλητών είναι οι «φύλο» και «θρήσκευμα».

Είναι εύκολο να αντιληφθούμε ότι στις διατάξιμες και ονοματικές μεταβλητές δε μπορεί να εκφραστεί έννοια απόστασης. Επίσης, είναι πολύ πιθανό να συναντήσουμε μια παραδοσιακά συνεχή μεταβλητή όπως το «εισόδημα», η οποία να συμπεριφέρεται ως διατάξιμη και τελικά να τη χρησιμοποιήσουμε έτσι στην ανάλυση μας. Για παράδειγμα, το «εισόδημα» μπορεί να θεωρηθεί ότι έχει κάποια επίπεδα από «χαμηλό» έως «υψηλό», ενώ η «ηλικία» από «νέος» ως «ηλικιωμένος».

Ο τύπος των μεταβλητών που μπορεί να αποδειχθούν σημαντικές ως προς την πληροφορία που προσφέρουν σε μια έρευνα καθορίζει και τη μέθοδο που θα επιλεγεί για την περαιτέρω ανάλυση. Πολλές φορές, δε μας απασχολεί πως συμπεριφέρθηκε ένα μεμονωμένο άτομο ηλικίας 28 ετών, αλλά ποια ήταν για παράδειγμα η συμπεριφορά των νέων ηλικίας 25 έως 30 ετών. Έτσι, έχει σημασία να ορίζουμε τον τύπο μεταβλητών με βάση το βέλτιστο αποτέλεσμα που μπορεί να προκύψει από την ερμηνεία τους.

Συχνά, οι μεταβλητές ενός μεγάλου συνόλου δεδομένων καταλήγουν να θεωρούνται όλες κατηγορικές. Επομένως, θεωρείται απαραίτητη η συμβολή της ΕΔ στην ανάλυση πολύ μεγάλων κατηγορικών συνόλων δεδομένων. Οι παραδοσιακές τεχνικές ανάλυσης κατηγορικών δεδομένων είναι μέσω **πινάκων συνάφειας** (*contingency tables*) και **λογιστικής παλινδρόμησης** (*logistic regression*). Στους πίνακες συνάφειας συγκεντρώνεται και καταγράφεται η αντιστοίχηση της συμπεριφοράς των ατόμων με βάση δύο ή περισσότερες μεταβλητές, ενώ η λογιστική παλινδρόμηση εκφράζει την επιρροή ενός συνόλου συνεχών ή / και κατηγορικών μεταβλητών πάνω σε μια κατηγορική αποκριτική μεταβλητή.

Η λογιστική παλινδρόμηση, καθώς και η πλειοψηφία των μοντέλων πινάκων συνάφειας αποτελούν μέλη της οικογένειας των **γενικευμένων γραμμικών μοντέλων** (*Generalized Linear Models – GLM*). Τα γενικευμένα γραμμικά μοντέλα παρουσιάστηκαν το 1972 από τους Nelder και Wedderburn. Στόχος ήταν η ενοποίηση διάφορων μοντέλων για κατηγορικά δεδομένα και η δημιουργία κάποιων νέων.

Η ανάπτυξη των GLM έδωσε νέα ώθηση στις δυνατότητες ανάλυσης κατηγορικών δεδομένων, μιας και στην οικογένεια αυτή εμπεριέχονται πολλά κλασσικά γραμμικά μοντέλα, καθώς και η εκτίμηση και οι έλεγχοι των μοντέλων αυτών. Ένα χαρακτηριστικό μοντέλο αυτής της ομάδας είναι η **παλινδρόμηση Poisson** (*Poisson regression*), με βάση την οποία μοντελοποιείται το αναμενόμενο πλήθος επιτυχιών ή αποτυχιών ενός πλήθους ερμηνευτικών μεταβλητών (*predictor variables*) μέσω μιας συνάρτησης παλινδρόμησης.

Άλλες ομάδες μοντέλων είναι τα **γενικευμένα προσθετικά μοντέλα** (*Generalized Additive models*) και τα **γραφικά μοντέλα** (*Graphical Models*), τα οποία θεωρούνται η «γέφυρα» μεταξύ της πολυμεταβλητής ανάλυσης από τη Στατιστική και τομέων όπως τεχνητή νοημοσύνη, αιτιολογική ανάλυση (*causal analysis*) και ΕΔ (Glymour et al., 1997).

Μια κατατοπιστική και ουσιαστική παρουσίαση των μοντέλων για κατηγορικά δεδομένα μέσω της οικογένειας των γενικευμένων γραμμικών μοντέλων δίνεται στα βιβλία του Agresti (2002, 2007), τα οποία αποτελούν τις σημαντικότερες εφαρμογές στην ανάλυση κατηγορικών δεδομένων.

Επίσης, αναφέρουμε ενδεικτικά κάποια κλασσικά συγγράμματα που ασχολούνται με την στατιστική ανάλυση κατηγορικών δεδομένων. Ένα από αυτά είναι των Bishop et al. (1975), καθώς και πιο σύγχρονα βιβλία, όπως των Andersen (2001) και Simonoff (2003).

ΚΕΦΑΛΑΙΟ 4

Προ-επεξεργασία δεδομένων

4.1 Γιατί να προ-επεξεργαστούμε τα δεδομένα;

Ακολουθώντας τη σύγχρονη εποχή και τις ανάγκες που προκύπτουν στη διαχείριση πραγματικών δεδομένων, υπάρχει συχνά η ανάγκη αποθήκευσης μεγάλου όγκου πληροφορίας σε οργανωμένες βάσεις δεδομένων. Πολλές φορές, όμως, οι βάσεις δεδομένων που διαχειρίζονται πραγματικά δεδομένα είναι ευαίσθητες σε θόρυβο (*noise*), ελλείπουσες τιμές (*missing values*) και ασυνεπή δεδομένα (*inconsistent data*). Για το λόγο αυτό, απαιτείται η **προ-επεξεργασία** των δεδομένων (*data preprocessing*) πριν την απόπειρα εξόρυξης χρήσιμων αποτελεσμάτων.

Ο ουσιαστικός στόχος της προ-επεξεργασίας των δεδομένων είναι η συνεισφορά στη βελτίωση της ποιότητας των διαθέσιμων δεδομένων ενός συνόλου δεδομένων, στο οποίο επιθυμούμε να εφαρμόσουμε μεθόδους ΕΔ. Έτσι, η διαδικασία εξόρυξης που θα πραγματοποιηθεί μεταγενέστερα, θα είναι περισσότερο ακριβής και αποτελεσματική (Han and Kamber, 2001).

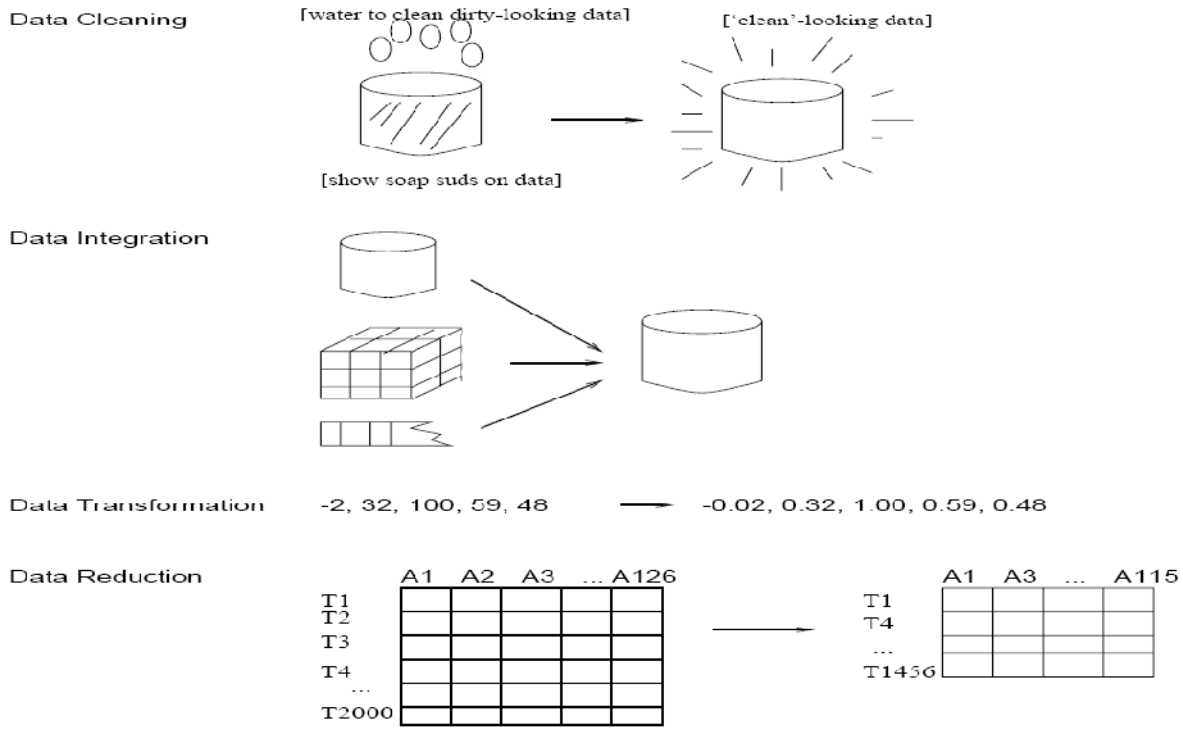
Υπάρχουν αρκετές τεχνικές προ-επεξεργασίας δεδομένων. Οι πιο χαρακτηριστικές είναι ο **καθαρισμός** δεδομένων (*data cleaning*), η **ενοποίηση** και ο **μετασχηματισμός** δεδομένων (*data integration and transformation*), καθώς και η **μείωση** δεδομένων (*data reduction*). Η εφαρμογή του καθαρισμού δεδομένων, οδηγεί στην απομάκρυνση το θορύβου και τη διόρθωση ασυνεπειών που μπορεί να εμφανίζονται στα δεδομένα. Η ενοποίηση δεδομένων συγχωνεύει δεδομένα από πολλαπλές πηγές σε μια συνεκτική (*coherent*) ομάδα δεδομένων, όπως είναι μια αποθήκη δεδομένων (*data warehouse*) ή ένας κύβος δεδομένων (*data cube*).

Επίσης, μιλώντας για μετασχηματισμό δεδομένων, εννοούμε εφαρμογές όπως η κανονικοποίηση δεδομένων (*data normalization*) που βοηθά στην ενιαία αντιμετώπιση και περαιτέρω ανάλυση των δεδομένων. Για παράδειγμα, μπορεί να μας απασχολούν οι διαφορετικές μονάδες μέτρησης. Τέλος, η μείωση των δεδομένων μπορεί να γίνει μέσω συνένωσης, απομάκρυνσης περιττών χαρακτηριστικών, συσταδοποίησης κ.λπ. Για να εμπεδώσουμε τα σχόλια που έγιναν πάνω στις τεχνικές προ-επεξεργασίας δεδομένων, παραθέτουμε το ακόλουθο Σχήμα.

ΣΧΗΜΑ 4.1

Τεχνικές προ-επεξεργασίας δεδομένων

[Πηγή: Han and Kamber, 2001]



Οι εικόνες αυτές είναι χαρακτηριστικές και αποδίδουν την έννοια και τη συνεισφορά κάθε τεχνικής προ-επεξεργασίας δεδομένων. Όσον αφορά την μείωση δεδομένων, που είναι η τελευταία τεχνική που απεικονίζεται στο Σχήμα 4.1, βλέπουμε ένα υποθετικό σύνολο δεδομένων, όπου T_i με i = 1, …, 2000 τα καταγεγραμμένα αντικείμενα και A_j με j = 1, …, 126 οι μεταβλητές.

Παρατηρούμε ότι, ύστερα από την εφαρμογή μιας μεθόδου μείωσης δεδομένων, έχουμε μείωση τόσο γραμμών όσο και στηλών του αρχικού συνόλου δεδομένων. Δηλαδή, μέσω της μείωσης δεδομένων μπορούμε να έχουμε μείωση γραμμών, στηλών ή και των δύο.

Στις ενότητες που ακολουθούν δίνουμε τα βασικά στοιχεία και τις μεθόδους εφαρμογής κάθε τεχνικής. Η αναφορά μας σε κάθε μέθοδο ή τεχνική δε θα είναι εκτενής. Το κεφάλαιο αυτό παρατίθεται ενημερωτικά, στα πλαίσια συνειδητοποίησης της ανάγκης προ-επεξεργασίας των δεδομένων πριν την ΕΔ, που είναι το κύριο αντικείμενο αυτής της εργασίας.

4.2 Καθαρισμός των δεδομένων

Η ποιότητα ενός συνόλου δεδομένων με πραγματικά δεδομένα βασίζεται σε πολλά ζητήματα. Όμως, ο πιο κρίσιμος παράγοντας είναι η πηγή των δεδομένων. Τα δεδομένα που καταγράφονται από τον πραγματικό κόσμο τείνουν πολλές φορές να είναι μη πλήρη, ασαφή και να περιλαμβάνουν θόρυβο.

Οι μέθοδοι καθαρισμού δεδομένων επιχειρούν να συμπληρώσουν **ελλείπουσες τιμές**, να εξομαλύνουν το **θόρυβο** που δημιουργείται κατά τον εντοπισμό έκτροπων παρατηρήσεων, αλλά και να διορθώσουν πιθανές **ασυνέπειες** στα δεδομένα (Han and Kamber, 2001).

Αξίζει να αναφέρουμε ότι δεν υπάρχει κάποιος κοινά αποδεκτός ορισμός για τον καθαρισμό δεδομένων. Υπάρχουν διάφοροι εναλλακτικοί ορισμοί, ανάλογα με την περιοχή στην οποία εφαρμόζεται η συγκεκριμένη διαδικασία. Οι βασικότεροι τομείς που περιλαμβάνουν τον καθαρισμό δεδομένων ως μέρος των διαδικασιών τους είναι η αποθήκευση δεδομένων (*data warehousing*), η ανακάλυψη γνώσης από βάσεις δεδομένων (*Knowledge Discovery in Databases – KDD*) και η διαχείριση της ποιότητας δεδομένων / πληροφορίας (*data / information quality management*).

Σε αυτή την ενότητα, θα ασχοληθούμε με την καταγραφή των πιο χρήσιμων μεθόδων καθαρισμού δεδομένων και προσδιορισμού των πιθανών λαθών σε ένα σύνολο δεδομένων. Η διαδικασία καθαρισμού αποτελείται από τις εξής φάσεις (Maimon and Rokach, 2005):

i) Ορισμός και εντοπισμός των τύπων λαθών
ii) Αναζήτηση και προσδιορισμός των περιπτώσεων λάθους
iii) Διόρθωση ακάλυπτων λαθών

Οι μέθοδοι που δίνουμε παρακάτω επικεντρώνονται στον προσδιορισμό των λαθών μέσω **στατιστικών εφαρμογών, συσταδοποίησης, προτύπων ή κανόνων συνάφειας**. Εφαρμόζοντας αυτές τις γενικές μεθόδους, είμαστε σε θέση να κατευθύνουμε το πρόβλημα και να επιχειρήσουμε τη λύση του.

Σύμφωνα με τους Maimon and Rokach (2005), η διαδικασία εντοπισμού πιθανών λαθών σε ένα σύνολο δεδομένων με τη χρήση διατακτικών κανόνων συνάφειας (*ordinal association rules*) αποτελείται από τα εξής δύο βήματα:

1. Εύρεση διατακτικών (*ordinal*) κανόνων με ένα ελάχιστο διάστημα c.
2. Προσδιορισμός των αντικειμένων των δεδομένων (*data items*) που σπάνε τους κανόνες και μπορούν να θεωρηθούν πιθανά λάθη (έκτροπες παρατηρήσεις).

Στο Σχήμα 4.2 δίνεται ο σχετικός αλγόριθμος για το πρώτο βήμα. Ακολουθεί συνοπτική ανάλυση των αλγορίθμων, ενώ στο Σχήμα 4.3 παραθέτουμε τον αλγόριθμο για το δεύτερο βήμα.

ΣΧΗΜΑ 4.2

Αλγόριθμος εντοπισμού λαθών μέσω κανόνων συνάφειας – πρώτο βήμα

[Πηγή: Maimon and Rokach, 2005]

```
Algorithm compare items.
    for each record in the data base (1 . . . N)
        normalize or convert data
        for each attribute x in (1 . . . M-1)
            for each attribute y in (x+1 . . . M-1)
                compare the values in x and y
                update the comparisons array
            end for.
        end for.
        output the record with normalized data
    end for.
    Output the camparisons array
end algorithm.
```

Με βάση τον αλγόριθμο, αρχικά γίνεται κανονικοποίηση των δεδομένων, εάν αυτό απαιτείται. Στη συνέχεια, γίνεται σύγκριση ανά ζεύγος μεταβλητών για κάθε εγγραφή, ενώ

απαιτείται μόνο ένας έλεγχος των δεδομένων. Τα αποτελέσματα των συγκρίσεων αποθηκεύονται σε ένα διάνυσμα.

Στη συνέχεια (Σχήμα 4.3) παραθέτουμε τον αλγόριθμο για το δεύτερο βήμα της διαδικασίας. Στο βήμα αυτό γίνεται εξαγωγή και αποθήκευση των δεδομένων που σχετίζονται με τους κανόνες. Ύστερα, για κάθε εγγραφή, ελέγχεται κάθε ζεύγος μεταβλητών που αντιστοιχεί σε ένα πρότυπο για να εξασφαλιστεί ότι οι τιμές των αντίστοιχων πεδίων είναι όμοιες με τη σχέση που δηλώνεται από το πρότυπο. Εάν δεν είναι, τότε κάθε πεδίο σημειώνεται ως πιθανό λάθος.

Στις περισσότερες περιπτώσεις, μόνο μια από τις δύο τιμές θα είναι πραγματικά λάθος. Στη συνέχεια υπολογίζεται ο μέσος αριθμός των πιθανών λαθών. Όσα πεδία έχουν σημειωθεί ως «πιθανώς λανθασμένα» περισσότερες φορές από το μέσο όρο χαρακτηρίζονται τελικά ως τα πιθανότερα λανθασμένα.

ΣΧΗΜΑ 4.3

Αλγόριθμος εντοπισμού λαθών μέσω κανόνων συνάφειας – δεύτερο βήμα

[Πηγή: Maimon and Rokach, 2005]

```
Algorithm analyze records.
    for each record in the data base (1N)
        for each rule in the pattern array
            determine rule type and pairs
            compare item pairs
            if item NOT holds
                then mark each item as possible error
            end for.
        compute average number of marks
        select the high probability marked errors
    end for.
end algorithm.
```

Τα «βρώμικα» δεδομένα μπορεί να προκαλέσουν σύγχυση στη διαδικασία εξόρυξης γνώσης. Βέβαια, σε πολλούς αλγορίθμους ΕΔ περιλαμβάνονται διαδικασίες ενασχόλησης με ασυνεπή δεδομένα ή δεδομένα με θόρυβο, αλλά είναι πιθανό να μη μπορεί να εκτελεστεί ο συγκεκριμένος αλγόριθμος ΕΔ που έχει επιλεγεί, λόγω προϋποθέσεων (βλ. Han and Kamber, 2001). Για παράδειγμα, ένας αλγόριθμος δεν είναι πάντα εύρωστος (*robust*). Έτσι, είναι πιθανό να προκύψουν δυσκολίες κατά την εφαρμογή του.

Συνεπώς, καταλαβαίνουμε ότι η διαδικασία του καθαρισμού δεδομένων και γενικότερα το στάδιο προ-επεξεργασίας των δεδομένων στην αλυσίδα KDD είναι ιδιαίτερα σημαντικό, καθώς μπορεί να εξυπηρετήσει στην ενδεχόμενη χρήση ενός αλγορίθμου.

4.2.1 Ελλείπουσες τιμές

Πολλές φορές, όταν καταγράφουμε διάφορα στοιχεία για ένα άτομο, υπάρχει περίπτωση να συναντήσουμε έλλειψη καταχώρησης σε κάποιες μεταβλητές. Για παράδειγμα, μπορεί να λείπουν εγγραφές για το εισόδημα κάποιων πελατών. Τι θα κάναμε σε αυτή την περίπτωση;

Υπάρχουν διάφορες μέθοδοι αντιμετώπισης αυτού του ζητήματος. Στόχος είναι η συμπλήρωση των ελλειπουσών τιμών. Από τις λιγότερο αποτελεσματικές μεθόδους θεωρούνται η εξαίρεση των διανυσμάτων των γνωρισμάτων (*tuples*) όπου συναντώνται ελλείπουσες τιμές από τη βάση δεδομένων ή η εισαγωγή τιμών από τον ερευνητή (*manually*). Επίσης, μπορεί να χρησιμοποιηθεί μια κοινώς αποδεκτή φράση ή τιμή, όπως “Unknown” ή $-\infty$, κίνηση που επίσης δεν προτείνεται.

Πέρα από αυτά, υπάρχουν μέθοδοι, οι οποίες θεωρούνται πιο αποτελεσματικές σε σχέση με τη συμπλήρωση των ελλειπουσών τιμών. Κάποιες από αυτές (Han and Kamber, 2001) είναι:

- Χρήση της **μέσης τιμής** της μεταβλητής ώστε να συμπληρωθεί η ελλίπουσα τιμή της μεταβλητής στη συγκεκριμένη παρατήρηση.
- Χρήση της **μέσης τιμής** της μεταβλητής, για όλα τα δείγματα της **ίδιας κατηγορίας**, με βάση το δεδομένο διάνυσμα γνωρισμάτων. Δηλαδή, αν υποθέσουμε ότι ταξινομούμε τα αντικείμενα μιας βάσης δεδομένων με βάση τις κατηγορίες μιας εκ των μεταβλητών, τότε αντικαθιστούμε την ελλείπουσα παρατήρηση με τη μέση τιμή της κατηγορίας όπου ανήκει.

- Χρήση της πιο **πιθανής τιμής** ώστε να καλυφθεί η ελλίπουσα τιμή. Η πιο πιθανή τιμή μπορεί να εκτιμηθεί μέσω εργαλείων που βασίζονται στη στατιστική συμπερασματολογία και χρησιμοποιούν τύπους του Bayes, ή μέσω της κατασκευής ενός δέντρου απόφασης. Για παράδειγμα, χρησιμοποιώντας τα χαρακτηριστικά των πελατών από ένα σύνολο δεδομένων, μπορούμε να κατασκευάσουμε ένα δέντρο απόφασης (βλ. Han and Kamber, 2001) και να προβλέψουμε για ένα γνώρισμα τις ελλείπουσες τιμές πελατών που καταγράφονται σε αυτό το σύνολο δεδομένων.

Οι μέθοδοι χρήσης μιας «κοινώς αποδεκτής» τιμής, οι δύο διαφορετικές προτάσεις που στηρίζονται στη «μέση τιμή», καθώς και η χρήση της «πιο πιθανής» τιμής θεωρούνται μεροληπτικές. Όμως, η τελευταία θεωρείται η καλύτερη στρατηγική, καθώς χρησιμοποιεί τη μέγιστη πληροφορία από τα υπάρχοντα δεδομένα, ώστε να προβλεφθούν οι ελλείπουσες τιμές.

4.2.2 Δεδομένα με θόρυβο

Μιλώντας για θόρυβο εννοούμε ένα τυχαίο σφάλμα ή διακύμανση σε μια μετρήσιμη μεταβλητή (*measured variable*). Με βάση τον Bramer (2007), θόρυβος είναι η τιμή ενός χαρακτηριστικού η οποία μπορεί να θεωρείται έγκυρη για ένα δεδομένο σύνολο δεδομένων, αλλά έχει καταγραφεί λανθασμένα.

Δοθέντος ενός αριθμητικού χαρακτηριστικού, όπως για παράδειγμα η «τιμή», μπορούμε να «εξομαλύνουμε» τα δεδομένα ώστε να απομακρυνθεί ο θόρυβος. Οι πιο χαρακτηριστικές μέθοδοι εξομάλυνσης (Han and Kamber, 2001) είναι:

- **Μέθοδοι Binning**

  Οι μέθοδοι αυτές εξομαλύνουν μια διατεταγμένη τιμή με βάση τη «γειτονικότητα» ή τις γύρω τιμές. Οι διατεταγμένες τιμές κατανέμονται σε ένα πλήθος (συνήθως ίσων) «ομάδων» (*bins*). Η εξομάλυνση που πραγματοποιείται είναι τοπική, καθώς στηριζόμαστε στη γειτονικότητα των τιμών των δεδομένων.

  Οι βασικότερες μέθοδοι binning είναι η εξομάλυνση με βάση τη μέση τιμή ανά bin (*smoothing by bin means*), ή με βάση τη διάμεσο (*smoothing by bin medians*) ή τα σύνορα του bin (*smoothing by bin boundaries*), δηλαδή τη μέγιστη και την ελάχιστη τιμή. Σε αυτή τη μέθοδο, κάθε τιμή αντικαθίσταται από την κοντινότερη οριακή τιμή (σύνορο) ανά bin.

Στον Πίνακα 4.1, που ακολουθεί, βλέπουμε μια απλή εφαρμογή εξομάλυνσης των δεδομένων μέσω μεθόδων binning.

ΠΙΝΑΚΑΣ 4.1

Εξομάλυνση δεδομένων μέσω μεθόδων binning

Διατεταγμένα δεδομένα (τιμές σε €)	5, 6, 10, 12, 15, 24, 29, 29, 35		
	Bin 1	**Bin 2**	**Bin 3**
Διαμέριση σε bins (ίσου εύρους)	5, 6, 10	12, 15, 24	29, 29, 35
Εξομάλυνση με βάση τη **μέση τιμή** ανά bin	7, 7, 7	17, 17, 17	31, 31, 31
Εξομάλυνση με βάση τα **σύνορα** του bin	5, 5, 10	12, 12, 24	29, 29, 35

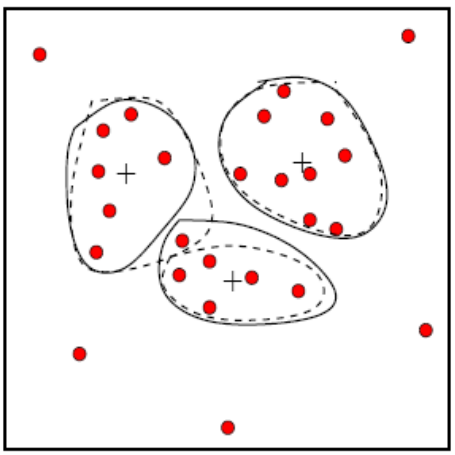
- **Συσταδοποίηση**

Οι έκτροπες παρατηρήσεις μπορούν να προσδιοριστούν μέσω της συσταδοποίησης. Οι όμοιες τιμές οργανώνονται σε ομάδες ή συστάδες. Παρατηρώντας ένα σχήμα σαν το ακόλουθο (βλ. Σχήμα 4.4), κρίνουμε με βάση τη διαίσθηση και την εμπειρία ότι οι τιμές που δεν τοποθετούνται σε καμμία συστάδα μπορούν να θεωρηθούν έκτροπες.

ΣΧΗΜΑ 4.4

Προσδιορισμός έκτροπων παρατηρήσεων μέσω συσταδοποίησης

[Πηγή: Han and Kamber, 2001]



- **Παλινδρόμηση**

  Η εξομάλυνση των δεδομένων μπορεί να γίνει μέσω της προσαρμογής τους σε κάποιο στατιστικό μοντέλο, όπως για παράδειγμα την παλινδρόμηση. Μπορούμε να εφαρμόσουμε απλή ή πολλαπλή γραμμική παλινδρόμηση, ώστε να προβλέψουμε κάποιο αποτέλεσμα. Η χρήση της παλινδρόμησης με στόχο την εύρεση συναρτησιακής σχέσης που να ταιριάζει στα δεδομένα βοηθά στην εξομάλυνση του θορύβου.

- **Συνδυασμένη επιθεώρηση από Η/Υ και υποκείμενο**

  Είναι σίγουρο ότι ο συνδυασμός της υποκειμενικής ανθρώπινης κρίσης και των υπολογιστικών συστημάτων μπορεί να προσφέρει τον προσδιορισμό των έκτροπων παρατηρήσεων. Η συμβολή του υποκειμενικού παράγοντα είναι στη διάκριση μεταξύ της χρήσιμης ή μη πληροφορίας που μπορεί να εξάγει ο υπολογιστής.

Ορισμένες μέθοδοι εξομάλυνσης δεδομένων, όπως οι μέθοδοι binning που παρατέθηκαν στην αρχή, θεωρούνται παράλληλα και μέθοδοι μείωσης δεδομένων, συμπεριλαμβανομένης της διακριτοποίησης. Οι μέθοδοι binning συμβάλλουν, για παράδειγμα, στη μείωση των διάφορων τιμών ανά χαρακτηριστικό. Οι τεχνικές μείωσης και η διακριτοποίηση θα σχολιαστούν εκτενέστερα παρακάτω.

4.2.3 Ασυνεπή δεδομένα

Πολλές φορές, υπάρχουν ασυνέπειες σε δεδομένα, όταν καταγράφονται κάποιες συναλλαγές. Η αντιμετώπισή τους μπορεί να γίνει προσωπικά από το χρήστη (*manually*) ή μέσω ορισμού αυτοματοποιημένων διαδικασιών. Για παράδειγμα, τα λάθη που δημιουργούνται κατά την εισαγωγή των δεδομένων μπορούν να διορθωθούν με τη χρήση ενός πλάνου δράσης (*paper trace*). Επίσης, η χρήση εργαλείων μηχανικής της γνώσης (*knowledge engineering tools*) μπορεί να προσδιορίσει την παραβίαση των περιορισμών στα δεδομένα.

Τέλος, είναι πιθανό να παρατηρηθούν ασυνέπειες στα δεδομένα λόγω της ενοποίησης που μπορεί να έχει πραγματοποιηθεί (*data integration*). Από τη διαδικασία αυτή, μπορεί να προκύψει ένα χαρακτηριστικό με διαφορετικές τιμές σε άλλες βάσεις δεδομένων. Ένα άλλο φαινόμενο μπορεί να είναι ο πλεονασμός (*redundancy*) χαρακτηριστικών. Παρακάτω γίνονται σχόλια για αυτά τα προβλήματα.

4.3 Ενοποίηση και μετασχηματισμός δεδομένων

Όπως είπαμε και στην αρχή του κεφαλαίου, μέσω της ενοποίησης των δεδομένων επιτυγχάνουμε τη **συγχώνευση** των δεδομένων σε μια «αποθήκη», η οποία επιθυμούμε να χαρακτηρίζεται από συνοχή και να έχει λογική δομή (Han and Kamber, 2001). Στα πλαίσια της ενοποίησης, μπορεί να μας απασχολήσουν θέματα όπως η ενοποίηση βασισμένη σε σχήμα (*schema integration*), ή ο πλεονασμός.

Ένα χαρακτηριστικό μπορεί να είναι πλεονάζων, εάν μπορεί να παραχθεί και από έναν άλλο πίνακα, όπως για παράδειγμα το ετήσιο εισόδημα. Επίσης, πλεονασμός μπορεί να προκληθεί από ασυνέπειες στην ονομασία χαρακτηριστικών ή διαστάσεων. Ο πλεονασμός μπορεί να προσδιοριστεί από την **ανάλυση συσχετίσεων** (*correlation analysis*).

Τέλος, ένα ακόμη θέμα που μας απασχολεί στην ενοποίηση δεδομένων είναι ο προσδιορισμός και η απόφαση για τις τιμές δεδομένων, όταν προκύψει σύγκρουση. Για παράδειγμα, οι τιμές ενός χαρακτηριστικού της ίδιας οντότητας μπορεί να διαφέρουν επειδή προέρχονται από διαφορετικές πηγές. Αυτή η διαφορά μπορεί να οφείλεται σε αντιθέσεις στην παρουσίαση, διάταξη ή κωδικοποίηση των τιμών (Han and Kamber, 2001).

Η προσεκτική ενοποίηση των δεδομένων από πολλαπλές πηγές μπορεί να βοηθήσει στη μείωση και αποφυγή ασυνεπειών και πλεονασμού στο προκύπτων σύνολο δεδομένων. Έτσι, θα έχουμε βελτιωμένη ακρίβεια και ταχύτητα για τη διαδικασία εξόρυξης που θα ακολουθήσει.

Όσον αφορά το **μετασχηματισμό** δεδομένων, τα δεδομένα μετασχηματίζονται ή παγιώνονται σε τύπους κατάλληλους για εξόρυξη. Οι βασικότεροι τύποι μετασχηματισμού είναι η **κανονικοποίηση** (*normalization*), η **εξομάλυνση** (*smoothing*), η **ενοποίηση** (*aggregation*) ή η **γενίκευση** (*generalization*).

4.4 Μείωση των δεδομένων

Στις περιπτώσεις όπου χρειάζεται να συγκεντρωθούν δεδομένα σε μια βάση δεδομένων, είναι πλέον σύνηθες φαινόμενο η δημιουργία τεράστιων συνόλων δεδομένων. Μια περίπλοκη διαδικασία ανάλυσης ή εξόρυξης σε πολύ μεγάλο όγκο δεδομένων είναι πιθανότατα χρονοβόρα και μη αποτελεσματική.

Για το λόγο αυτό, απαιτείται συχνά η μείωση του μεγέθους του συνόλου δεδομένων, χωρίς όμως να κινδυνεύσουν τα αποτελέσματα της ΕΔ. Ας δούμε όμως με ποιους τρόπους μπορεί να επιτευχθεί αυτό.

4.4.1 Μέθοδοι μείωσης δεδομένων

Οι τεχνικές μείωσης δεδομένων οδηγούν σε πιο ευέλικτη παρουσίαση των δεδομένων, καθώς μειώνεται κατά πολύ ο όγκος, χωρίς να χάνεται η ακεραιότητα των αρχικών δεδομένων.

Οι βασικότερες τεχνικές μείωσης δεδομένων, σύμφωνα με τους Han και Kamber (2001), είναι:

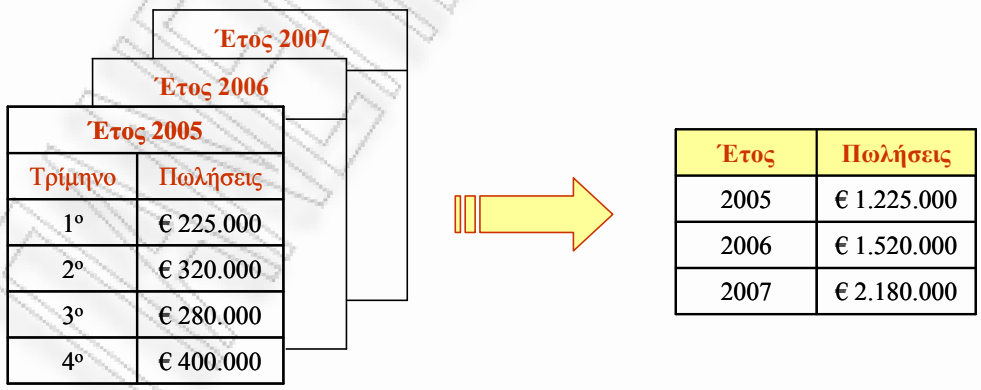
1. Ενοποίηση δεδομένων σε κύβο (*Data cube aggregation*)

Ας υποθέσουμε ότι έχουμε συλλέξει τα δεδομένα για την ανάλυση που θέλουμε να πραγματοποιήσουμε. Έστω ότι έχουμε στοιχεία για τις πωλήσεις μιας εταιρείας (σε ευρώ) ανά τρίμηνο, για τα έτη 2005 έως και 2007.

Όμως, θεωρούμε ότι στην ανάλυση αυτή μας απασχολούν οι ετήσιες πωλήσεις συνολικά. Τότε, μπορούμε να προχωρήσουμε σε **ενοποίηση**, ώστε να έχουμε συνοπτικά τα στοιχεία που χρειαζόμαστε. Το Σχήμα 4.5, που παραθέτουμε αμέσως μετά, εκφράζει την έννοια της συσσωμάτωσης, καθώς βρίσκουμε τις πωλήσεις ανά έτος.

ΣΧΗΜΑ 4.5

Διαδικασία ενοποίησης



Το προκύπτων σύνολο δεδομένων (βλ. Σχήμα 4.5) έχει σαφώς μικρότερο μέγεθος, χωρίς να υπάρχει απώλεια της απαραίτητης πληροφορίας. Άλλωστε, στο παράδειγμα αυτό, δεν ενδιαφερόμαστε για τις πωλήσεις ανά τρίμηνο.

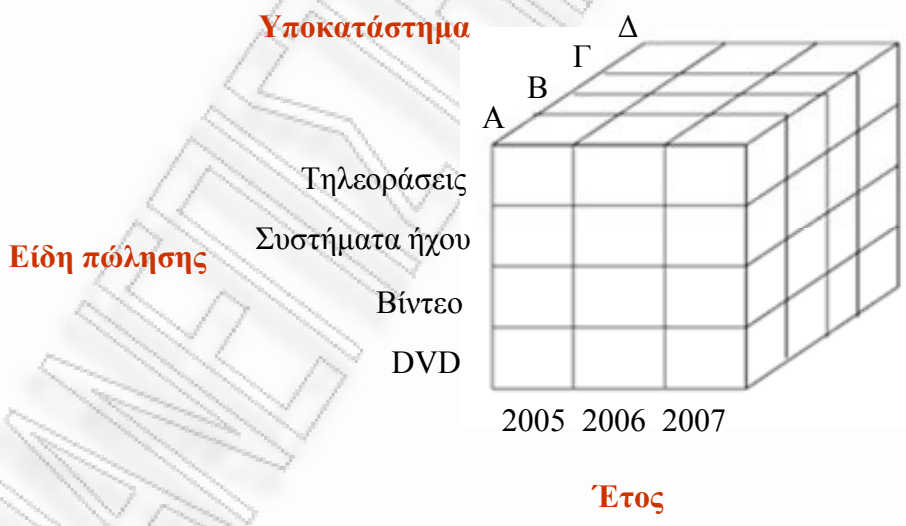
Η συνεισφορά ενός **κύβου δεδομένων** (*data cube*) σε μια τέτοια ανάλυση είναι στην αποθήκευση πολυδιάστατης ενοποιημένης πληροφορίας. Για παράδειγμα, μπορεί να μας ενδιαφέρουν οι ετήσιες πωλήσεις για τα έτη αυτά, ανά αντικείμενο (*item type*) που πουλά η εταιρεία, αλλά και ανά υποκατάστημα (*branch*).

Έτσι, θα μπορούσαμε να έχουμε ένα κύβο σαν αυτό του Σχήματος 4.6, όπου υποθέτουμε ότι έχουμε συλλέξει δεδομένα για τις πωλήσεις της εταιρείας, από τα τρία πιο πρόσφατα έτη. Από τα είδη που πουλά η εταιρεία, μας ενδιαφέρει να δούμε τις πωλήσεις των συστημάτων ήχου και εικόνας και πιο συγκεκριμένα των τηλεοράσεων, ηχοσυστημάτων, βίντεο και DVD. Θεωρούμε ότι η εταιρεία έχει τέσσερα υποκαταστήματα, έστω Α, Β, Γ και Δ.

Να σημειώσουμε ότι κάθε κελί του κύβου δεδομένων (βλ. Σχήμα 4.6) αποτελεί μια ενοποιημένη τιμή.

ΣΧΗΜΑ 4.6

Κύβος δεδομένων



Για κάθε χαρακτηριστικό, μπορεί να υπάρχουν «ιεραρχίες», οι οποίες θα επιτρέπουν την ανάλυση των δεδομένων σε πολλαπλά αποσπασματικά επίπεδα. Για παράδειγμα, μια ιεραρχία για τα υποκαταστήματα (βλ. Σχήμα 4.6) θα επέτρεπε την ομαδοποίησή τους σε περιοχές, ανάλογα με τη διεύθυνση.

Στο ζήτημα αυτό θα γίνει εκτενέστερη αναφορά, σε σχετική υποενότητα, όπου γίνεται και διαχωρισμός με βάση τον τύπο δεδομένων (συνεχή / κατηγορικά). Οι κύβοι δεδομένων παρέχουν ταχύτερη πρόσβαση σε ήδη υπολογισμένα ή ενοποιημένα δεδομένα και βελτίωση στις διαδικασίες ανάλυσης και εξόρυξης.

2. Μείωση διαστάσεων (*Dimension reduction*)

Μέσω αυτής της μεθόδου, επιχειρούμε να προσδιορίσουμε και να απομακρύνουμε χαρακτηριστικά ή διαστάσεις που δε σχετίζονται μεταξύ τους, ή εμφανίζουν ελαφρά συσχέτιση ή και πλεονασμό. Έτσι, οδηγούμαστε στην ελαχιστοποίηση του μεγέθους του δεδομένου συνόλου δεδομένων, αλλά και στη δημιουργία περισσότερο κατανοητών προτύπων.

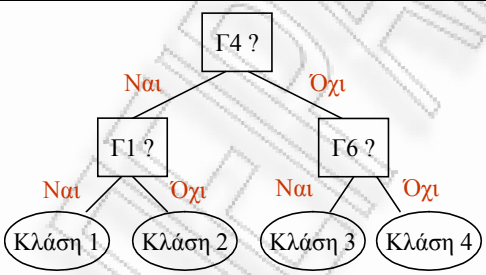
Υπάρχουν αρκετές ευρετικές (*heuristic*) μέθοδοι μείωσης διαστάσεων, όπως η κατά βήματα **«προς τα εμπρός» επιλογή** ή **«προς τα πίσω» απαλοιφή** (*step-wise forward selection or backward elimination*), ή ο συνδυασμός αυτών των δύο. Στην «προς τα εμπρός» επιλογή, ξεκινάμε με ένα άδειο σύνολο χαρακτηριστικών και προσθέτουμε όσο προχωράμε σε βήματα, ενώ στην «προς τα πίσω» απαλοιφή ξεκινάμε με το πλήρες σύνολο χαρακτηριστικών και αφαιρούμε αυτές που δε χρειαζόμαστε.

Οι δύο αυτές μέθοδοι απεικονίζονται στον Πίνακα 4.2, που παρατίθεται στην επόμενη σελίδα. Στην περίπτωση που θέλουμε να πραγματοποιήσουμε το συνδυασμό των μεθόδων αυτών, τότε ασχολούμαστε σε κάθε βήμα με τα εναπομείναντα χαρακτηριστικά και από αυτά επιλέγουμε το βέλτιστο και απομακρύνουμε το χειρότερο. Όλες αυτές οι μεθοδοι ποικίλουν στα κριτήρια τερματισμού.

Στον Πίνακα 4.2, βλέπουμε και άλλη μια μέθοδο, αυτή της δημιουργίας **δέντρου αποφάσεων** (*decision tree induction*). Υπάρχουν αλγόριθμοι κατασκευής δέντρων, όπως οι ID3 και C4.5, οι οποίοι αρχικά είχαν σχεδιαστεί για την ταξινόμηση (*classification*).

ΠΙΝΑΚΑΣ 4.2

Ευρετικές μέθοδοι επιλογής χαρακτηριστικών

<table>
<tr><td colspan="3">Αρχικό σύνολο γνωρισμάτων: {Γ1, Γ2, Γ3, Γ4, Γ5, Γ6}</td></tr>
<tr><td>«Προς τα εμπρός» επιλογή</td><td>«Προς τα πίσω» απαλοιφή</td><td>Δέντρο αποφάσεων</td></tr>
<tr><td>{ }</td><td rowspan="2">{Γ1, Γ3, Γ4, Γ5, Γ6}</td><td rowspan="3"> 
</td></tr>
<tr><td>{Γ1}</td></tr>
<tr><td>{Γ1, Γ4}</td><td>{Γ1, Γ4, Γ5, Γ6}</td></tr>
<tr><td colspan="3">Μειωμένο σύνολο γνωρισμάτων</td></tr>
<tr><td>{Γ1, Γ4, Γ6}</td><td>{Γ1, Γ4, Γ6}</td><td>{Γ1, Γ4, Γ6}</td></tr>
</table>

Σε κάθε κόμβο του δέντρου αποφάσεων (βλ. Πίνακα 4.2) γίνεται ένας έλεγχος χαρακτηριστικού και κάθε κλαδί δίνει ένα αποτέλεσμα. Επίσης, στους εξωτερικούς κόμβους δίνονται οι προβλέψεις των ομάδων, ενώ σε κάθε κόμβο, ο αλγόριθμος διαμερίζει τα δεδομένα σε μεμονωμένες κλάσεις, με βάση το βέλτιστο χαρακτηριστικό.

3. Συμπίεση δεδομένων (*Data compression*)

Στη συμπίεση δεδομένων, εφαρμόζονται κωδικοποιήσεις ή μετασχηματισμοί δεδομένων, με σκοπό να έχουμε μείωση ή «συμπίεση» των αρχικών δεδομένων. Βέβαια, το επιθυμητό είναι να μπορούμε να επιστρέψουμε μέσω ανακατασκευής από τα συμπιεσμένα δεδομένα στα αρχικά. Τότε, θεωρούμε ότι η μέθοδος συμπίεσης που μπορεί να εφαρμοστεί δεν οδηγεί σε απώλεια πληροφορίας. Όμως, οι μέθοδοι αυτές έχουν περιορισμένες δυνατότητες χειρισμού των δεδομένων.

Για το λόγο αυτό, προτιμούμε τη χρήση μεθόδων που μπορεί να χάνουν περισσότερη πληροφορία σε σχέση με τις προηγούμενες, αλλά είναι καλύτερα συντονισμένες. Οι πιο γνωστές μέθοδοι αυτού του τύπου είναι ο **κυματοειδής μετασχηματισμός** (*wavelet transforms*) και η **ανάλυση κυρίων συνιστωσών** (*principal component analysis*)

4. Μείωση πλήθους δεδομένων (*Numerosity reduction*)

Ένα θέμα που τίθεται συχνά είναι εάν μπορούμε να μειώσουμε τον όγκο των δεδομένων επιλέγοντας εναλλακτικούς ή πιο σύντομους τρόπους αντιπροσώπευσης. Για το λόγο αυτό, δημιουργήθηκαν οι μέθοδοι μείωσης πλήθους δεδομένων, κάποιες εκ των οποίων είναι παραμετρικές και κάποιες άλλες όχι.

Στις παραμετρικές μεθόδους αυτής της ομάδας, χρησιμοποιείται ένα μοντέλο ώστε να γίνει εκτίμηση των δεδομένων και αποθηκεύονται μόνο οι εκτιμήσεις των δεδομένων αντί για τα πλήρη πραγματικά δεδομένα. Επιπλέον, αποθηκεύονται και οι έκτροπες παρατηρήσεις. Ένα παράδειγμα είναι τα **λογαριθμογραμμικά μοντέλα** (*loglinear models*), τα οποία εκτιμούν διακριτές πολυδιάστατες κατανομές πιθανότητας.

Ως μη παραμετρικές μεθόδους αποθήκευσης των αντιπροσωπεύσεων (μειωμένα δεδομένα) αναφέρουμε τα **ιστογράμματα**, τη **συσταδοποίηση** και τη **δειγματοληψία**.

5. Διακριτοποίηση και παραγωγή ιεραρχικής έννοιας (*Discretization and concept hierarchy generation*)

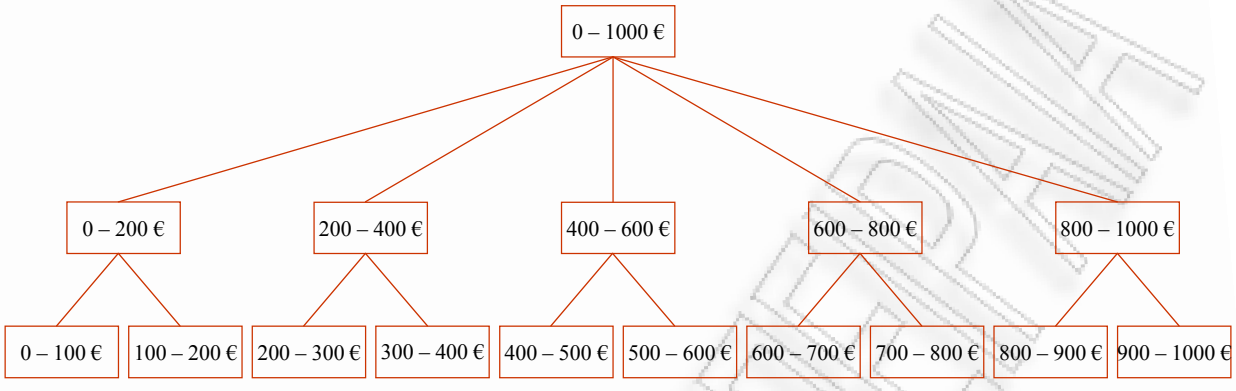
Χρησιμοποιώντας τις μεθόδους διακριτοποίησης, μπορούμε να μειώσουμε τον αριθμό των τιμών μιας δοθείσας συνεχούς μεταβλητής, χωρίζοντάς την σε διαστήματα. Αυτή η εφαρμογή χρησιμοποιείται αρκετά συχνά, όταν πρόκειται να διενεργηθεί μια μεγάλη ανάλυση, καθώς οδηγεί και σε πιο κατανοητά αποτελέσματα.

Οι τεχνικές διακριτοποίησης είναι συχνά επαναληπτικές, ενώ θεωρείται ιδιαίτερα χρονοβόρα η διάταξη των δεδομένων κάθε φορά. Πολλές τεχνικές αυτού του τύπου μπορούν να χρησιμοποιηθούν πολλές φορές ώστε να οδηγήσουν σε μια ιεραρχική ή πολυτμηματική διαμέριση των τιμών της μεταβλητής. Η διαμέριση αυτή είναι γνωστή ως μια ιεραρχία έννοιας ή εννοιολογική ιεραρχία (*concept hierarchy*).

Ας δούμε για παράδειγμα, στο Σχήμα 4.7, πως θα μπορούσαμε να ορίσουμε μια έννοια ιεραρχίας για τη μεταβλητή «τιμή πώλησης σε €». Έστω ότι οι τιμές για τα προϊόντα μιας εταιρείας κυμαίνονται από 0 έως 1000 €. Για τη διενέργεια μιας ανάλυσης, είναι πιο ενδιαφέρον να αντιμετωπίσουμε αυτή τη συνεχή μεταβλητή με βάση κάποιες κατηγορίες της και να δημιουργήσουμε υποκατηγορίες, αν κριθεί απαραίτητο.

ΣΧΗΜΑ 4.7

Παραγωγή εννοιολογικής ιεραρχίας



Στην υποενότητα που ακολουθεί εκφράζουμε την ιδέα της ιεραρχίας εννοιών και διαχωρίζουμε την εφαρμογή της σε συνεχή και κατηγορικά δεδομένα.

4.4.2 Παραγωγή εννοιολογικής ιεραρχίας σε συνεχή και κατηγορικά δεδομένα

Όπως είπαμε και προηγουμένως, η παραγωγή της ιεραρχίας εννοιών οδηγεί στη διαμέριση των τιμών ενός χαρακτηριστικού (βλ. Σχήμα 4.7). Συχνά, η τεχνική αυτή είναι αρκετά χρήσιμη. Όσον αφορά τα συνεχή δεδομένα, μια ιεραρχία έννοιας για ένα δοθέν χαρακτηριστικό ορίζει μια διακριτοποίηση του χαρακτηριστικού.

Οι εννοιολογικές ιεραρχίες μπορούν να χρησιμοποιηθούν ώστε να καταλήξουμε σε μείωση των δεδομένων. Αυτό επιτυγχάνεται μέσα από τη συλλογή και αντικατάσταση των χαμηλότερου επιπέδου εννοιών, όπως είναι για παράδειγμα κάποιες μεμονωμένες αριθμητικές τιμές για τη μεταβλητή «ηλικία». Δηλαδή, θα ήταν καλύτερα εάν μεταβλητής αυτή παρουσιαζόταν μέσω εννοιών υψηλότερου επιπέδου, όπως «νέος», «μεσήλικας» και «ηλικιωμένος», οι οποίες θα ήταν και πιο χρήσιμες.

Με αυτή τη γενίκευση, είναι σίγουρο ότι χάνεται ένα μέρος της πληροφορίας, αλλά θα έχουμε πιο άμεσα αποτελέσματα κατά τη διεξαγωγή της έρευνας. Η παραγωγή ιεραρχίας εννοιών σε συνεχή χαρακτηριστικά γίνεται αυτόματα, με βάση την κατανομή των δεδομένων.

Οι πιο σημαντικές μέθοδοι παραγωγής εννοιολογικών ιεραρχιών σε συνεχή δεδομένα είναι:

- Η διαδικασία **binning**
- Η **ανάλυση ιστογράμματος** (*histogram analysis*)

- Η **ανάλυση συσταδοποίησης** (*clustering analysis*)
- Η **διακριτοποίηση** βασιζόμενη στην **εντροπία** (*entropy-based discretization*)
- Η **τμηματοποίηση** μέσω φυσικής διαμέρισης (*segmentation by natural partitioning*)

Στο βιβλίο των Han και Kamber (2001) γίνεται εκτενής παρουσίαση των μεθόδων αυτών, καθώς και μεθόδων κατάλληλων για κατηγορικά δεδομένα.

Η αυτόματη παραγωγή εννοιολογικών ιεραρχιών για κατηγορικά δεδομένα μπορεί να στηριχτεί στον αριθμό των ευδιάκριτων τιμών της μεταβλητής. Οι τιμές αυτές ορίζουν την ιεραρχία στη μεταβλητή. Η ιεραρχία εκφράζεται από το χρήστη, καθώς συνήθως δεν υπάρχει σαφής διάταξη μεταξύ των τιμών μιας κατηγορικής μεταβλητής.

Οι βασικότερες μέθοδοι παραγωγής ιεραρχίας εννοιών για κατηγορικά δεδομένα (Han and Kamber, 2001) είναι:

- Ο προσδιορισμός από το χρήστη μιας μερικής διάταξης των χαρακτηριστικών, με βάση τα επίπεδα ενός σχήματος (για παράδειγμα: δρόμος < πόλη < νομός)
- Ο ορισμός ενός τμήματος (*portion*) ιεραρχίας από ρητή ομαδοποίηση δεδομένων
- Ο προσδιορισμός ενός συνόλου χαρακτηριστικών, αλλά όχι της μερικής διάταξής τους
- Η επιλογή ενός μερικού συνόλου χαρακτηριστικών

ΚΕΦΑΛΑΙΟ 5

Συσταδοποίηση και αναγνώριση προτύπων

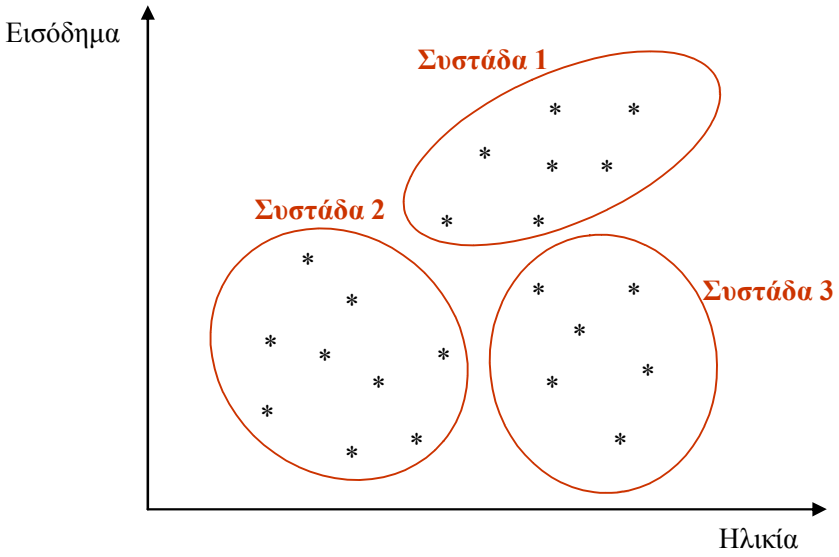
5.1 Εισαγωγικά στοιχεία

Μια από τις πιο σημαντικές διεργασίες στη διαδικασία εξόρυξης γνώσης είναι η **συσταδοποίηση** (*clustering*). Πρόκειται για μια μεθοδολογία ανακάλυψης συστάδων και κατανομών ή προτύπων (*patterns*) που παρουσιάζουν ενδιαφέρον στα υπό μελέτη δεδομένα. Ως **συστάδα** (*cluster*) ορίζεται μια συλλογή αντικειμένων (*objects*) από τα δεδομένα, με βάση τη μεταξύ τους ομοιότητα. Για την ακρίβεια, τα αντικείμενα μιας συστάδας αναμένεται να έχουν όμοια συμπεριφορά μεταξύ τους και ανόμοια σε σχέση με τα αντικείμενα άλλων συστάδων.

Κάθε συστάδα θεωρείται ότι έχει όμοια αντίδραση ως σύνολο σε μια συγκεκριμένη εφαρμογή. Ένα υποθετικό παράδειγμα δίνεται στο Σχήμα 5.1, όπου με βάση τα χαρακτηριστικά ηλικία και εισόδημα, προέκυψαν τρεις συστάδες για το συγκεκριμένο σύνολο δεδομένων. Κάθε δημιουργηθείσα συστάδα έχει τις δικές της «ετικέτες» (*labels*).

ΣΧΗΜΑ 5.1

Παράδειγμα συσταδοποίησης



Οι απαρχές της ανάλυσης κατά συστάδες (*Cluster analysis*) εντοπίζονται από πολύ παλιά, συγκεκριμένα από τη δεκαετία του 1960 (βλ. Anderberg, 1973). Σήμερα, οι διαδικασίες της συσταδοποίησης θεωρούνται απαραίτητο συστατικό για πάρα πολλές εφαρμογές ετερόκλητων κλάδων, όπως η αναγνώριση προτύπων (*pattern recognition*), η ανάλυση δεδομένων, η μεταποίηση εικόνας (*image processing*), η έρευνα αγοράς κ.λπ.

Σε αντίθεση με την ταξινόμηση (*classification*), όπως σχολιάσαμε και στο δεύτερο κεφάλαιο, η συσταδοποίηση δε στηρίζεται σε εκ των προτέρων ορισμένες κλάσεις και εκπαιδευτικά παραδείγματα με διαμορφωμένα χαρακτηριστικά ανά κλάση (*class-labeled training examples*). Αντίθετα, «αφήνει» τα δεδομένα να αυτοπροσδιοριστούν με βάση τις μεταξύ τους ομοιότητες και διαφορές και να καθορίσουν το πλήθος των κλάσεων και τα ποιοτικά τους χαρακτηριστικά.

Ένα προσόν της διαδικασίας συσταδοποίησης είναι ότι παράγει τις αρχικές κατηγορίες στις οποίες οι τιμές ενός συνόλου δεδομένων μπορούν να κατηγοριοποιηθούν κατά την εφαρμογή της ταξινόμησης. Για το λόγο αυτό, στη συγκεκριμένη εργασία επικεντρωνόμαστε στη συσταδοποίηση όσον αφορά το μέρος της ΕΔ από την αλυσίδα KDD (βήματα 5, 6, 7 – υποενότητα 2.1.2). Ορισμένα σχόλια για την εποπτευόμενη μάθηση ύστερα από τη συσταδοποίηση γίνονται στο βιβλίο του Alpaydin (2004).

Στη μηχανική εκμάθηση (*machine learning*) και την αναγνώριση προτύπων, η ανάλυση συστάδων αναφέρεται συχνά ως **μη εποπτευόμενη εκμάθηση** (*unsupervised learning*). Άλλες ονομασίες με τις οποίες μπορεί να συναντήσουμε τη συσταδοποίηση είναι αριθμητική ταξονομία (*numerical taxonomy*) στη βιολογία, τυπολογία (*typology*) στις κοινωνικές επιστήμες και τμηματοποίηση (*partition*) στη θεωρία γράφων (Theodoridis and Koutroumbas, 2003).

5.1.1 Μέτρα απόστασης / ομοιότητας (*Distance / similarity measures*)

Το βασικότερο στοιχείο για να εκτελέσουμε μια επιτυχημένη συσταδοποίηση είναι η ύπαρξη κατάλληλων ποσοτήτων, οι οποίες θα υποδεικνύουν εάν δύο άτομα (παρατηρήσεις) είναι όμοια ή ανόμοια μεταξύ τους. Το χαρακτηριστικό αυτών των ποσοτήτων είναι ότι οι παρατηρήσεις που μοιάζουν πολύ μεταξύ τους, θα πρέπει να δίνουν πολύ μικρή τιμή στην **απόσταση** (Κούτρας, 2007).

Ας υποθέσουμε ένα τυπικό πρόβλημα της πολυμεταβλητής ανάλυσης. Έστω ότι έχουμε ένα δείγμα n ατόμων (αντικειμένων) από ένα πληθυσμό. Για κάθε άτομο, παρατηρούμε p χαρακτηριστικά (τυχαίες μεταβλητές), με $p \geq 2$, έστω τις $X_1, X_2, ..., X_p$. Τότε, θα έχουμε ένα πίνακα $X=(x_{ij})$ διάστασης $n \times p$, όπου n και p το πλήθος των γραμμών και στηλών αντίστοιχα, σαν τον ακόλουθο:

$$\begin{array}{c} \\ \text{n άτομα} \end{array} \begin{array}{c} \text{p μεταβλητές} \\ \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \end{array}$$

Έτσι, για το υποκείμενο i, με i = 1, …, n , έχουμε την p-διάστατη παρατήρηση $x_i = (x_{i1}, x_{i2}, ..., x_{ip})'$. Η **εγγύτητα** (*proximity*) των n υποκειμένων προκύπτει βάσει κάποιας απόστασης. Για την επιλογή του κατάλληλου μέτρου απόστασης λαμβάνουμε υπόψη (βλ. Κατέρη, 2006):

- Τη φύση των μεταβλητών: δηλαδή, εάν τα δεδομένα είναι συνεχή, διακριτά ή δίτιμα
- Την κλίμακα μέτρησης
- Τα ίδια τα δεδομένα και τη φύση τους: για παράδειγμα, η ομοιότητα μπορεί να κρίνεται άλλοτε από την εμφάνιση ενός χαρακτηριστικού και άλλοτε από την απουσία του

Εάν έχουμε n άτομα / αντικείμενα, τότε τοποθετούμε τις αποστάσεις τους, έστω $d_{ij} = d(x_i, x_j)$, σε έναν πίνακα $D = [d_{ij}]$, ο οποίος θα έχει n γραμμές και n στήλες. Ο πίνακας αυτός καλείται **πίνακας αποστάσεων** (ή πίνακας εγγύτητας) των n σημείων. Τα διαγώνια στοιχεία του πίνακα αποστάσεων είναι ίσα με μηδέν και ισχύει $d_{ij} = d_{ji}$, για i≠j.

Σε αυτό το σημείο, είμαστε σε θέση να δώσουμε τις σημαντικότερες εκφράσεις απόστασης, έτσι όπως καταγράφονται στη σχετική βιβλιογραφία (βλ. Johnson and Wichern, 1998). Ο διαχωρισμός μας θα γίνει με βάση τον τύπο δεδομένων.

Αποστάσεις για ζεύγη υποκειμένων με συνεχή δεδομένα

Στον Πίνακα 5.1, αναφέρουμε τις ονομασίες και τους τύπους των σημαντικότερων αποστάσεων που έχουν καταγραφεί για συνεχή δεδομένα. Ακολουθεί σχολιασμός των τύπων.

ΠΙΝΑΚΑΣ 5.1

Αποστάσεις για συνεχή δεδομένα

	Ονομασία	Τύπος
1.	Ευκλείδεια απόσταση	$d(x_i, x_j) = \sqrt{(x_i - x_j)'(x_i - x_j)} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$
2.	Τυποποιημένη ευκλείδεια απόσταση ή απόσταση του Pearson	$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (\frac{x_{ik} - x_{jk}}{s_k})^2}$
3.	Τετραγωνική ευκλείδεια απόσταση	$d(x_i, x_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$
4.	Απόσταση Mahalanobis	$d(x_i, x_j) = \sqrt{(x_i - x_j)'S^{-1}(x_i - x_j)}$
5.	Μετρική Manhattan ή City-block	$d(x_i, x_j) = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$
6.	Μετρική Minkowski	$d(x_i, x_j) = \left[\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^m \right]^{1/m}$
7.	Μετρική Power	$d(x_i, x_j) = \left[\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^m \right]^{1/r}$
8.	Μετρική Canberra	$d(x_i, x_j) = \sum_{k=1}^{p} \frac{\lvert x_{ik} - x_{jk} \rvert}{x_{ik} + x_{jk}}$
9.	Απόσταση max ή Chebyshev	$d(x_i, x_j) = \max_k \lvert x_{ik} - x_{jk} \rvert$
10.	Συντελεστής του Czekanowski	$d(x_i, x_j) = 1 - \frac{2\sum_{k=1}^{p} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{p} (x_{ik} + x_{jk})}$
11.	Στατιστική απόσταση	$d(x_i, x_j) = \sqrt{(x_i - x_j)'A(x_i - x_j)}$

Οι πρώτοι τρεις τύποι του Πίνακα 5.1 αποτελούν τις ευκλείδειες αποστάσεις. Οι αποστάσεις αυτές έχουν εύκολη γεωμετρική ερμηνεία, εξαρτώνται από την κλίμακα μέτρησης, ενώ επηρεάζονται αρκετά από τις έκτροπες παρατηρήσεις (*outliers*). Όταν έχουμε διαφορετική κλίμακα μέτρησης μεταξύ των παρατηρήσεων, ενδείκνυται η χρήση της απόστασης του Pearson (τύπος 2), καθώς γίνονται καλύτερες συγκρίσεις μεταξύ των μεταβλητών (βλ. Κούτρας, 2007).

Όσον αφορά τη μετρική Minkowski (τύπος 6), για την οποία θεωρείται ότι το m είναι δεδομένο, παρατηρούμε ότι για m=1 οδηγούμαστε στην απόσταση city-block (τύπος 5), ενώ για m=2 έχουμε την ευκλείδεια απόσταση (τύπος 1). Οι αποστάσεις 6 και 7 έχουν όμοια αποτελέσματα με την 1, εκτός εάν υπάρχουν έκτροπες παρατηρήσεις.

Η απόσταση Chebyshev (τύπος 9) θεωρεί δύο παρατηρήσεις διαφορετικές εάν διαφέρουν τουλάχιστον σε μια μεταβλητή και αποτελεί ειδική περίπτωση της Minkowski. Επιπλέον, οι τύποι 8 και 10 ισχύουν μόνο για μη αρνητικές μεταβλητές (Johnson and Wichern, 1998).

Επίσης, αναφέρουμε για τη στατιστική απόσταση (τύπος 11) ότι λαμβάνει υπόψη τις συσχετίσεις μεταξύ των μεταβλητών. Για $A=S^{-1}$, που είναι και η συνηθέστερη επιλογή, οδηγούμαστε στην απόσταση Mahalanobis (τύπος 4), όπου επίσης λαμβάνονται υπόψη οι συσχετίσεις μεταξύ των μεταβλητών. Ενώ, όταν ο πίνακας S είναι διαγώνιος η στατιστική απόσταση ταυτίζεται με την τυποποιημένη ευκλείδεια απόσταση.

Με S (τύποι 4 και 10, στην περίπτωση που μπει στη θέση του A) συμβολίζουμε τον πίνακα διακύμανσης – συνδιακύμανσης που αντιστοιχεί στα διανύσματα $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ και $x_j = (x_{j1}, x_{j2}, ..., x_{jp})$.

Τέλος, πρέπει να σχολιάσουμε το γεγονός ότι για τη συσταδοποίηση συνεχών δεδομένων συνίσταται η χρήση «πραγματικών» αποστάσεων, δηλαδή αποστάσεων που ικανοποιούν τις εξής ιδιότητες:

$$d(x_i, x_j) = d(x_j, x_i)$$

$$d(x_i, x_j) > 0 \text{ , εάν } x_i \neq x_j$$

$$d(x_i, x_j) = 0 \text{ , εάν } x_i = x_j$$

$$d(x_i, x_j) \leq d(x_i, x_r) + d(x_r, x_j) \text{ (τριγωνική ανισότητα)}$$

Όμως, οι περισσότεροι αλγόριθμοι συσταδοποίησης δέχονται και αποστάσεις που μπορεί να μην ικανοποιούν μια από τις ιδιότητες αυτές, όπως για παράδειγμα την τριγωνική ανισότητα (βλ. Κατέρη, 2006).

Αποστάσεις για δίτιμες μεταβλητές

Έστω ότι x_{ik} είναι η τιμή της k-στης δίτιμης μεταβλητής (με k=1,...,p) και ισούται με 1 ή 0, ανάλογα με το εάν το i-στό υποκείμενο (με i=1,...,n) εμφανίζει ή όχι το υπό μελέτη χαρακτηριστικό. Συγκρίνοντας τα άτομα i και j, έχουμε:

$$(x_{ik} - x_{jk})^2 = \begin{cases} 0, & \text{αν } x_{ik}=x_{jk}=0 \text{ ή } x_{ik}=x_{jk}=1 \\ 1, & \text{αν } x_{ik} \neq x_{jk} \end{cases}$$

και αν χρησιμοποιηθεί η ευκλείδεια απόσταση, τότε θα λαμβάνει υπόψη το πλήθος των ασυμφωνιών μεταξύ δύο υποκειμένων, όσον αφορά στην εμφάνιση των υπό μελέτη χαρακτηριστικών.

Σε αυτό το μέτρο ομοιότητας, η συμφωνία αντιμετωπίζεται όμοια, είτε πρόκειται για συμφωνία στο «όχι», είτε στο «ναι». Όμως, το γεγονός ότι δύο άτομα εμφανίζουν ένα χαρακτηριστικό αποτελεί συχνά μεγαλύτερη ένδειξη ομοιότητάς τους απ'το να μην το εμφανίζει κανένας από τους δύο.

Ας θεωρήσουμε ότι έχουμε ένα ζεύγος παρατηρήσεων $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ και $x_j = (x_{j1}, x_{j2}, ..., x_{jp})$. Τότε, το μέτρο ομοιότητας, έστω $s_{ij} = s(x_i, x_j)$, είναι ένας πραγματικός αριθμός έτσι ώστε να ισχύουν οι παρακάτω ιδιότητες (βλ. Κούτρας, 2007):

i) $s_{ij} \geq 0$ για κάθε i, j και $i = j \Rightarrow s_{ij} = 1$

ii) $s_{ij} \leq 1$

iii) $s_{ij} = s_{ji}$ (συμμετρική ιδιότητα)

Στον Πίνακα 5.2 δίνουμε τους συντελεστές ομοιότητας (*similarity coefficients*) που επιτρέπουν το διαφορετικό χειρισμό της συμφωνίας στο 0 και το 1, αναφέροντας και μια σύντομη επεξήγηση ανά μέτρο. Πρώτα, όμως, παραθέτουμε ένα πίνακα του πλήθους «ομοιοτήτων – ανομοιοτήτων» για δύο υποκείμενα i και j, με p το συνολικό πλήθος μεταβλητών.

ΠΙΝΑΚΑΣ 5.2

Πίνακας «ομοιοτήτων – ανομοιοτήτων»

		Υποκείμενο j	
Υποκείμενο i	1	0	**Σύνολο**
1	a	b	a+b
0	c	d	c+d
Σύνολο	a+c	b+d	a+b+c+d=p

Τα βασικότερα μέτρα ομοιότητας (βλ. Johnson and Wichern, 1998), τα οποία χρησιμοποιούνται για τη συσταδοποίηση ατόμων στα οποία παρατηρούνται δίτιμες μεταβλητές, είναι αυτά που παραθέτουμε στον Πίνακα 5.3.

ΠΙΝΑΚΑΣ 5.3

Αποστάσεις για δίτιμα δεδομένα

[Πηγή: Κούτρας, 2007]

	Ονομασία	Τύπος	Επεξήγηση του μέτρου
1.	*Simple Matching*	$s_{ij} = \frac{a+d}{a+b+c+d}$	Ίσα βάρη για συμφωνίες «1-1» και «0-0»
2.	*Rogers and Tanimoto*	$s_{ij} = \frac{a+d}{a+d+2(b+c)}$	Διπλάσιο βάρος για τις ασυμφωνίες
3.	*Sokal and Sneath*	$s_{ij} = \frac{2(a+d)}{2(a+d)+(b+c)}$	Διπλάσιο βάρος για τις συμφωνίες «1-1» και «0-0»
4.	*Jaccard coefficient*	$s_{ij} = \frac{a}{a+b+c}$	Απουσία συμφωνιών «0-0» από αριθμητή και παρονομαστή
5.	*Dice and Sorensen*	$s_{ij} = \frac{2a}{2a+b+c}$	Απουσία συμφωνιών «0-0» από αριθμητή και παρονομαστή. Διπλάσιο βάρος για τις συμφωνίες «1-1»
6.	*Russel and Rao*	$s_{ij} = \frac{a}{a+b+c+d}$	Όχι συμφωνίες «0-0» στον αριθμητή
7.	*Sokal and Sneath II*	$s_{ij} = \frac{a}{a+2(b+c)}$	Απουσία συμφωνιών «0-0» από αριθμητή και παρονομαστή. Διπλάσιο βάρος στις ασυμφωνίες.
8.	*Sokal and Sneath III*	$s_{ij} = \frac{a+d}{b+c}$	Λόγος: συμφωνίες προς ασυμφωνίες
9.	*Kulczynski*	$s_{ij} = \frac{a}{b+c}$	Λόγος «συμφωνίες προς ασυμφωνίες» έχοντας εξαιρέσει τις συμφωνίες «0-0»

Αποστάσεις για ονοματικές μεταβλητές

Ο τύπος Simple Matching του Πίνακα 5.3 (τύπος 1) γενικεύεται και για ονοματικές κλίμακες. Έστω x και y δύο υποκείμενα του δείγματός μας. Τότε,

$$s(x, y) = \frac{u}{p}$$

όπου u το πλήθος των μεταβλητών με την ίδια τιμή για τα υποκείμενα x και y, ενώ p είναι το σύνολο των μεταβλητών (βλ. Κατέρη, 2006).

Αποστάσεις για διατάξιμες μεταβλητές

Στην περίπτωση που θέλουμε να μετρήσουμε την απόσταση μεταξύ διατάξιμων μεταβλητών, τότε η κλίμακα μέτρησης των κατηγοριών των μεταβλητών αντιμετωπίζεται ως συνεχής. Έτσι, χρησιμοποιούμε τις αποστάσεις που δηλώσαμε στην παράγραφο των συνεχών δεδομένων (Πίνακας 5.1).

Όμως, πρέπει να τονίσουμε ότι απαιτείται ιδιαίτερη προσοχή στην κλίμακα των μεταβλητών που χρησιμοποιούνται. Για την ακρίβεια, εάν όλες οι μεταβλητές δεν μετρούνται στην ίδια κλίμακα, τότε τις μετασχηματίζουμε στο διάστημα (0,1).

Γενικές παρατηρήσεις:

1. Εάν έχουμε ορίσει μια απόσταση d_{ij}, τότε μπορούμε πάντα να δημιουργήσουμε ένα αντίστοιχο μέτρο ομοιότητας, με βάση τον τύπο:

$$s_{ij} = \frac{1}{1 + d_{ij}} \text{ με } 0 < s_{ij} \leq 1$$

2. Εάν έχουμε ορίσει ένα μέτρο ομοιότητας s_{ij}, τότε μπορούμε να δημιουργήσουμε μια αντίστοιχη απόσταση, μέσω του τύπου:

$$d_{ij} = \sqrt{2(1 - s_{ij})}$$

Όμως, αυτός ο τύπος δεν εξασφαλίζει την ισχύ της τριγωνικής ανισότητας. Σύμφωνα με τον Gower, εάν ο πίνακας ομοιοτήτων είναι θετικά ημιορισμένος και η μέγιστη ομοιότητα είναι μετασχηματισμένη έτσι ώστε να ισχύει $s_{ij} = 1$, τότε μπορούμε να εφαρμόσουμε την παραπάνω σχέση.

3. Όταν έχουμε μεταβλητές διαφορετικής τάξης μεγέθους, τις τυποποιούμε ή τις μετασχηματίζουμε ώστε να παίρνουν ίδιες τιμές στο ίδιο διάστημα.

4. Όταν οι προς μελέτη μεταβλητές είναι κατηγορικές μη διατάξιμες, όπως «χρώμα ματιών» ή «τύπος αυτοκινήτου», τότε μια προτεινόμενη τεχνική είναι να κατασκευάσουμε αρχικά δίτιμες μεταβλητές, μια για κάθε επίπεδο της κατηγορικής μεταβλητής, και ύστερα να υπολογίσουμε την αντίστοιχη απόσταση για τα δίτιμα δεδομένα, με βάση τον τύπο (*simple matching distance*):

$$d_{ij} = \frac{p - u}{p} = 1 - \frac{u}{p}$$

όπου u το πλήθος των συμφωνιών (αριθμός των μεταβλητών για τις οποίες τα αντικείμενα i και j εμφανίζουν την ίδια κατάσταση) και p ο αριθμός των μεταβλητών.

5. Έστω ότι θέλουμε να αναλύσουμε ένα σύνολο μεταβλητών που δεν ανήκουν όλες στον ίδιο τύπο (συνεχείς – κατηγορικές – δίτιμες). Τότε, σύμφωνα με τον Gower, η ομοιότητα δύο υποκειμένων x και y υπολογίζεται μέσω του συντελεστή:

$$s(x, y) = \frac{\sum_{i=1}^{p} w_i(x, y) \cdot s_i(x, y)}{\sum_{i=1}^{p} w_i(x, y)}$$

όπου για συνεχείς μεταβλητές:

$$s_i(x, y) = 1 - \frac{|x_i - y_i|}{R_i}$$

με R_i το εύρος της i-στης μεταβλητής, ενώ για διακριτές μεταβλητές:

$$s_i(x, y) = \begin{cases} 1, \text{ για συμφωνίες} \\ 0, \quad \text{διαφορετικά} \end{cases}$$

Τα βάρη w_i παίρνουν την τιμή 1 ή 0, ανάλογα με το αν η σύγκριση στη i-στή μεταβλητή έχει νόημα ή όχι.

5.1.2 «Απόσταση» μεταξύ κατηγορικών δεδομένων

Ανατρέχοντας στη βιβλιογραφία για τρόπους μέτρησης της «απόστασης» μεταξύ κατηγορικών δεδομένων, βρήκαμε διαφορετικές προσεγγίσεις γύρω από το θέμα. Βέβαια, σε αυτή την υποενότητα, δε μας απασχολεί η μέτρηση της απόστασης μέσω ενός τύπου, όπως αυτοί που παρουσιάσαμε στην παράγραφο 5.1.1.

Μιλώντας για «απόσταση», εδώ, εννοούμε την ποσοτικοποίηση της κλίμακας και την έκφραση της διαφορετικότητας μεταξύ των τιμών μιας κατηγορικής μεταβλητής. Στο σημείο αυτό, δίνουμε τρεις εναλλακτικές προτάσεις:

i) Ένα συχνό φαινόμενο για τη μέτρηση των αποστάσεων μεταξύ κατηγορικών δεδομένων είναι η αυθαίρετη θεώρηση ότι δύο ίδιες τιμές (κατηγορίες της μεταβλητής) απέχουν μεταξύ τους 0, ενώ δύο διαφορετικές απέχουν 1 (βλ. Bramer, 2007).

Για παράδειγμα, εάν καταγράφουμε την προτίμηση ατόμων σε χρώματα, η διαφορά του κόκκινου από το κόκκινο είναι μηδενική, ενώ η διαφορά «κόκκινο – μπλε» ισούται με 1.

ii) Επίσης, εάν υπάρχει διαταξιμότητα (ή μερική διαταξιμότητα) μεταξύ των μεταβλητών, μπορούμε να ορίσουμε αποστάσεις κυρίως με βάση την κρίση μας. Ας υποθέσουμε ότι έχουμε μια μεταβλητή με τρεις κατηγορίες: «χαμηλό», «μέτριο» και «υψηλό».

Τότε, η απόσταση μεταξύ «χαμηλού» και «μέτριου» ή «μέτριου» και «υψηλού» θα μπορούσε να οριστεί ότι είναι ίση με 0,5. ενώ, η απόσταση μεταξύ «χαμηλού» και «υψηλού» θα μπορούσε να είναι 1 (βλ. Bramer, 2007).

iii) Τέλος, μια εναλλακτική δυνατότητα που έχουμε είναι να χρησιμοποιούμε ως διαφορά (απόσταση) μεταξύ δύο αντικειμένων το άθροισμα των γνωρισμάτων που δεν είναι κοινά μεταξύ τους.

Δηλαδή, εάν καταγράφουμε διάφορα χαρακτηριστικά για ένα πλήθος ατόμων, η απόσταση μεταξύ δύο αντικειμένων (ατόμων) θα είναι 2 εάν διαφέρουν σε δύο από τα καταγραφόμενα χαρακτηριστικά κ.ο.κ. (βλ. Βαζιργιάννης και Χαλκίδη, 2005).

5.2 Διαδικασία συσταδοποίησης

Η διαδικασία της συσταδοποίησης μπορεί να οδηγήσει σε διαφορετικές τμηματοποιήσεις ενός συνόλου δεδομένων, ανάλογα με το κριτήριο που χρησιμοποιείται για τη συσταδοποίηση. Όπως καταλαβαίνουμε, είναι αναγκαία η προ-επεξεργασία των δεδομένων πριν εφαρμοστεί η συσταδοποίηση σε ένα σύνολο δεδομένων.

Στο προηγούμενο κεφάλαιο, της προ-επεξεργασίας των δεδομένων, αναφερθήκαμε στο ρόλο της συσταδοποίησης στο συγκεκριμένο της αλυσίδας KDD. Ας δούμε όμως, σε αυτή την ενότητα, ποια είναι τα βήματα της διαδικασίας συσταδοποίησης, καθώς και ποιες είναι οι απαιτήσεις που έχουμε από έναν αλγόριθμο συσταδοποίησης. Επίσης, είναι σημαντικό να εντοπίσουμε τη συνεισφορά της συσταδοποίησης στην αναγνώριση προτύπων (*pattern recognition*).

5.2.1 Βασικά βήματα

Για την πραγματοποίηση μιας επιτυχημένης συσταδοποίησης, πρέπει να εκτελεστούν τα ακόλουθα βήματα (βλ. Fayyad et al., 1996-a):

1ο ΒΗΜΑ: Επιλογή χαρακτηριστικών γνωρισμάτων

Στόχος του Βήματος αυτού είναι να επιλεγούν κατάλληλα τα γνωρίσματα (*attributes*) στα οποία πρόκειται να εφαρμοστεί η συσταδοποίηση. Έτσι, θα μπορέσει να κωδικοποιηθεί όσο το δυνατόν περισσότερη πληροφορία σχετικά με το ζήτημα που μας απασχολεί.

Κατά συνέπεια, η προ-επεξεργασία των δεδομένων μπορεί να είναι απαραίτητη πριν τη χρησιμοποίησή τους στη διαδικασία της συσταδοποίησης.

2ο ΒΗΜΑ: Αλγόριθμος συσταδοποίησης

Μέσα από το δεύτερο Βήμα, γίνεται η επιλογή ενός αλγορίθμου που οδηγεί στον καθορισμό ενός καλού σχήματος συσταδοποίησης (*clustering scheme*) για ένα σύνολο δεδομένων. Η δυνατότητα του αλγορίθμου να κριθεί κατάλληλος και να καθορίσει ένα σχήμα συσταδοποίησης που να ταιριάζει στο σύνολο δεδομένων καθορίζεται από:

i) **Μέτρο εγγύτητας** (*proximity measure*): το μέτρο αυτό προσδιορίζει πόσο «όμοια» είναι δύο αντικείμενα (δηλαδή διανύσματα γνωρισμάτων). Στις περισσότερες περιπτώσεις πρέπει να εξασφαλίσουμε ότι όλα τα γνωρίσματα που επιλέχθηκαν στο πρώτο Βήμα συμβάλλουν εξίσου στον υπολογισμό του μέτρου εγγύτητας και δεν υπάρχει κανένα γνώρισμα που να υπερισχύει των άλλων.
ii) **Κριτήριο συσταδοποίησης:** το κριτήριο που θα καθοριστεί μπορεί να εκφραστεί μέσω μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων. Επίσης, πρέπει να ληφθεί υπόψη ο τύπος συστάδων που αναμένονται να εμφανιστούν στο σύνολο δεδομένων. Η επιλογή ενός κατάλληλου κριτηρίου οδηγεί σε μια τμηματοποίηση που να ταιριάζει στο συγκεκριμένο σύνολο δεδομένων.

3ο ΒΗΜΑ: Επικύρωση αποτελεσμάτων

Χρησιμοποιώντας κατάλληλα κριτήρια και τεχνικές, μπορούμε να προσδιορίσουμε την ακρίβεια των αποτελεσμάτων του αλγορίθμου συσταδοποίησης. Στις περισσότερες εφαρμογές, η τελική τμηματοποίηση των δεδομένων απαιτεί κάποιου είδους αξιολόγηση,

καθώς οι αλγόριθμοι συσταδοποίησης καθορίζουν τις συστάδες που δεν είναι γνωστές εκ των προτέρων (ανεξάρτητα από τις μεθόδους συσταδοποίησης).

4ο ΒΗΜΑ: Ερμηνεία αποτελεσμάτων

Στο τελευταίο Βήμα, ερμηνεύονται τα αποτελέσματα. Για το σκοπό αυτό, είναι σύνηθες να ενώνονται τα αποτελέσματα της συσταδοποίησης με άλλα πειραματικά στοιχεία και αποτελέσματα προηγούμενης ανάλυσης των υπό μελέτη στοιχείων. Στόχος είναι να προκύψει το σωστό συμπέρασμα.

5.2.2 Απαιτήσεις

Η εφαρμογή της συσταδοποίησης αποτελεί ένα ενδιαφέρον πεδίο έρευνας, όπου κάθε αλγόριθμος μπορεί να έχει συγκεκριμένες απαιτήσεις. Οι σημαντικότερες από αυτές καταγράφονται παρακάτω (βλ. Han and Kamber, 2001).

Υπάρχουν αρκετοί αλγόριθμοι συσταδοποίησης οι οποίοι είναι αποτελεσματικοί μόνο εάν πρόκειται να εφαρμοστούν σε μικρά σύνολα δεδομένων (έως 200 αντικείμενα). Το ζήτημα, όμως, είναι να υπάρχει δυνατότητα **κλιμάκωσης** (*scalability*) ενός αλγορίθμου, καθώς μία αποκαλούμενη «μεγάλη» βάση δεδομένων μπορεί να περιέχει εκατομμύρια αντικείμενα! Έτσι, η εφαρμογή συσταδοποίησης σε ένα δείγμα που έχει ληφθεί από μια τεράστια βάση δεδομένων μπορεί να οδηγήσει σε μεροληπτικά (*biased*) αποτελέσματα.

Επομένως, ένας αλγόριθμος συσταδοποίησης πρέπει να χαρακτηρίζεται από δυνατότητες εξελιξιμότητας, ώστε να μπορεί να αντιμετωπίσει και μεγάλα σύνολα δεδομένων. Επίσης, από έναν αλγόριθμο απαιτείται η δυνατότητα χειρισμού **διαφορετικού τύπου μεταβλητών**. Είναι πολύ πιθανό να συναντήσουμε μία βάση δεδομένων, η οποία να περιέχει πλήθος διαφορετικού τύπου μεταβλητών, όπως αριθμητικές, δίτιμες, διατάξιμες κ.λπ.

Μια άλλη απαίτηση που θα είχαμε από έναν αλγόριθμο είναι να οδηγεί σε συστάδες **αυθαίρετου σχήματος**. Πολλοί αλγόριθμοι βασίζονται στην ευκλείδεια ή τη city-block απόσταση (τύποι 1 και 5 αντίστοιχα, Πίνακας 5.1). Οι αλγόριθμοι που βασίζονται σε αυτούς τους τύπους απόστασης, τείνουν να δημιουργούν συστάδες σφαιρικού σχήματος και με όμοιο μέγεθος και πυκνότητα. Όμως, μια δημιουργηθείσα συστάδα μπορεί να έχει οποιοδήποτε σχήμα. Επομένως, είναι σημαντικό για έναν αλγόριθμο να μπορεί να οδηγεί σε συστάδες ανεξάρτητου σχήματος και πυκνότητας.

Για τους περισσότερους αλγόριθμους συσταδοποίησης, είναι πιθανή η απαίτηση εισαγωγής κάποιων **παραμέτρων**, όπως ο επιθυμητός αριθμός των συστάδων που θέλουμε να δημιουργηθούν. Όμως, είναι καλό για έναν αλγόριθμο να έχει τις λιγότερες δυνατές προαπαιτούμενες παραμέτρους εισόδου, καθώς τα αποτελέσματα συχνά επηρρεάζονται από αυτές. Πολλές φορές, είναι δύσκολο να προσδιοριστούν ορισμένες παράμετροι εισόδου, ειδικά όταν πρόκειται για πολυδιάστατα αντικείμενα, πράγμα που μειώνει την ευελιξία του χρήστη αλλά και την ποιότητα της συσταδοποίησης.

Όσον αφορά την εισαγωγή των στοιχείων σε ένα σύνολο δεδομένων, είναι πολύ πιθανό να δημιουργηθούν τελείως διαφορετικές συστάδες, εάν εισαχθούν τα δεδομένα με άλλο τρόπο. Δηλαδή, η **διάταξη** των δεδομένων παίζει σημαντικό ρόλο. Αυτό σημαίνει ότι απαιτείται για έναν αλγόριθμο να μην είναι ευαίσθητος στη σειρά με την οποία εισάγονται τα δεδομένα.

Επιπλέον, αυτό που επιθυμούμε από έναν αλγόριθμο είναι η ικανότητά του να χειρίζεται δεδομένα με **θόρυβο** (παράγραφος 4.2.2) ή πολύ μεγάλο **πλήθος διαστάσεων ή μεταβλητών.** Όπως παρατηρούμε, έχουμε αναφέρει πολλές φορές ως εδώ το ζήτημα που προκύπτει όταν υπάρχει μεγάλο πλήθος μεταβλητών ή διαστάσεων, είτε σαν μεμονωμένο θέμα που χρειάζεται να αντιμετωπίζει ένας αλγόριθμος συσταδοποίησης, είτε σαν πρόβλημα που περιπλέκει άλλες απαιτήσεις που σχολιάστηκαν προηγουμένως.

Κλείνοντας, συμπληρώνουμε ότι ο ιδανικός αλγόριθμος συσταδοποίησης πρέπει να είναι **κατανοητός** και **χρήσιμος**, αλλά και να καλύπτει τους **περιορισμούς** που μπορεί να τίθενται στα πλαίσια εφαρμογής πραγματικών δεδομένων.

5.2.3 Συνεισφορά στην αναγνώριση προτύπων

Μιλώντας για την αναγνώριση προτύπων, εννοούμε τον επιστημονικό κλάδο που ασχολείται με μεθόδους περιγραφής ή ταξινόμησης αντικειμένων (Marques de Sá, 2001). Τα τελευταία χρόνια, έχει αναπτυχθεί ιδιαίτερο ενδιαφέρον στο σχεδιασμό και εγκατάσταση αλγορίθμων που περιγράφουν ή ταξινομούν αντικείμενα.

Στην περίπτωση που θέλουμε να πραγματοποιήσουμε μεθόδους ταξινόμησης ή πρόβλεψης, εκφράζουμε την ομοιότητα σε μορφή απόστασης, η οποία θεωρείται αριθμητική ποσότητα. Όμως, υπάρχουν περιπτώσεις που χρειάζεται, τουλάχιστον ως πρώτο βήμα, να εκτελέσουμε **εργασίες περιγραφής** (*description tasks*).

Οι τρεις βασικές προσεγγίσεις της αναγνώρισης προτύπων είναι η στατιστική, η προσέγγιση μέσω νευρωνικών δικτύων και η δομική προσέγγιση (*structural approach*). Στο

Σχήμα 5.2, βλέπουμε ένα διάγραμμα απεικόνισης των διαφορετικών προσεγγίσεων της αναγνώρισης προτύπων. Για την καλύτερη κατανόηση του διαγράμματος, παραθέτουμε πρώτα ένα πίνακα (βλ. Πίνακα 5.4) με τις επεξηγήσεις των συμβολισμών του διαγράμματος, καθώς και μια σύντομη περιγραφή ανά συμβολισμό, ώστε να υπενθυμίσουμε όσα έχουν αναφερθεί σε προηγούμενα κεφάλαια.

ΠΙΝΑΚΑΣ 5.4

Επεξήγηση των διαφορετικών προσεγγίσεων της αναγνώρισης προτύπων

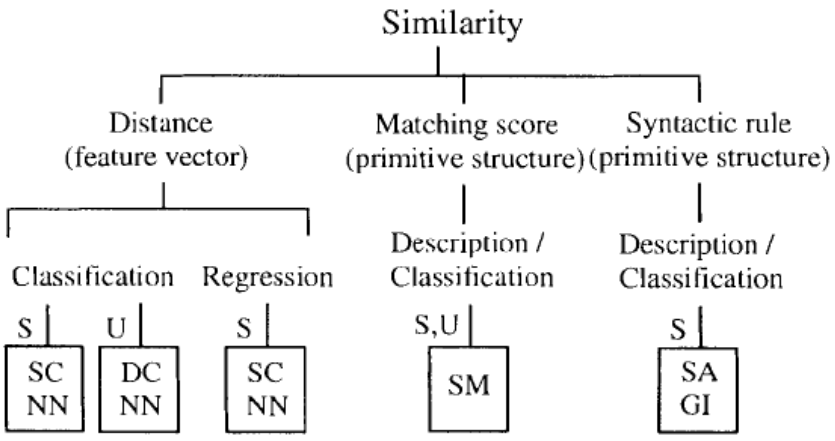
Συμβολισμός	Ερμηνεία	Επεξήγηση
S	supervised	Επιβλεπόμενη διαδικασία
U	unsupervised	Μη επιβλεπόμενη διαδικασία
SC	statistical classification	Στατιστική ταξινόμηση: εργασία πρόβλεψης
NN	neural networks	Νευρωνικά δίκτυα (σχολιάζονται στην παράγραφο 5.4.1)
DC	data clustering	Συσταδοποίηση δεδομένων: εργασία περιγραφής (αντικείμενο κεφαλαίου 5)
SM	structural matching	Συνδυασμός με βάση τη δομή
SA	syntactic analysis	Συντακτική ανάλυση
GI	grammatical inference	Συμπερασματολογία βασισμένη στη γραμματική

Ακολουθεί το Σχήμα 5.2, όπου απεικονίζονται οι εναλλακτικές προσεγγίσεις της αναγνώρισης προτύπων.

ΣΧΗΜΑ 5.2

Προσεγγίσεις της αναγνώρισης προτύπων

[Πηγή: Marques de Sá, 2001]



Όπως βλέπουμε, η συσταδοποίηση κατατάσσεται στην **μη επιβλεπόμενη ταξινόμηση** (βλ. Σχήμα 5.2). Αυτή είναι και η συνεισφορά της στην αναγνώριση προτύπων. Χρησιμοποιώντας ένα μέτρο ομοιότητας, η συσταδοποίηση είναι σε θέση να οργανώσει τα δεδομένα / πρότυπα σε ενδιαφέρουσες ομάδες, χωρίς να έχει σχετική εκ των προτέρων πληροφορία. Επίσης, η συνεισφορά της συσταδοποίησης είναι σημαντική στην διερευνητική εύρεση ομοιοτήτων μεταξύ των δεδομένων.

5.3 Εφαρμογές συσταδοποίησης

Οι μέθοδοι συσταδοποίησης αποτελούν ένα πολύ σημαντικό μέσο για την εξαγωγή χρήσιμων συμπερασμάτων σε πολλές εφαρμογές, τόσο στον τομέα των επιστημών, όσο και των επιχειρήσεων. Το πλαίσιο εφαρμογών της συσταδοποίησης (Theodoridis and Koutroumbas, 2003) καλύπτει τις παρακάτω διαδικασίες :

- **Μείωση δεδομένων**

  Η προσφορά της συσταδοποίησης σε αυτό το πεδίο είναι η συμβολή στη συμπίεση της πληροφορίας των δεδομένων. Ο χωρισμός ενός μεγάλου όγκου πληροφορίας σε «ενδιαφέρουσες» συστάδες μειώνει τη δυσκολία στο χειρισμό του συνόλου δεδομένων, καθώς ο «αντιπρόσωπος» της συστάδας εκπροσωπεί αυτή την ομάδα δεδομένων στην ανάλυση.

- **Συμβολή στη διατύπωση υποθέσεων**

  Πολλές φορές είναι χρήσιμο να προκύψουν μερικές υποθέσεις για τα δεδομένα. Για παράδειγμα, σε μια βάση δεδομένων στεγαστικών δανείων, μπορεί να προκύψουν δύο σημαντικές συστάδες πελατών με βάση την οικογενειακή τους κατάσταση (π.χ. οικογενειάρχης, εργένης) και το χρόνο εκπλήρωσης των υποχρεώσεών τους.

  Τότε, μπορεί να προκύψουν ορισμένες ενδιαφέρουσες υποθέσεις, όπως: «Οι οικογενειάρχες είναι πιο συνεπείς στην πληρωμή της δόσης».

- **Έλεγχος υπόθεσης**

  Η εγκυρότητα μιας συγκεκριμένης υπόθεσης επαληθεύεται μέσα από την ανάλυση συστάδων. Για παράδειγμα, για να ελέγξουμε την υπόθεση που έγινε προηγουμένως για τους οικογενειάρχες και τη συνέπεια στην πληρωμή της δόσης, μπορούμε να

εφαρμόσουμε τη διαδικασία συσταδοποίησης σε ένα αντιπροσωπευτικό σύνολο καναλιών πώλησης (καταστήματα τράπεζας ή σημεία προώθησης δανείων).

Συγκεντρώνοντας από κάθε κανάλι σχετικές πληροφορίες, θα θεωρήσουμε ότι η υπόθεση που έχει γίνει θα είναι έγκυρη εάν δημιουργηθεί μια συστάδα που να αντιστοιχεί στο «οι οικογενειάρχες είναι πιο συνεπείς στην πληρωμή της δόσης».

Στη συνέχεια, μπορεί να εφαρμοστεί και στατιστικός έλεγχος της υπόθεσης. Η μηδενική υπόθεση μπορεί να έχει τη μορφή:

$$H_0: P_{\text{οικογενειάρχης}} \geq P_{\text{εργένης}}$$

ως προς την εναλλακτική:

$$H_1: P_{\text{οικογενειάρχης}} < P_{\text{εργένης}}$$

όπου P το ποσοστό των συνεπών δανειοληπτών.

- **Πρόβλεψη βασισμένη σε συστάδες**

  Οι συστάδες που προκύπτουν ύστερα από την ανάλυση ενός συνόλου δεδομένων, έχουν ως «ετικέτες» τα χαρακτηριστικά των προτύπων που ανήκουν σε αυτές. Έτσι, κάθε συστάδα έχει κάποια συγκεκριμένα χαρακτηριστικά. Τα άγνωστα πρότυπα μπορούν να ταξινομηθούν στις προσδιοριζόμενες συστάδες, με βάση την ομοιότητά τους στα χαρακτηριστικά των συστάδων.

  Με τον τρόπο αυτό μπορεί να εξαχθεί χρήσιμη γνώση, η οποία θα χρησιμοποιηθεί μελλοντικά. Για παράδειγμα, ας σκεφτούμε τους πελάτες ενός τμήματος στεγαστικών δανείων. Έστω ότι δημιουργούνται για αυτούς όμοιες ως προς τη συμπεριφορά ομάδες με βάση το εάν παρέμειναν πελάτες της τράπεζας ή όχι.

  Τότε, εάν έρθει ένας νέος πελάτης, μπορούμε να προσδιορίσουμε τη συστάδα στην οποία εκτιμούμε ότι θα ταξινομηθεί και να προβλέψουμε τη συμπεριφορά του. Έτσι, θα αποφασίσουμε ανάλογα για το αίτημά του.

Μέσα από το τελευταίο παράδειγμα κατανοούμε όσα σχολιάστηκαν στο δεύτερο κεφάλαιο, αλλά και στην αρχή αυτού του κεφαλαίου. Συνήθως, η συσταδοποίηση προηγείται της ταξινόμησης, καθώς πρώτα περιγράφεται η βάση δεδομένων, προκύπτουν τα πρότυπα και στη συνέχεια αυτό το σύνολο χρησιμοποιείται ως εκπαιδευτικό δείγμα για να πραγματοποιηθεί η ταξινόμηση.

Δηλαδή, η συσταδοποίηση μπορεί να χρησιμεύσει ως βήμα προεπεξεργασίας για άλλους αλγορίθμους. Οι χαρακτηριστικότερες εφαρμογές της συσταδοποίησης, σύμφωνα με τους Han και Kamber (2001), είναι στους παρακάτω κλάδους:

- ✓ **Επιχειρήσεις**

  Έχουμε ήδη συνειδητοποιήσει τη χρησιμότητα των διαδικασιών συσταδοποίησης στον επιχειρηματικό κλάδο. Η συσταδοποίηση μπορεί, για παράδειγμα, να βοηθήσει τους εμπόρους να ανακαλύψουν σημαντικές συστάδες στη βάση δεδομένων των πελατών τους και να τις χαρακτηρίσουν με βάση τα αγοραστικά πρότυπα.

- ✓ **Βιολογία**

  Η χρήση της συσταδοποίησης στη βιολογία μπορεί να καθορίσει τις ταξονομίες (*taxonomies*), να κατηγοριοποιήσει τα γονίδια με παρόμοια λειτουργία και να μελετηθούν σε βάθος οι υπό μελέτη δομές των πληθυσμών.

- ✓ **Χωρική ανάλυση στοιχείων**

  Τα χωρικά δεδομένα είναι αυτά που μπορούν να ληφθούν από δορυφορικές εικόνες, ιατρικό εξοπλισμό, γεωγραφικά συστήματα πληροφοριών (*GIS*), εξερεύνηση βάσεων δεδομένων εικόνας κ.λπ. Επομένως, κατανοούμε ότι ο όγκος τέτοιου τύπου δεδομένων είναι τεράστιος και η εξέτασή τους με λεπτομέρεια είναι ακριβή και δύσκολη.

  Αυτό που μπορεί να προσφέρει η συσταδοποίηση στην περίπτωση αυτή είναι η συμβολή στην αυτοματοποίηση της διαδικασίας ανάλυσης και κατανόησης των χωρικών δεδομένων. Μέσω της συσταδοποίησης εξάγονται ενδιαφέροντα χαρακτηριστικά και πρότυπα που μπορούν να υπάρξουν σε μεγάλες χωρικές βάσεις δεδομένων.

- ✓ **Εξόρυξη στον παγκόσμιο ιστό**

  Σε αυτήν την περίπτωση, η συσταδοποίηση εξυπηρετεί την ανακάλυψη σημαντικών συστάδων εγγράφων στην τεράστια συλλογή ημι-δομημένων εγγράφων του παγκόσμιου ιστού (*WWW*). Έτσι, ύστερα από την κατηγοριοποίηση των εγγράφων του παγκόσμιου ιστού εξυπηρετούμαστε στην ανακάλυψη χρήσιμης πληροφορίας.

5.4 Διάκριση μεθόδων και σχετικών αλγορίθμων

Στα πλαίσια της βιβλιογραφικής μας ανασκόπησης, διαπιστώσαμε ότι υπάρχουν αρκετοί αλγόριθμοι συσταδοποίησης. Συνεπώς, η κατηγοριοποίησή τους σε συγκεκριμένες ομάδες μας εξυπηρετεί ώστε να έχουμε σαφέστερη εικόνα. Η επιλογή του κατάλληλου αλγορίθμου ανά περίπτωση βασίζεται στον τύπο δεδομένων που είναι διαθέσιμα, αλλά και στο στόχο μας και την εφαρμογή που θέλουμε να πραγματοποιήσουμε.

Οι υποενότητες που ακολουθούν διαχωρίζουν τους αλγορίθμους συσταδοποίησης σε πέντε ομάδες. Ο διαχωρισμός αυτός έγινε με βάση τη μέθοδο (ή εφαρμογή) συσταδοποίησης που θέλουμε να εκτελέσουμε (Han and Kamber, 2001).

5.4.1 Διαιρετική συσταδοποίηση (*Partitional clustering*)

Δεδομένου ότι έχουμε μια βάση δεδομένων n αντικειμένων, μπορούμε να δημιουργήσουμε k διαμερίσεις (*partitions*) των δεδομένων, όπου κάθε διαμέριση παριστάνει μια συστάδα, ενώ πρέπει $k \leq n$. Τα βασικά χαρακτηριστικά των ομάδων που δημιουργούνται είναι ότι κάθε ομάδα πρέπει να περιέχει τουλάχιστον ένα αντικείμενο, αλλά και ότι κάθε αντικείμενο πρέπει να ανήκει αυστηρά σε μια ομάδα. Οι συστάδες που προκύπτουν δεν σχετίζονται μεταξύ τους.

Σε αυτή την κατηγορία ανήκουν αλγόριθμοι που προσπαθούν να ελαχιστοποιήσουν μια συνάρτηση, η οποία μπορεί να δίνει έμφαση στην τοπική δομή των δεδομένων, αναθέτοντας συστάδες στα άκρα της συνάρτησης (ελάχιστο ή μέγιστο) ή στη γενική δομή των δεδομένων. Βασικότερος στόχος των αλγορίθμων διαιρετικής συσταδοποίησης είναι η **ελαχιστοποίηση** των μέτρων ανομοιότητας μεταξύ των δειγμάτων μέσα σε κάθε συστάδα, καθώς και η **μεγιστοποίηση** της ανομοιότητας μεταξύ των διαφορετικών συστάδων.

Στην παράγραφο 5.1.1 αναφέραμε τα σημαντικότερα μέτρα ομοιότητας, ενώ παραθέσαμε και ένα πίνακα του πλήθους «ομοιοτήτων - ανομοιοτήτων» μεταξύ δύο υποκειμένων (βλ. Πίνακα 5.2). Αναλόγως ορίζεται η ανομοιότητα μεταξύ των δειγμάτων κάθε συστάδας, καθώς και μεταξύ των συστάδων. Ως ανομοιότητα, δηλαδή, εννοούμε τη **διαφορά** ματαξύ των υποκειμένων i και j.

Μέσα από τη διαιρετική συσταδοποίηση δημιουργείται μια αρχική διαμέριση, αλλά υπάρχει πιθανότητα να απαιτηθεί **επανάληψη** της διαδικασίας ώστε να υπάρξει βελτίωση στη δομή των συστάδων. Αυτό μπορεί να σημαίνει μετακίνηση ενός αντικειμένου από μια

συστάδα σε μια άλλη, ώστε να περιλαμβάνονται «όμοια» αντικείμενα σε κάθε μια από τις τελικές συστάδες.

Οι κλασσικοί αλγόριθμοι διαμέρισης είναι οι k-means και k-medoid. Στον k-means, κάθε συστάδα αντιπροσωπεύεται από τη μέση τιμή των αντικειμένων της. Αυτό είναι ένα από τα μειονεκτήματα αυτού του αλγορίθμου, καθώς απαιτείται η γνώση ή εύρεση της μέσης τιμής. Επιπλέον, ο αλγόριθμος k-means είναι ευαίσθητος στην παρουσία έκτροπων παρατηρήσεων (*outliers*).

Στην περίπτωση ύπαρξης θορύβου και outliers, καταλληλότερος είναι ο αλγόριθμος k-medoid. Στον αλγόριθμο αυτό, κάθε συστάδα αντιπροσωπεύεται από ένα από τα αντικείμενα που βρίσκεται κοντά στο κέντρο της συστάδας. Για την ακρίβεια, ορίζουμε μια **«κεντρική»** τιμή, η οποία μπορεί να είναι είτε ένα από τα υπάρχοντα αντικείμενα, είτε μια «φανταστική» τιμή στο κέντρο των παρατηρήσεων.

Γενικά, τα αρνητικά χαρακτηριστικά των διαιρετικών αλγορίθμων είναι ότι δημιουργούν συστάδες σφαιρικού σχήματος, πράγμα που δεν είναι πάντα επιθυμητό. Επίσης, οι διαιρετικοί αλγόριθμοι είναι κατά βάση αποτελεσματικοί για σύνολα δεδομένων μικρού μεγέθους.

Ως παραλλαγές του k-means γνωρίζουμε τον αλγόριθμο k-modes, ο οποίος χειρίζεται κατηγορικά δεδομένα, αλλά και τον k-prototypes, ο οποίος είναι κατάλληλος για μικτά δεδομένα. Επίσης, μια παραλλαγή του k-medoid (βλ. Han and Kamber, 2001) είναι ο PAM (*Partition around medoids*).

Για να αντιμετωπίσουμε πολύ μεγάλα σύνολα δεδομένων, αντί των αλγορίθμων τύπου k-medoid χρησιμοποιούμε τους CLARA (*Clustering large applications*) και CLARANS (*Clustering large applications upon randomized search*). Ο αλγόριθμος CLARA αναπτύχθηκε από τους Kaufman και Rousseeuw (1990), ενώ ο CLARANS είναι η βελτίωση του CLARA (βλ. Ng and Han, 1994).

5.4.2 Ιεραρχική συσταδοποίηση (*Hierarchical clustering*)

Μία ιεραρχική μέθοδος συσταδοποίησης προκαλεί μια ιεραρχική αποσύνθεση ενός δοθέντος συνόλου αντικειμένων. Η ιεραρχική συσταδοποίηση χωρίζεται σε δύο μέρη, ανάλογα με τον τρόπο που επιθυμούμε να κατασκευάσουμε τις συστάδες. Το πρώτο μέρος είναι η **συσσωρευτική προσέγγιση** (*agglomerative approach*), η οποία καλείται και «από κάτω προς τα πάνω» προσέγγιση (*"bottom-up" approach*).

Με βάση την ονομασία, καταλαβαίνουμε εύκολα ότι ο πρώτος αυτός τρόπος ιεραρχικής συσταδοποίησης ξεκινά θεωρώντας κάθε αντικείμενο ως μια μεμονωμένη συστάδα και προχωρά «προς τα πάνω» ενώνοντας όσο μπορεί σε ομάδες (υπάρχει περίπτωσει να τα οδηγήσει όλα σε μια συστάδα).

Το δεύτερο μέρος της ιεραρχικής συσταδοποίησης είναι η **διαχωριστική προσέγγιση** (*divisive approach*) ή αλλιώς «από πάνω προς τα κάτω» προσέγγιση (*"top-down" approach*). Σε αυτή την περίπτωση, ξεκινάμε θεωρώντας όλα τα αντικείμενα σε μια συστάδα και επιχειρούμε το διαχωρισμό τους σε ομοιογενείς ομάδες. Το τέλος της διαδικασίας βρίσκει κάθε αντικείμενο σε μια συστάδα ή σταματέει κάπου ενδιάμεσα, ανάλογα με τις συνθήκες που πρέπει να πληρούνται.

Το ελλάτωμα των ιεραρχικών μεθόδων συσταδοποίησης είναι ότι κάθε διαχωρισμός ή συσσώρευση που πραγματοποιείται δε μπορεί να ανακληθεί. Αυτό βέβαια μπορεί να θεωρηθεί και το συστατικό που τονώνει την επιτυχία αυτής της κατηγορίας συσταδοποίησης, αφού τελικά υπάρχει λιγότερο κόστος και προβληματισμός, εφόσον ξέρουμε πως ότι έγινε δεν αλλάζει.

Ένας ιδιαίτερα ενδιαφέρων συνδυασμός ιδιοτήτων είναι η επαναληπτική τοποθέτηση (*iterative relocation*) και η ιεραρχική συσσώρευση (*hierarchical agglomeration*). Θα ήταν χρήσιμο, δηλαδή, να γίνει αρχικά η εφαρμογή ενός συσσωρευτικού αλγορίθμου και στη συνέχεια να πραγματοποιηθεί εκλέπτυνση του αποτελέσματος μέσω μιας επαναληπτικής τοποθέτησης. Για το λόγο αυτό, κατασκευάστηκαν κλιμακούμενοι αλγόριθμοι συσταδοποίησης, όπως οι BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) και CURE (*Clustering Using REpresentatives*).

Οι αλγόριθμοι αυτοί δημιουργήθηκαν στα πλαίσια του ενδιαφέροντος για βελτίωση της ποιότητας συσταδοποίησης. Ο BIRCH είναι διαχωριστικός (βλ. Zhang et al., 1996), ενώ ο CURE συσσωρευτικός (βλ. Guha et al., 1998). Ένας άλλος αλγόριθμος που κατασκευάστηκε με τις ίδιες προθέσεις είναι ο ROCK (*Robust Clustering for Categorical Attributes*), ο οποίος μάλιστα αφορά στα κατηγορικά δεδομένα (βλ. Guha et al., 2000) και θα παρουσιαστεί στο επόμενο κεφάλαιο.

Επιπλέον, ένας σημαντικός αλγόριθμος ιεραρχικής συσταδοποίησης είναι ο CHAMELEON (βλ. Karypis et al., 1999), ο οποίος εξερευνά τη δυναμική μοντελοποίηση

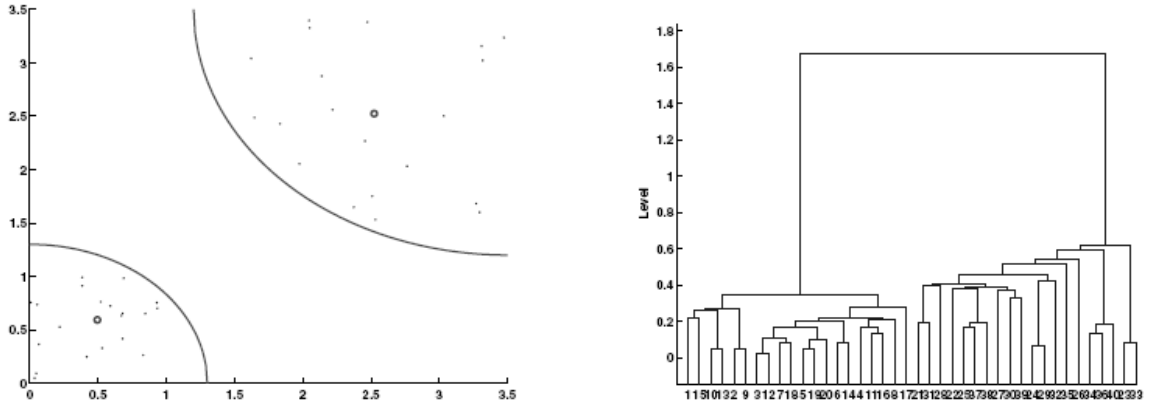
(*dynamic modeling*) στα πλαίσια της ιεραρχικής συσταδοποίησης. Ο αλγόριθμος αυτός είναι κατάλληλος για όλους τους τύπους δεδομένων για τους οποίους μπορεί να κατασκευαστεί πίνακας εγκύτητας (βλ. παράγραφο 5.1.1).

Κλείνοντας την ενότητα αυτή, παραθέτουμε ένα διπλό σχήμα όπου απεικονίζεται η διαφορά μεταξύ διαιρετικής και ιεραρχικής συσταδοποίησης, όσον αφορά στα εξαγόμενα αποτελέσματα. Σε αντίθεση με τη διαιρετική συσταδοποίηση που παρέχει απλά ένα διαχωρισμό των δεδομένων, η ιεραρχικοί αλγόριθμοι έχουν σαν αποτέλεσμα ένα δέντρο από συστάδες, το οποίο καλείται **δενδρογράφημα**.

ΣΧΗΜΑ 5.2

Αποτελέσματα διαιρετικής και ιεραρχικής συσταδοποίησης

[Πηγή: Abonyi and Feil, 2007]



Το δέντρο αυτό (βλ. Σχήμα 5.2) δίνει μια **απεικόνιση** της σχέσης μεταξύ των τελικών συστάδων και δεν πρέπει να συγχέεται με τα δέντρα ταξινόμησης ή παλινδρόμησης (*classification or regression trees*), τα οποία χρησιμοποιούνται για τη διαμόρφωση προβλέψεων.

Με βάση το Σχήμα 5.2, μπορούμε να διαπιστώσουμε ότι η διαιρετική συσταδοποίηση είναι καταλληλότερη εάν έχουμε ένα μεγάλο σύνολο δεδομένων, καθώς η δημιουργία ενός δενδρογραφήματος θα ήταν αρκετά πολύπλοκη.

5.4.3 Μέθοδοι βασισμένες στην πυκνότητα (*Density-based methods*)

Οι αλγόριθμοι αυτής της κατηγορίας καταλήγουν στη δημιουργία αλγορίθμων αυθαίρετου (*arbitrary*) σχήματος, πράγμα στο οποίο υστερούν οι διαιρετικοί αλγόριθμοι. Στόχος τους είναι η ένωση περιοχών όμοιας πυκνότητας σε συστάδες ή η ένωση των αντικειμένων των συστάδων με βάση τη συνάρτηση κατανομής πυκνότητας (*density function distribution*).

Χαρακτηριστικοί αλγόριθμοι αυτής της κατηγορίας είναι οι DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) και DENCLUE (*an abbreviation of DENsity-based CLUstering*). Ο DBSCAN (βλ. Ester et al., 1996) ενώνει περιοχές με αρκετά υψηλή πυκνότητα, ενώ ο DENCLUE (βλ. Hinneburg and Keim, 1998) στηρίζεται σε ένα σύνολο συναρτήσεων κατανομής πυκνότητας.

Ένα ακόμη προσόν των αλγορίθμων αυτής της κατηγορίας είναι ότι μπορούν να απομακρύνουν το θόρυβο από τα δεδομένα.

5.4.4 Μέθοδοι βασισμένες στο πλέγμα (*Grid-based methods*)

Ένας αλγόριθμος που ανήκει σε αυτή την κατηγορία συσταδοποίησης μετατρέπει σε κβάντα (*quantizes*) το χώρο του αντικειμένου σε ένα περιορισμένο αριθμό κελιών, τα οποία δημιουργούν μια δομή πλέγματος. Στη συνέχεια, οι λειτουργίες συσταδοποίησης εφαρμόζονται σε αυτή τη δομή.

Το προσόν αυτής της κατηγορίας είναι ότι οι αλγόριθμοι έχουν γρήγορο χρόνο εκτέλεσης, ανεξαρτήτως του πλήθους των αντικειμένων. Ο χρόνος εκτέλεσης εξαρτάται μόνο από τον αριθμό κελιών που δημιουργείται σε κάθε διάσταση του κβαντωμένου χώρου. Χαρακτηριστικότερος αλγόριθμος είναι ο STING (*Statistical Information Grid* – βλ. Wang et al., 1997).

Επίσης, δύο αλγόριθμοι που βασίζονται ταυτόχρονα στην πυκνότητα αλλά και στο πλέγμα είναι οι CLIQUE (*Clustering In QUest* – βλ. Agrawal et al., 1998) και Wave-Cluster (βλ. Sheikholeslami et al., 1998). Ο CLIQUE είναι κατάλληλος για τεράστιες βάσεις δεδομένων, όπου περιλαμβάνονται πολυδιάστατα δεδομένα.

5.4.5 Μέθοδοι βασισμένες σε μοντέλο (*Model-based methods*)

Μία μέθοδος που βασίζεται σε μοντέλο υποθέτει ένα συγκεκριμένο μοντέλο για κάθε μια από τις συστάδες. Στη συνέχεια, η μέθοδος αυτή επιχειρεί να βρει τη βέλτιστη εφαρμογή μεταξύ μοντέλου και δεδομένων.

Οι μέθοδοι αυτής της κατηγορίας στηρίζονται συνήθως στην υπόθεση ότι τα δεδομένα παράγονται από μια ανάμειξη υποκείμενων κατανομών πιθανότητας. Υπάρχουν δύο διαφορετικές μεθοδολογίες προσέγγισης τέτοιων μεθόδων: η στατιστική προσέγγιση και η προσέγγιση μέσω νευρωνικών δικτύων (*neural network*). Στην ενότητα 5.1.4 αναφέρθηκαν τρεις προσεγγίσεις της αναγνώρισης προτύπων, οι δύο εκ των οποίων αναφέρθηκαν και εδώ.

Σχετικά με τη στατιστική προσέγγιση, για το σχηματισμό και την περιγραφή των συστάδων στηριζόμαστε στη θεωρία πιθανοτήτων, αλλά και σε έννοιες όπως ανεξαρτησία, συσχέτιση και δέντρα πιθανοτήτων (*probability-based trees*). Ένας τέτοιος αλγόριθμος είναι ο COBWEB (βλ. Fisher, 1987), τον οποίο θα παρουσιάσουμε στην ενότητα 6.1.4.

Επίσης, υπάρχουν αλγόριθμοι, όπως ο Autoclass (βλ. Cheeseman and Stutz, 1996) όπου εφαρμόζεται μέθοδος συσταδοποίησης κατά Bayes (*Bayesian clustering*). Η εκτίμηση του αριθμού συστάδων στον Autoclass γίνεται μέσω Bayesian στατιστικής ανάλυσης (βλ. Bernardo and Smith, 1994).

Για τα νευρωνικά δίκτυα σχολιάζουμε παρακάτω. Ο πιο γνωστός αλγόριθμος από αυτή την ομάδα είναι ο SOM (*self-organizing feature map* – βλ. Kohonen, 1982).

5.5 Άλλες μέθοδοι συσταδοποίησης

Με βάση τη σχετική βιβλιογραφία, υπάρχουν πολλοί τρόποι να διαχωρίσουμε τους αλγορίθμους συσταδοποίησης (βλ. Βαζιργιάννης και Χαλκίδη, 2005). Για παράδειγμα, μπορούμε να δημιουργήσουμε κατηγορίες αλγορίθμων με βάση τη μέθοδο συσταδοποίησης. Έτσι ακριβώς έγινε η διάκριση των μεθόδων στην προηγούμενη ενότητα, ενώ στην υποενότητα που ακολουθεί αναφέρουμε και κάποιες ακόμα κατηγορίες αυτού του τρόπου διαχωρισμού.

Επιπλέον, μπορούμε να διαχωρίσουμε τους αλγορίθμους με βάση τον τύπο δεδομένων που εισάγονται στον αλγόριθμο. Οι δύο κατηγορίες αλγορίθμων εδώ είναι αλγόριθμοι για αριθμητικά δεδομένα και αλγόριθμοι για δεδομένα κειμένου (ή λεκτικά δεδομένα).

Τέλος, υπάρχει ένας ακόμη τρόπος διαχωρισμού των αλγορίθμων. Στην περίπτωση αυτή κατηγοριοποιούμε τους αλγορίθμους με βάση τη θεωρία του ορισμού συστάδας. Περισσότερα σχόλια γίνονται στις ακόλουθες υποενότητες.

5.5.1 Συσταδοποίηση με βάση τη μέθοδο

Πέρα από τις κατηγορίες της προηγούμενης ενότητας, εκ των οποίων οι σημαντικότερες είναι η διαιρετική και η ιεραρχική συσταδοποίηση, παραθέτουμε και άλλες κατηγορίες αλγορίθμων με βάση τη μέθοδο συσταδοποίησης. Οι κατηγορίες αυτές είναι:

- **Ασαφής συσταδοποίηση (*Fuzzy clustering*)**

  Στην κατηγορία αυτή χρησιμοποιούνται μέθοδοι ασαφούς λογικής για την ομαδοποίηση των δεδομένων, ενώ θεωρείται ότι ένα αντικείμενο μπορεί να ταξινομηθεί σε περισσότερες από μια συστάδες. Τα σχήματα συσταδοποίησης που δημιουργούνται είναι συμβατά με την εμπειρία μας από την καθημερινή ζωή, καθώς χειρίζονται την αβεβαιότητα πραγματικών δεδομένων.

  Ο πιο σημαντικός αλγόριθμος ασαφούς συσταδοποίησης είναι ο Fuzzy C-Means. Για μια εκτενή παρουσίαση της μεθοδολογίας και των τεχνικών της ασαφούς συσταδοποίησης παραπέμπουμε στους Abonyi και Feil (2007).

- **Μη ασαφής συσταδοποίηση (*Crisp clustering*)**

  Με βάση τη μέθοδο αυτή, θεωρούνται μη επικαλυπτόμενα χωρίσματα τα οποία δείχνουν ότι ένα στοιχείο του συνόλου δεδομένων είτε ανήκει σε μια κατηγορία, είτε όχι. Εδώ ανήκουν οι περισσότεροι αλγόριθμοι συσταδοποίησης, καθώς οδηγούν σε σαφείς συστάδες.

- **Συσταδοποίηση βασισμένη στα δίκτυα Kohonen (*Kohonen net clustering*)**

  Αυτή η κατηγορία αλγορίθμων στηρίζεται στις έννοιες των νευρωνικών αλγορίθμων. Το δίκτυο Kohonen έχει κόμβους εισόδου και εξόδου. Το επίπεδο εισόδου (κόμβοι εισόδου) έχει ένα κόμβο για κάθε γνώρισμα μιας εγγραφής. Τα γνωρίσματα αυτά συνδέονται με κάθε κόμβο εξόδου.

Μεταξύ των δύο επιπέδων (είσοδος – έξοδος) υπάρχει σύνδεση, η οποία σχετίζεται με ένα βάρος που καθορίζει τη θέση του κόμβου εξόδου. Με βάση ένα σχετικό αλγόριθμο, ο οποίος αλλάζει τα βάρη, οι κόμβοι εξόδου τείνουν να σχηματίζουν συστάδες.

Περισσότερα στοιχεία για τα τα δίκτυα Kohonen και γενικά τα νευρωνικά δίκτυα μπορεί να δει κανείς σε σχετικά βιβλία, όπως αυτό του Marques de Sá (2001).

- **Συσταδοποίηση υποχώρων (*Subspace clustering)***

  Σε αυτή την κατηγορία ανήκουν αλγόριθμοι που προσπαθούν να βρουν τα υποσύνολα του αρχικού χώρου όπου τα αποτελέσματα συσταδοποίησης είναι καλύτερα.

5.5.2 Συσταδοποίηση με βάση τον τύπο δεδομένων

Όπως αναφέραμε και στην αρχή της ενότητας, με βάση αυτή τη μέθοδο διαχωρισμού των αλγορίθμων, έχουμε δύο κατηγορίες (βλ. Βαζιργιάννης και Χαλκίδη, 2005):

i) Συσταδοποίηση αριθμητικών δεδομένων

Οι αλγόριθμοι αυτής της κατηγορίας παράγουν συστάδες με βάση κάποια μέτρα αριθμητικής ομοιότητας μεταξύ των αντικειμένων και μπορούν να εφαρμοστούν σε ΒΔ με τύπο γνωρισμάτων αριθμητικές (συνεχείς) τιμές.

Για παράδειγμα, ας υποθέσουμε ότι η περιγραφή ενός ατόμου είναι:

Γνώρισμα	Ύψος	Βάρος	IQ
Τιμή	1,87	85	100

Τα αντικείμενα μπορούν να αναπαρασταθούν μέσω ενός διανύσματος. Η ομοιότητα ή η απόσταση μεταξύ των αντικειμένων υπολογίζεται με τη χρήση ενός μέτρου απόστασης (βλ. παράγραφο 5.1.1).

Στην κατηγορία αυτή ανήκει και η **στατιστική συσταδοποίηση** (*statistical clustering*), η οποία έχει τις ρίζες της στο πεδίο της στατιστικής ανάλυσης.

ii) Εννοιολογική συσταδοποίηση

Η συσταδοποίηση αυτού του τύπου μπορεί να εφαρμοστεί σε βάσεις δεδομένων με τύπο γνωρισμάτων μόνο κείμενο (*text*). Ένα αντικείμενο αυτής της μορφής μπορεί να είναι:

Γνώρισμα	Ύψος	Βάρος	IQ
Τιμή	Ψηλός	Αδύνατος	Υψηλό

Οι γεωμετρικές αποστάσεις δεν είναι κατάλληλες σε αυτή την περίπτωση. Για την εκτίμηση της απόστασης μεταξύ αντικειμένων της παραπάνω μορφής μπορούμε να χρησιμοποιήσουμε τον αριθμό γνωρισμάτων που δεν είναι κοινά στα δύο αντικείμενα (βλ. παράγραφο 5.1.2).

Για παράδειγμα, έστω τρία αντικείμενα με τα ακόλουθα διανύσματα γνωρισμάτων:

[Ψηλός, Αδύνατος, Υψηλό]

[Κοντός, Αδύνατος, Υψηλό]

[Κοντός, Βαρύς, Μέτριο]

Η απόσταση μεταξύ των πρώτων δύο αντικειμένων είναι 1, αφού διαφέρουν μόνο στην τιμή του γνωρίσματος «Ύψος». Ενώ, η απόσταση μεταξύ του πρώτου και του τρίτου αντικειμένου είναι 3, γιατί διαφέρουν και στα τρία γνωρίσματα.

5.5.3 Συσταδοποίηση με βάση τη θεωρία και τις θεμελιώδεις έννοιες

Αυτή η μέθοδος συσταδοποίησης έχει ως κριτήριο διάκρισης των αλγορίθμων τον τρόπο που η συσταδοποίηση χειρίζεται την αβεβαιότητα από την άποψη της επικάλυψης των συστάδων. Στη βιβλιογραφία συναντήσαμε αρκετούς αλγορίθμους που θα μπορούσαμε να πούμε ότι ανήκουν στη ομάδα αυτή, αλλά δεν υπήρχε σαφής υπαινιγμός για την ένταξή τους σε κάποια ομάδα.

Στο επόμενο κεφάλαιο, όπου καταγράφουμε τους υπάρχοντες αλγορίθμους συσταδοποίησης κατηγορικών δεδομένων, δίνουμε και αλγορίθμους που θεωρούμε ότι ανήκουν στην κατηγορία αυτή, καθώς λειτουργούν με βάση τη θεωρία και στηρίζονται σε θεμελιώδεις έννοιες όπως αβεβαιότητα, εντροπία κ.οκ.

ΚΕΦΑΛΑΙΟ 6

Συσταδοποίηση κατηγορικών και μικτών δεδομένων

6.1 Αλγόριθμοι συσταδοποίησης για κατηγορικά δεδομένα

Με βάση την ανασκόπησή μας σε σχετικά βιβλία, άρθρα, αλλά και ιστοσελίδες στο διαδίκτυο, είμαστε σε θέση να συμφωνήσουμε με την άποψη ότι οι τεχνικές συσταδοποίησης αποτελούν χρήσιμο εργαλείο για πολλούς επιστημονικούς και επιχειρηματικούς κλάδους. Άλλωστε, η συσταδοποίηση θεωρείται μια από τις σημαντικότερες εφαρμογές της ΕΔ.

Σύμφωνα με τους Han και Kamber (2001), μία ειδική ομάδα μεθόδων που συμπεριλαμβάνεται στις μεθόδους συσταδοποίησης, τις οποίες παρουσιάσαμε στο προηγούμενο κεφάλαιο, είναι η **συσταδοποίηση κατηγορικών δεδομένων** (*Categorical Data Clustering*). Οι αλγόριθμοι της ομάδας αυτής αναπτύσσονται ειδικά για δεδομένα όπου δε μπορεί να υπολογιστεί η ευκλείδεια απόσταση ή άλλες ανάλογες αποστάσεις που βασίζονται σε αριθμητικά δεδομένα (βλ. παράγραφο 5.1.1)

Υπάρχουν διάφορες προσεγγίσεις για την αντιμετώπιση του ζητήματος της συσταδοποίησης κατηγορικών δεδομένων, οι περισσότερες εκ των οποίων στηρίζονται σε διαχωριστικές ή ιεραρχικές μεθόδους (Andritsos, 2002). Στην ενότητα αυτή, θα παρουσιάσουμε τους χαρακτηριστικότερους αλγορίθμους που έχουν κατασκευαστεί έτσι ώστε να διαχειρίζονται αποτελεσματικά τα κατηγορικά δεδομένα.

Παραδοσιακά, η ανάπτυξη μεθόδων ανάλυσης κατηγορικών δεδομένων υστερεί χρονικά των αντίστοιχων εξελίξεων για τα συνεχή δεδομένα. Όμως, οι απαιτήσεις της εποχής καθιστούν αναγκαία την εξέλιξη του κλάδου και την ανάπτυξη αλγορίθμων για τη διαχείριση τεράστιων συνόλων δεδομένων με κατηγορικές μεταβλητές. Έχουμε ήδη αναφέρει ποιοι είναι οι τύποι δεδομένων που συγκαταλέγονται στα **κατηγορικά δεδομένα** (δίτιμα, διακριτά, διατάξιμα, ονοματικά, διαστηματικά).

Επίσης, υπάρχει μια ακόμη ομάδα δεδομένων, πέρα από τα συνεχή και τα κατηγορικά. Η ομάδα αυτή αποτελείται από τα **μικτά δεδομένα**, τα οποία σχολιάζονται στην επόμενη

ενότητα, ενώ παρουσιάζουμε και τις σχετικές μεθόδους συσταδοποίησης. Ας δούμε όμως, αρχικά, τους βασικότερους αλγορίθμους συσταδοποίησης κατηγορικών δεδομένων, έτσι όπως καταγράφονται στη σχετική βιβλιογραφία.

6.1.1 k-modes

Ο αλγόριθμος k-modes φέρεται να είναι ο πρώτος αλγόριθμος που δημοσιεύτηκε (Huang, 1998) ώστε να είναι δυνατή η διαχείριση ενός πολύ μεγάλου συνόλου δεδομένων με κατηγορικές μεταβλητές. Πρόκειται για μια επέκταση του γνωστού αλγορίθμου k-means, τον οποίο παρουσιάζουμε στο Παράρτημα Α, μιας και αφορά στη συσταδοποίηση συνεχών δεδομένων. Επομένως, είναι ένας **διαιρετικός** αλγόριθμος, όπως και ο k-means.

Η ιδέα και η δομή του k-modes δεν αλλάζουν σε σχέση με τον k-means. Η μόνη διαφορά είναι στον τρόπο μέτρησης της ομοιότητας, ώστε να έχουμε συγκρίσιμα αντικείμενα. Με αυτή την αφορμή, δημιουργήθηκαν οι παρακάτω τροποποιήσεις (βλ. Βαζιργιάννης και Χαλκίδη, 2005) σε σχέση με τον k-means:

- Χρησιμοποιούνται διαφορετικά μέτρα ανομοιότητας, τα οποία μπορούν να εφαρμοστούν σε κατηγορικά δεδομένα
- Αντικαταστάθηκαν τα k κέντρα με τα k modes
- Χρησιμοποιούνται μέθοδοι βασισμένες στη συχνότητα (*frequency-based methods*) εμφάνισης των τιμών προκειμένου να ενημερώνονται τα κέντρα των συστάδων, δηλαδή τα modes.

Για να συνειδητοποιήσουμε την έννοια του **μέτρου ανομοιότητας**, έτσι όπως χρησιμοποιείται στον αλγόριθμο αυτό, ας υποθέσουμε δύο αντικείμενα Χ και Υ. Το μέτρο ανομοιότητας μεταξύ τους μπορεί να οριστεί με βάση τη συνολική ανομοιότητα μεταξύ των κατηγοριών των χαρακτηριστικών των δύο αντικειμένων.

Όσο μικρότερος είναι ο αριθμός των αταίριαστων τιμών των αντίστοιχων γνωρισμάτων των δύο αντικειμένων, τόσο περισσότερο όμοια μπορούν να θεωρηθούν τα δύο αντικείμενα. Μια σχέση που εκφράζει τα προηγούμενα είναι (Huang, 1998):

$$d(\hat{x}, \hat{y}) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

$$\text{όπου } \delta(x_j, y_j) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases}$$

Όπως καταλαβαίνουμε, η σχέση αυτή μετρά τις αταίριαστες τιμές μεταξύ των αντίστοιχων χαρακτηριστικών των δύο αντικειμένων. Να σημειώσουμε ότι σε κάθε κατηγορία ενός χαρακτηριστικού δίνεται το ίδιο βάρος (σημαντικότητα). Λαμβάνοντας υπόψη τις συχνότητες των τιμών ενός συνόλου δεδομένων, τότε μια έκφραση για την μέτρηση της ανομοιότητας είναι:

$$d(\hat{x}, \hat{y}) = \sum_{i=1}^{n} \frac{n_{xi} + n_{yi}}{n_{xi} \cdot n_{yi}} \delta(x_i, y_i)$$

όπου n_{xi} και n_{yi} οι αριθμοί των αντικειμένων ενός συνόλου δεδομένων, τα οποία έχουν κατηγορίες x_i και y_i για το χαρακτηριστικό i, αντίστοιχα.

Το **mode** ενός συνόλου δεδομένων είναι η τιμή που εμφανίζεται περισσότερο σε αυτό. Για κάθε σύνολο δεδομένων διάστασης n, κάθε συστάδα c, με $1 \leq c \leq k$, έχει ένα mode που ορίζεται από ένα διάνυσμα $Q^c = (x_1^c, x_2^c, ..., x_n^c)$. Το σύνολο των Q^c που ελαχιστοποιούν την εξίσωση:

$$E = \sum_{c=1}^{k} \sum_{\hat{x} \in c} d(\hat{x}, Q^c) \qquad (6.1)$$

είναι το επιθυμητό αποτέλεσμα της μεθόδου. Τα βασικά βήματα του αλγορίθμου k-modes, έτσι όπως περιγράφονται από τους Βαζιργιάννη και Χαλκίδη (2005), είναι:

1. Επιλογή k αρχικών modes, ένα για κάθε συστάδα

2. Ανάθεση ενός αντικειμένου στη συστάδα της οποίας το mode είναι πιο κοντά στο αντικείμενο σύμφωνα με την απόσταση όπως ορίστηκε στην εξίσωση (6.1). Ενημέρωση του mode της συστάδας μετά από κάθε ανάθεση σύμφωνα με το θεώρημα

3. Αφού όλα τα αντικείμενα έχουν τοποθετηθεί σε συστάδες, γίνεται επανέλεγχος της ανομοιότητας των αντικειμένων ως προς τα τρέχοντα modes. Εάν για ένα αντικείμενο βρεθεί ότι βρίσκεται πιο κοντά στο mode μιας άλλης συστάδας από ότι στο mode της

τρέχουσας συστάδας, επανατοποθετείται στο αντικείμενο της άλλης συστάδας και ενημερώνονται ανάλογα τα modes των συστάδων.

4. Επαναλαμβάνεται το βήμα 3 μέχρις ότου κανένα αντικείμενο να μην αλλάζει συστάδες μετά από τον πλήρη έλεγχο όλου του συνόλου δεδομένων.

Σε αναλογία με τον αλγόριθμο k-means (βλ. Παράρτημα Α), έτσι και ο k-modes παράγει βέλτιστες λύσεις, οι οποίες εξαρτώνται από τα αρχικά modes και τη διάταξη των αντικειμένων στο σύνολο δεδομένων. Μια τεχνική για τον ορισμό των αρχικών k modes είναι να επιλεγούν οι k πρώτες εγγραφές ως τα k αρχικά modes του αλγορίθμου. Μια άλλη τεχνική (βλ. Huang, 1998) αποτελείται από τα παρακάτω βήματα:

1. Υπολογισμός των συχνοτήτων εμφάνισης όλων των κατηγοριών για όλα τα γνωρίσματα και αποθήκευση σε έναν πίνακα με φθίνουσα σειρά των συχνοτήτων, ως εξής:

$$\begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & c_{1,4} \\ c_{2,1} & c_{2,2} & c_{2,3} & c_{2,4} \\ c_{3,1} & & c_{3,3} & c_{3,4} \\ c_{4,1} & & c_{4,3} & \\ & & c_{5,3} & \end{bmatrix}$$

Στον παραπάνω πίνακα, θεωρούμε ότι έχουμε ένα σύνολο δεδομένων με $j = 4$ γνωρίσματα, τα οποία έχουν 4, 2, 5, 3 κατηγορίες αντίστοιχα. Επίσης, $c_{i,j}$ δηλώνει την κατηγορία i του γνωρίσματος j και $f(c_{i,j}) \geq f(c_{i+1,j})$ όπου $f(c_{i,j})$ είναι η συχνότητα της κατηγορίας $c_{i,j}$

2. Ανάθεση των πιο συχνών κατηγοριών ομοιόμορφα στα k αρχικά modes. Για παράδειγμα, έστω για k=3, αναθέτουμε:

$$Q_1 = [q_{1,1} = c_{1,1} \ , \ q_{1,2} = c_{2,2} \ , \ q_{1,3} = c_{3,3} \ , \ q_{1,4} = c_{1,4}]$$
$$Q_2 = [q_{2,1} = c_{2,1} \ , \ q_{2,2} = c_{1,2} \ , \ q_{2,3} = c_{4,3} \ , \ q_{2,4} = c_{2,4}]$$
$$Q_3 = [q_{3,1} = c_{3,1} \ , \ q_{3,2} = c_{2,2} \ , \ q_{3,3} = c_{1,3} \ , \ q_{3,4} = c_{3,4}]$$

3. Έναρξη με Q_l. Επιλογή της εγγραφής που είναι περισσότερο όμοια με το Q_l και αντικατάσταση του Q_l με την εγγραφή αυτή σαν το πρώτο αρχικό mode. Ομοίως για τις υπόλοιπες εγγραφές. Στόχος είναι η αποφυγή εμφάνισης κενών συστάδων

Γενικά ο στόχος αυτής της μεθόδου επιλογής είναι να έχουμε ποικιλία στα αρχικά modes, ώστε να καταλήξουμε σε καλύτερα αποτελέσματα συσταδοποίησης.

6.1.2 ROCK (*RObust Clustering using linKs*)

Ο αλγόριθμος ROCK είναι ένας **ιεραρχικός** αλγόριθμος συσταδοποίησης κατηγορικών δεδομένων. Οι Guha et al. (2000) προτείνουν μια καινοτόμο προσέγγιση, η οποία βασίζεται σε μια νέα ιδέα: τους **συνδέσμους** (*links*) μεταξύ των αντικειμένων.

Η ιδέα αυτή βοηθά να ξεπεραστούν τα προβλήματα που προκύπτουν όταν απαιτείται η χρήση μέτρων απόστασης μεταξύ διανυσμάτων, πράγμα που δεν είναι εφικτό όταν ασχολούμαστε με κατηγορικά δεδομένα. Για το σκοπό αυτό, οι Guha et al. εισάγουν τις έννοιες των γειτόνων και των συνδέσμων, οι οποίες περιγράφονται ως εξής (Βαζιργιάννης και Χαλκίδη, 2005):

- ✓ **Γείτονες**

  Οι γείτονες (*neighbors*) ενός σημείου είναι εκείνα τα σημεία τα οποία παρουσιάζουν σημαντική ομοιότητα με αυτό. Θεωρούμε την $\mathrm{sim}(p_i, p_j)$ ως τη συνάρτηση ομοιότητας με βάση την οποία εκτιμούμε την εγγύτητα μεταξύ δύο σημείων και η οποία κυμαίνεται μεταξύ του 0 και του 1.

  Η συνάρτηση μπορεί να είναι ένα οποιοδήποτε καλά ορισμένο μέτρο απόστασης ή ακόμα και μια μετρική συνάρτηση (βλ. ενότητα 5.1). Για παράδειγμα, μπορεί να είναι μια συνάρτηση ομοιότητας που παρέχεται από ειδικούς στο πεδίο που ανήκουν τα στοιχεία που συγκρίνουμε. Επομένως, δεδομένων μιας συνάρτησης ομοιότητας και ενός ορίου θ με $\theta \in [0,1]$, ένα ζεύγος σημείων p_i, p_j είναι γείτονες εάν ισχύει η ανισότητα:

  $$\mathrm{sim}(p_i, p_j) \geq \theta$$

- ✓ **Σύνδεσμοι**

  Ο σύνδεσμος $\mathrm{link}(p_i, p_j)$ ορίζεται ως ο αριθμός των κοινών γειτόνων μεταξύ των στοιχείων p_i και p_j.

Πέρα από αυτά, η αλληλοσύνδεση (*interconnectivity*) μεταξύ δύο συστάδων C_1 και C_2 δίνεται από το πλήθος των διασταυρώσεων (*cross links*) μεταξύ τους, το οποίο ισούται με:

$$\sum_{p_q \in C_1, p_r \in C_2} \text{link}(p_q, p_r)$$

Επίσης, το αναμενόμενο πλήθος των συνδέσμων σε μια συστάδα C_i είναι:

$$n_i^{1+2f(\theta)}$$

$$\text{όπου } f(\theta) = \frac{1-\theta}{1+\theta}$$

Ο αλγόριθμος ROCK μετρά την ομοιότητα δύο συστάδων, συγκρίνωντας τη συνολική αλληλοσύνδεση δύο συστάδων, σε αντίθεση με ένα στατικό μοντέλο αλληλοσύνδεσης που μπορεί να προσδιοριστεί από το χρήστη (Andritsos, 2002). Αντικείμενο του συγκεκριμένου αλγορίθμου είναι η μεγιστοποίηση της ακόλουθης έκφρασης:

$$E = \sum_{i=1}^{k} n_i \cdot \sum_{p_q, p_r \in C_i} \frac{\text{link}(p_q, p_r)}{n_i^{1+2f(\theta)}}$$

Η συνάρτηση αυτή αποκαλείται **συνάρτηση «κριτήριο»** (*criterion function*) και χρησιμοποιείται για την εκτίμηση της ποιότητας των συστάδων. Έτσι, μέσω της συνάρτησης αυτής, μπορούμε να εντοπίσουμε τις βέλτιστες συστάδες.

Όμως, για να καθορίσουμε τα καλύτερα ζευγάρια σημείων ώστε να συνδυαστούν σε κάθε βήμα του αλγορίθμου ROCK, θα χρησιμοποιήσουμε ένα παρόμοιο **μέτρο ποιότητας**. Έστω ότι η $\text{link}[C_i, C_j]$ αποθηκεύει τον αριθμό των δεσμών μεταξύ των συστάδων C_i και C_j και ότι αυτό ο αριθμός δεσμών είναι ίσος με:

$$\text{link}[C_i, C_j] = \sum_{p_q \in C_i, p_r \in C_j} \text{link}(p_q, p_r)$$

Τότε, μπορούμε να χρησιμοποιήσουμε ως μέτρο ποιότητας για τη συσχέτιση των συστάδων C_i και C_j την εξής σχέση:

$$g(C_i, C_j) = \frac{\text{link}[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \qquad (6.2)$$

Τα ζεύγη των συστάδων για τα οποία το παραπάνω μέτρο ποιότητας παίρνει τη μέγιστη τιμή είναι το καλύτερο ζεύγος των συστάδων που πρόκειται να συσχετιστούν σε κάθε βήμα του αλγορίθμου συσταδοποίησης. Στο Σχήμα 6.1 που ακολουθεί, δίνουμε μια σύνοψη του αλγορίθμου ROCK.

ΣΧΗΜΑ 6.1

Σύνοψη αλγορίθμου ROCK

[Πηγή: Guha et al., 2000]

Όπως βλέπουμε και στο Σχήμα 6.1, ο αλγόριθμος εφαρμόζεται σε ένα τυχαίο δείγμα του συνόλου δεδομένων. Αρχικά δίνουμε το σύνολο S των n σημείων του δείγματος, στο οποίο πρόκειται να γίνει συσταδοποίηση, αλλά και τον αριθμό k των συστάδων. Η διαδικασία ξεκινά με τον υπολογισμό του αριθμού των δεσμών ανάμεσα στα ζεύγη των σημείων.

Αυτά φαίνονται και μέσω του αλγορίθμου στο Σχήμα 6.2, που ακολουθεί. Ο υπολογισμός του αριθμού δεσμών γίνεται στο βήμα 1 (βλ. Σχήμα 6.2). Αρχικά, κάθε σημείο θεωρείται μια μεμονωμένη συστάδα. Για κάθε συστάδα i, χτίζουμε μια τοπική στοίβα (*build_local_heap*), την οποία διατηρούμε κατά την εκτέλεση του αλγορίθμου (βλ. Guha et al., 2000). Έστω q[i] ο συμβολισμός της στοίβας. Το q[i] περιέχει κάθε συστάδα j, έτσι ώστε link[i,j] $\neq$ 0. Οι συστάδες j στο q[i] διατάσσονται σε φθίνουσα σειρά, με βάση το μέτρο ποιότητας που δηλώθηκε στην Εξίσωση 6.2, σε σχέση με το i. Δηλαδή, το μέτρο ποιότητας θα είναι το g(i,j).

Εκτός από τη δημιουργία των στοιβάδων q[i] για κάθε συστάδα i, υπάρχει στον αλγόριθμο (Βήμα 4, Σχήμα 6.2) μια «ολική στοιβάδα» Q, η οποία περιέχει όλες τις συστάδες, οι οποίες μάλιστα διατάσσονται σε φθίνουσα σειρά με βάση τα αντίστοιχα καλύτερα μέτρα ποιότητάς τους. Έτσι, το μέτρο g(j , max(q[j])) χρησιμοποιείται για να διατάξει τις διάφορες συστάδες j στο Q, όπου max(q[j]) είναι η καλύτερη συστάδα ώστε να συγχωνευθεί με τη συστάδα j. Σε κάθε βήμα, η μέγιστη συστάδα j του σωρού Q και η μέγιστη συστάδα του q[j] είναι το ιδανικό προς συγχώνευση ζεύγος συστάδων.

Η συνθήκη **while** του Βήματος 5 (Σχήμα 6.2) επαναλαμβάνεται μέχρι τη στιγμή που θα παραμείνουν k συστάδες στον ολικό σωρό Q. Επίσης, αυτή η συνθήκη **while** διακόπτει τη

συσταδοποίηση εάν ο αριθμός των συνδέσμων ανάμεσα σε κάθε ζεύγος εναπομείνοντων συστάδων γίνεται μηδέν.

ΣΧΗΜΑ 6.2

ROCK: Αλγόριθμος συσταδοποίησης

[Πηγή: Guha et al., 2000]

```
procedure cluster(S,k)
begin
1.  link := compute_links(S)
2.  for each s ∈ S do
3.      q[s] := build_local_heap(link,s)
4.  Q := build_global_heap(S,q)
5.  while size(Q) > k do {
6.      u := extract_max(Q)
7.      v := max(q[u])
8.      delete(Q,v)
9.      w := merge(u,v)
10.     for each x ∈ q[u] ∪ q[v] do {
11.       link[x,w] := link[x,u]+link[x,v]
12.       delete(q[x],u); delete(q[x],v)
13.       insert(q[x],w,g(x,w)); insert(q[w],x,g(x,w))
14.       update(Q,x,q[x])
15.     }
16.     insert(Q,w,q[w])
17.     deallocate(q[u]); deallocate(q[v])
18. }
end
```

Σε κάθε Βήμα της συνθήκης **while** (Σχήμα 6.2), η μέγιστη συστάδα u εξάγεται από τη στοίβα Q μέσω της εντολής extract_max, ενώ το q[u] χρησιμοποιείται για τον προσδιορισμό της καλύτερης συστάδας της στοίβας Q, η οποία συμβολίζεται με v. Από τη στιγμή που οι

συστάδες u και v θα συγχωνευθούν (στο Βήμα 9), δεν απαιτούνται πλέον άλλες είσοδοι για τα u και v και μπορούν να διαγραφούν από Q. Οι συστάδες u και v που συγχωνεύονται στο Βήμα 9 (Σχήμα 6.2) δημιουργούν μια συστάδα w, η οποία περιέχει |u|+|v| σημεία. Από τη στιγμή που συγχωνεύονται οι συστάδες u και v, πρέπει να γίνουν δύο εργασίες:

1. για κάθε συστάδα που περιέχει u ή v στον τοπικό σωρό της, τα στοιχεία u και v πρέπει να αντικατασταθούν από τη συστάδα w και να ανανεωθεί ο τοπικός σωρός
2. χρειάζεται η δημιουργία ενός νέου τοπικού σωρού για τη συστάδα w

Οι δύο αυτές εργασίες περιγράφονται από τον αλγόριθμο στα Βήματα 10 έως 15, στο Σχήμα 6.2. Ο αριθμός των συνδέσμων ανάμεσα στις συστάδες x και w είναι το άθροισμα του αριθμού των συνδέσμων μεταξύ x και u και των συνδέσμων μεταξύ x και v. Αυτός ο αριθμός συνδέσμων χρησιμοποιείται για τον υπολογισμό του μέτρου ποιότητας g(x,w) των συστάδων x και w, ενώ οι δύο συστάδες εισάγονται η μια στον τοπικό σωρό της άλλης.

Να σημειώσουμε ότι το q[w] μπορεί να συμπεριλάβει μόνο συστάδες που ήταν πριν είτε στο q[u], είτε στο q[v], καθώς αυτές είναι και οι μοναδικές συστάδες που έχουν μη μηδενικούς συνδέσμους με τη συστάδα w. Επίσης, ως αποτέλεσμα της συγχώνευσης των u και v, είναι πιθανό για τη συστάδα u ή v να ήταν προηγουμένως η καλύτερη που συγχωνεύθηκε με τη x και τώρα η w γίνεται η καλύτερη προς συγχώνευση.

Επιπλέον, είναι πιθανό να μην ήταν καμμία από τις u και v η καλύτερη προς συγχώνευση με τη x, αλλά να είναι τώρα η w η καλύτερη συστάδα για να συγχωνευθεί με τη x. Για τις περιπτώσεις αυτές, όταν αλλάζει η μέγιστη συστάδα στην τοπική στοίβα του x, ο αλγόριθμος χρειάζεται να επανατοποθετήσει (*relocate*) το x στο Q, ώστε να προκύψουν πληροφορίες σχετικά με τη νέα καλύτερη συστάδα για το x (Βήμα 14, Σχήμα 6.2). Τέλος, μέσω της διαδικασίας πρέπει να είναι βέβαιο ότι το Q περιέχει την καλύτερη συστάδα προς συγχώνευση για τη συστάδα w.

Ολοκληρώνοντας την αναφορά στον αλγόριθμο ROCK, παραθέτουμε έναν αλγόριθμο που προτείνεται από τους Guha et al. (2000) ως αποτελεσματική λύση για τον υπολογισμό των συνδέσμων που απαιτείται στο Βήμα 1 του Σχήματος 6.2.

Ο αλγόριθμος αυτός δίνεται στο Σχήμα 6.3, που ακολουθεί. Στο πρώτο Βήμα του αλγορίθμου, ύστερα από τον υπολογισμό μιας λίστας των γειτόνων (*nbrlist*) κάθε σημείου i, λαμβάνονται υπόψη όλα τα ζεύγη των γειτόνων του σημείου αυτού.

ΣΧΗΜΑ 6.3

ROCK: Αλγόριθμος υπολογισμού συνδέσμων

[Πηγή: Guha et al., 2000]

```
procedure compute_links(S)
begin
1.  Compute nbrlist[i] for every point i in S
2.  Set link[i,j] to be zero for all i, j
3.  for i := 1 to n do {
4.      N := nbrlist[i]
5.      for j := 1 to |N|-1 do
6.          for l := j+1 to |N| do
7.              link[N[j],N[l]] := link[N[j],N[l]] + 1
8.  }
end
```

Για κάθε ζεύγος γειτόνων του σημείου i, υπάρχει ένας σύνδεσμος. Εάν η διαδικασία επαναληφθεί για κάθε σημείο και έχουμε αύξηση της μέτρησης στη λίστα για κάθε ζεύγος γειτόνων, τότε θα έχουμε βρει στο τέλος το πλήθος των συνδέσμων για όλα τα ζεύγη των σημείων (βλ. Σχήμα 6.3).

Με βάση τη βιβλιογραφία που μελετήσαμε, παρατηρούμε ότι οι αλγόριθμοι που είναι κοινά αποδεκτοί ως οι ικανότεροι για τη συσταδοποίηση κατηγορικών δεδομένων είναι οι ROCK και k-modes, που παρουσιάσαμε παραπάνω. Στη συνέχεια καταγράφουμε και άλλους αλγορίθμους που θεωρούνται σημαντικοί, χωρίς όμως να επεκταθούμε ιδιαίτερα στην ανάλυσή τους.

6.1.3 Ο αλγόριθμος STIRR και η βελτίωσή του: CACTUS

Σε αυτή την υποενότητα, θα παρουσιάσουμε έναν αλγόριθμο που εισάγει την ιδέα της γενικευμένης φασματικής διαμέρισης γραφήματος στη συσταδοποίηση κατηγορικών δεδομένων. Ο αλγόριθμος αυτός καλείται STIRR, θεωρείται αρκετά ισχυρός και προσφέρει τη δυνατότητα χαρτογράφησης κατηγορικών δεδομένων σε μη-γραμμικά δυναμικά συστήματα, μέσω επαναληπτικής προσέγγισης.

Επιπλέον, θα αναφέρουμε έναν αλγόριθμο που βασίζεται στον STIRR και θεωρείται η βελτίωσή του. Ο αλγόριθμος αυτός καλείται CACTUS.

STIRR (*Sieving Through Iterated Reinforcement*)

Ο αλγόριθμος STIRR είναι μια από τις πιο ισχυρές μεθόδους συσταδοποίησης κατηγορικών συνόλων δεδομένων (Gibson et al., 1998). Βασικό του στοιχείο είναι ότι χρησιμοποιεί μια προσέγγιση με βάση την οποία τα αντικείμενα θεωρούνται όμοια εάν τα στοιχεία μαζί με τα οποία εμφανίζονται στη βάση δεδομένων έχουν μια μεγάλη επικάλυψη (*overlap*), ασχέτως εάν αυτά τα αντικείμενα θα συνυπήρχαν ποτέ από μόνα τους ή όχι.

Για να κατανοήσουμε την έννοια του όρου «επικάλυψη», ας υποθέσουμε ότι έχουμε δύο τύπους αυτοκινήτου, τα Civic και Accord, μοντέλα της εταιρείας Honda για το έτος 1998. Τότε, τα διανύσματα γνωρισμάτων (*tuples*) που περιλαμβάνουν αυτά τα στοιχεία είναι:

[Honda, Civic, 1998]

[Honda, Accord, 1998]

Με βάση τον STIRR, θεωρούμε τους δύο αυτούς τύπους αυτοκινήτου όμοιους, καθώς παρατηρούμε επικάλυψη σε δύο από τα τρία στοιχεία των διανυσμάτων.

Άλλα βασικά στοιχεία της προσέγγισης των Gibson et al., πέρα από το παραπάνω, είναι (βλ. Andritsos, 2002):

1. Δεν υπάρχει **εκ των προτέρων ποσοτικοποίηση** (*a-priori quantization*). Δηλαδή, η συσταδοποίηση των κατηγορικών δεδομένων γίνεται απευθείας μέσω των προτύπων της συνύπαρξης (*co-occurrence*), χωρίς την απόπειρα έκφρασης δομής ή σχέσης.
2. Θεωρώντας κάθε σύνολο γνωρισμάτων (διάνυσμα) ως ένα σύνολο τιμών, τότε το σύνολο των διανυσμάτων αντιμετωπίζεται ως ένα νοερό σύστημα (*abstract set system*) ή υπεργράφημα (*hyper-graph*). Αυτή η θεώρηση απεικονίζεται στο Σχήμα 6.4, που παραθέτουμε στην επόμενη σελίδα.

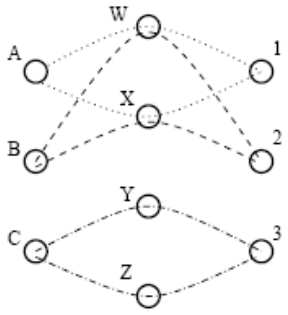
ΣΧΗΜΑ 6.4

STIRR: Παρουσίαση ενός συνόλου διανυσμάτων

[Πηγή: Gibson et al., 1998]

Tuple	Attribute a	b	c
1.	A	W	1
2.	A	X	1
3.	B	W	2
4.	B	X	2
5.	C	Y	3
6.	C	Z	3

is represented as



Πέρα από τα προηγούμενα, οι φασματικές (*spectral*) μέθοδοι συσχετίζουν τις «καλές» διαμερίσεις ενός μη κατευθυνόμενου γραφήματος με τις ιδιοτιμές και τα ιδιοδιανύσματα συγκεκριμένων πινάκων που παράγονται από το γράφημα. Ο αλγόριθμος STIRR παρέχει φασματική διαμέριση σε συσταδοποίηση υπεργραφήματος (*spectral partitioning on hyper-graph clustering*), χρησιμοποιώντας μη-γραμμικά δυναμικά συστήματα (*non-linear dynamical systems*) αντί για ιδιοδιανύσματα.

Επίσης, ο STIRR προτείνει μια μέθοδο αναπαραγωγής στάθμισης (*weight-propagation method*), με βάση την οποία (βλ. Andritsos, 2002):

- Πρώτα δίνεται μια χαμηλή στάθμιση (βάρος) σε ένα στοιχείο που μας ενδιαφέρει. Για παράδειγμα, το Honda από τα διανύσματα γνωρισμάτων που αναφέραμε προηγουμένως. Βέβαια, αυτό δεν είναι απαραίτητο, καθώς συχνά ως αρχικό βάρος θεωρείται η μονάδα.
- Αυτή η στάθμιση μεταδίδεται σε στοιχεία που συνυπάρχουν με το παραπάνω στοιχείο, που στην περίπτωσή μας είναι το Honda, από τα παραπάνω διανύσματα γνωρισμάτων.
- Τα στοιχεία αυτά μεταδίδουν αλλού τη συγκεκριμένη στάθμιση, μέχρι να τερματιστεί η διαδικασία.

Η στάθμιση αυξάνεται στην περίπτωση που θέλουμε να επικεντρωθούμε σε κάποιο βάρος. Στο άρθρο των Gibson et al. περιλαμβάνονται ορισμένα θεωρήματα από τη φασματική θεωρία γραφημάτων, τα οποία είναι απαραίτητα για την απόδειξη της αποτελεσματικότητας του STIRR, ακόμη και όταν υπάρχει στάθμιση με διαφορετικό πρόσημο μεταξύ κάποιων τιμών.

CACTUS (*Clustering Categorical Data Using Summaries*)

Ο συγκεκριμένος αλγόριθμος (Ganti et al., 1999) αποτελεί μια βελτίωση του STIRR. Βασική ιδέα του CACTUS είναι ότι η συνολική πληροφορία που εκφράζεται από το σύνολο δεδομένων είναι επαρκής για την ανακάλυψη καλώς ορισμένων συστάδων.

Σε αντίθεση με τον STIRR, ο CACTUS μπορεί να ανακαλύψει ιδιαίτερους τύπους συστάδων, όπως συστάδες που υπερκαλύπτουν η μια την άλλη από πλευράς αποτελεσμάτων ή συστάδες που μοιράζονται την ίδια προβολή (*projection*), δηλαδή η μία αποτελεί προβολή της άλλης. Μια σύνοψη του CACTUS αποτελείται από τα παρακάτω στάδια (Andritsos, 2002):

1. **Σύνοψη (*Summarization*)**

   Υπάρχουν δύο τύποι σύνοψης:

   α) Σύνοψη μεταξύ μεταβλητών (*inter-attribute summaries*): μετρά όλα τα ζεύγη τιμών των μεταβλητών που συνδέονται ισχυρά, από διαφορετικές μεταβλητές

   β) Σύνοψη εντός των μεταβλητών (*intra-attribute summaries*): υπολογισμός των ομοιοτήτων μεταξύ τιμών της ίδιας μεταβλητής

2. **Συσταδοποίηση (*Clustering*)**

   Στη φάση αυτή αναλύεται κάθε μεταβλητή, έτσι ώστε να υπολογιστούν όλες οι απεικονίσεις των συστάδων. Η απεικόνιση, συμβ. S_i^j, κάθε μεταβλητής A_i είναι ένα υποσύνολο τιμών της μεταβλητής, το οποίο συνδέεται ισχυρά με τις τιμές της μεταβλητής για κάθε άλλη μεταβλητή A_j, με $i \neq j$.

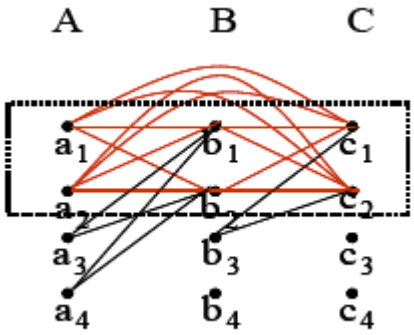
   Για παράδειγμα, ας παρατηρήσουμε το Σχήμα 6.5. Θεωρούμε τη μεταβλητή Α. Τότε, μπορούμε να υπολογίσουμε τα $S_A^B = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ και $S_A^C = \{\alpha_1, \alpha_2\}$. Στην περίπτωση αυτή, η απεικόνιση του Α είναι $S_A^B \cap S_A^C$.

   Συνεχίζοντας στη φάση συσταδοποίησης, υπολογίζονται οι υποψήφιες συστάδες πάνω σε σύνολα συστάδων, με βάση τις απεικονίσεις των συστάδων των μεμονωμένων μεταβλητών. Το στάδιο της συσταδοποίησης θεωρείται ότι επεκτείνει το προηγούμενο, καθώς αυξάνει κατά ένα τη διάσταση των συστάδων.

ΣΧΗΜΑ 6.5

Συσταδοποίηση στον CACTUS

[Πηγή: Andritsos, 2002]



3. **Επικύρωση (*Validation*)**

   Ο αλγόριθμος αναγνωρίζει τις λανθασμένα υποψήφιες συστάδες, ελέγχοντας εάν η υποστήριξη από κάθε υποψήφια συστάδα είναι μεγαλύτερη από το απαιτούμενο όριο (*threshold*).

Στην ακόλουθη υποενότητα, θα παρουσιάσουμε συνοπτικά άλλους αλγορίθμους που υπάρχουν στη βιβλιογραφία και έχουν σχεδιαστεί για τη διενέργεια συσταδοποίησης κατηγορικών δεδομένων. Πρόκειται για τους πιο αξιόλογους και αποτελεσματικούς αλγορίθμους που έχουν προταθεί, κυρίως κατά την τελευταία δεκαετία.

6.1.4 Άλλοι αλγόριθμοι για κατηγορικά δεδομένα

Η αναφορά των αλγορίθμων θα γίνει με χρονολογική σειρά. Παράλληλα με την παρουσίαση των αλγορίθμων, παραπέμπουμε στα σχετικά άρθρα για την εκτενέστερη μελέτη τους. Οι αλγόριθμοι αυτοί είναι οι εξής:

COBWEB

Ο αλγόριθμος COBWEB προτάθηκε από τον Fisher (1987). Πρόκειται για έναν αλγόριθμο που ανήκει στην **εννοιολογική συσταδοποίηση** (*conceptual clustering*) και μάλιστα στην ομάδα των αλγορίθμων που βασίζονται σε μοντέλο (*model based*). Ο COBWEB προσφέρει μια μη επιβλεπόμενη, αυξητική (*incremental*) ταξινόμηση των εννοιών (*concepts*) σε κατηγορικά σύνολα δεδομένων, με κατεύθυνση «από πάνω προς τα κάτω» (*top-down*).

Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί μια μετρική, η οποία καλείται «**χρησιμότητα κατηγορίας**» (*category utility*) και συμβάλλει στην αξιολόγηση κάθε νέου εισερχόμενου κόμβου, έτσι ώστε να τοποθετηθεί στο κατάλληλο σημείο στην ιεραρχία. Όμως, ο COBWEB έχει και αρκετούς περιορισμούς. Οι βασικότεροι είναι οι εξής (Han and Kamber 2001):

- Απαιτείται η υπόθεση ότι οι κατανομές πιθανότητας των γνωρισμάτων είναι μεταξύ τους στατιστικά ανεξάρτητες
- Η επαναλαμβανόμενη παρουσίαση των κατανομών πιθανότητας ανά συστάδα ανεβάζει το κόστος και μετατρέπει τη μέθοδο σε μια ακριβή διαδικασία
- Το δέντρο που δημιουργείται στηριζόμενο στην πιθανότητα (*probability-based tree*), με στόχο τον εντοπισμό των συστάδων, δεν χαρακτηρίζεται από ιδιαίτερη ισσοροπία (σταθερότητα)

Squeezer

Ο αλγόριθμος Squeezer είναι ένας αποτελεσματικός αλγόριθμος, ο οποίος απαιτεί μια ανάγνωση (πέρασμα) των δεδομένων (*one-pass algorithm*). Η ιδιότητά του είναι ότι διαβάζει το διάνυσμα των γνωρισμάτων (*tuple*) ενός αντικειμένου διαδοχικά για κάθε αντικείμενο, ένα προς ένα. Όταν ολοκληρώσει την ανάγνωση του πρώτου διανύσματος γνωρισμάτων, δημιουργεί μια μεμονωμένη συστάδα.

Τα επόμενα σύνολα γνωρισμάτων τοποθετούνται σε υπάρχουσες συστάδες ή απορρίπτονται από αυτές και δημιουργούν μια νέα συστάδα, δοθείσης της συνάρτησης ομοιότητας (*similarity function*) μεταξύ συνόλου γνωρισμάτων και συστάδας. Για περισσότερες λεπτομέρειες, παραπέμπουμε στους He et al. (2002).

COOLCAT

Το 2002 προτάθηκε από τους Barbará et al. ένας αλγόριθμος που στηρίζεται στη θεωρία της πληροφορίας (*information-theoretic algorithm*). Ο αλγόριθμος COOLCAT έχει πολλές ομοιότητες με τον k-means, από πλευράς αντίληψης. Όμως, περιλαμβάνει μια αρχική φάση δειγματοληψίας, όπου στόχος είναι η επιλογή των k κατάλληλων αντιπροσώπων των συστάδων.

Ως μέτρο ομοιότητας στον COOLCAT χρησιμοποιείται η **εντροπία** (*entropy*), με στόχο να εκτιμηθεί η ομοιότητα μεταξύ των αντικειμένων. Η εντροπία θεωρείται ένα μέτρο της αβεβαιότητας που υπάρχει κατά την πρόβλεψη των τιμών μιας τυχαίας μεταβλητής.

Η αντικειμενική συνάρτηση του αλγορίθμου επιχειρεί να ελαχιστοποιήσει την εντροπία μεταξύ των συστάδων. Δεδομένου του επιθυμητού πλήθους συστάδων, ο αλγόριθμος έχει ως αντικείμενο τη διαμέριση του συνόλου δεδομένων, με την ελάχιστη εντροπία μεταξύ των τελικών συστάδων, ή αλλιώς με τη μέγιστη βεβαιότητα να προβλεθούν οι τιμές που συμπεριλήφθηκαν σε κάθε συστάδα..

Ο αλγόριθμος COOLCAT στηρίζεται στη δειγματοληψία και είναι **μη-ιεραρχικός** (Barbará et al., 2002). Βασική προϋπόθεση είναι να θεωρηθεί ότι κάθε αντικείμενο είναι ανεξάρτητο του άλλου. Στην αρχή του αλγορίθμου, προσδιορίζεται η εντροπία μιας συστάδας και επιλέγεται ένα δείγμα από τα σημεία. Ύστερα, με τη χρήση μιας ευρετικής μεθόδου (*heuristic method*), εντοπίζεται ένα σύνολο των k αρχικών συστάδων, έτσι ώστε οι εντροπίες ανά ζεύγη να είναι μεγάλες (Andritsos, 2004).

Στο επόμενο βήμα του αλγορίθμου, τοποθετούνται τα εναπομείναντα σύνολα γνωρισμάτων (*tuples*) των υπόλοιπων αντικειμένων του συνόλου δεδομένων σε συστάδες, με στόχο την ελαχιστοποίηση της συνολικής εντροπίας. Για την έναρξη του αλγορίθμου απαιτούνται το πλήθος k των συστάδων και το μέγεθος του αρχικού δείγματος, καθώς θεωρείται ότι η ποιότητα της συσταδοποίησης βασίζεται σε αυτό.

LIMBO (*scaLable InforMation BOttleneck clustering*)

Ο συγκεκριμένος αλγόριθμος προτάθηκε από τους Andritsos et al. (2003) και μοιάζει περισσότερο στον COOLCAT, σε σχέση με τους υπόλοιπους αλγορίθμους για κατηγορικά δεδομένα. Επίσης, δανείζεται κάποια υπολογιστικά στοιχεία από τον BIRCH, ο οποίος είναι κατάλληλος για συνεχή δεδομένα.

Πρόκειται για έναν **ιεραρχικό** αλγόριθμο με δυνατότητες **κλιμάκωσης** (*scalable*). Είναι ο πρώτος αλγόριθμος που κατασκευάστηκε βασιζόμενος στην μέθοδο **IB** (*Information Bottleneck*) και αφορά στα κατηγορικά δεδομένα. Η μέθοδος IB αποτελεί εναλλακτικό τρόπο έκφρασης ενός μέτρου απόστασης για κατηγορικά διανύσματα γνωρισμάτων (βλ. Tishby et al., 1999).

Ο LIMBO έχει τη δυνατότητα να χρησιμοποιηθεί για τη συσταδοποίηση τόσο συνόλων γνωρισμάτων όσο και μεμονωμένων τιμών των γνωρισμάτων, ανάλογα με τις απαιτήσεις της έρευνας. Επίσης, μπορεί να διαχειριστεί πολύ μεγάλα σύνολα δεδομένων, δημιουργώντας ένα συνοπτικό μοντέλο (*summary model*) για τα δεδομένα. Τέλος, οι Andritsos et al. (2003)

δίνουν δύο διαφορετικές εκδόσεις του LIMBO, ανάλογα με το εάν θέλουμε να ελέγξουμε το μέγεθος ή την ακρίβεια του μοντέλου.

k-histogram

Πρόκειται για ένα νέο αποτελεσματικό αλγόριθμο συσταδοποίησης κατηγορικών δεδομένων, ο οποίος προτάθηκε το από τους He et al. (2005-a). Ο k-histogram επεκτείνει τον k-means στον τομέα των κατηγορικών δεδομένων, αντικαθιστώντας τα κέντρα των συστάδων με ιστογράμματα. Κατά τη διάρκεια της διαδικασίας συσταδοποίησης, τα ιστογράμματα αναβαθμίζονται δυναμικά.

Σε αντίθεση με τον αλγόριθμο k-modes, που παρουσιάσαμε παραπάνω, ο k-histogram χρησιμοποιεί τη δομή που περιγράφεται από τα ιστογράμματα για να περιγράψει τις συστάδες και όχι τα modes. Όμως, ο k-histogram έχει και πολλές ομοιότητες με τον k-modes. Με βάση την αναλυτική σύγκριση των He et al., τα αποτελέσματα της συσταδοποίησης από την εφαρμογή του k-histogram ήταν πολύ ακριβή.

Στα πλαίσια της αναζήτησης αλγορίθμων συσταδοποίησης κατηγορικών δεδομένων, συγκεντρώσαμε αρκετό υλικό και αλγορίθμους, οι περισσότεροι από τους οποίους παρουσιάστηκαν παραπάνω. Για παράδειγμα, οι αλγόριθμοι που θεωρούνται περισσότερο αποδεκτοί και αποτελεσματικοί είναι οι k-modes (Huang, 1998), ROCK (Guha et al., 2000), STIRR (Gibson et al., 2000), καθώς και ο CACTUS (Ganti et al., 1999).

Όμως, αξίζει να αναφέρουμε και τη δυνατότητα συσταδοποίησης μέσω του αλγορίθμου **EM** (*Expectation – Maximization*), που εισήγαγαν οι Dempster et al. (1977). Πρόκειται για μια μέθοδο διαιρετικής συσταδοποίησης., η οποία αναθέτει τυχαία διαφορετικές πιθανότητες σε κάθε κατηγορία κάθε συστάδας. Στη συνέχεια, οι πιθανότητες αυτές προσαρμόζονται ώστε να μεγιστοποιηθεί η πιθανοφάνεια των δεδομένων, δοθέντος του αριθμού συστάδων (βλ. Parmar et al., 2007).

Στον αλγόριθμο EM θεωρείται ότι κάθε παρατήρηση ανήκει σε κάθε συστάδα με μια συγκεκριμένη πιθανότητα, καθώς ο αλγόριθμος αυτός υπολογίζει τις πιθανότητες ταξινόμησης (*classification probabilities*). Η πραγματική ανάθεση των παρατηρήσεων σε μια συστάδα προσδιορίζεται με βάση τη μεγαλύτερη πιθανότητα ταξινόμησης. Σύμφωνα με τον Andritsos (2002), ο EM μπορεί να χειριστεί σύνολα δεδομένων με μικτά δεδομένα.

Στο σημείο αυτό, αξίζει να αναφέρουμε ένα σημαντικό συμπέρασμα των Halkidi et al. (2001): για όλους τους αλγορίθμους υπάρχει μια κοινή υπόθεση, ότι «κάθε αντικείμενο μπορεί να ταξινομηθεί μόνο σε μια συστάδα και όλα τα αντικείμενα έχουν τον ίδιο βαθμό εμπιστοσύνης (*degree of confidence*) όταν ομαδοποιούνται σε μια συστάδα». Βέβαια, μια τέτοια υπόθεση είναι σχεδόν απίθανο να επαληθευθεί στην πραγματικότητα. Σε μια εφαρμογή από τον πραγματικό κόσμο, είναι δύσκολο να υπάρξουν σαφή όρια μεταξύ των συστάδων, καθώς υπάρχει πάντα το πρόβλημα της **αβεβαιότητας**.

Μια πρώτη απόπειρα αντιμετώπισης του προβλήματος αυτού ήταν μέσω εφαρμογής **ασαφών συνόλων** (*fuzzy sets*) στην συσταδοποίηση κατηγορικών δεδομένων. Ο Huang (1998) πρότεινε τον αλγόριθμο **fuzzy K-modes**, ενώ οι Kim et al. (2004) επιχείρησαν την επέκτασή του. Το μειονέκτημα αυτών των αλγορίθμων είναι ότι απαιτούν πολλές επαναλήψεις.

Στη συνέχεια, στα πλαίσια της ανάγκης δημιουργίας ενός **εύρωστου** αλγορίθμου για κατηγορικά δεδομένα, ώστε να επιτευχθεί η διαχείριση της αβεβαιότητας, προτάθηκε από τους Parmar et al. (2007) ένας αλγόριθμος που βασίζεται στη θεωρία ασαφών συνόλων (*Rough Set Theory*). Ο αλγόριθμος αυτός καλείται **MMR** (*Min-Min Roughness*), και έχει την ικανότητα να διαχειρίζεται την αβεβαιότητα κατά τη συσταδοποίηση κατηγορικών δεδομένων, καθώς και να δίνει σταθερά αποτελέσματα, δοθέντος του επιθυμητού αριθμού συστάδων. Επίσης, ο αλγόριθμος αυτός μπορεί να χειρίζεται μεγάλα σύνολα δεδομένων.

Κλείνοντας, αναφέρουμε και κάποιους άλλους αλγορίθμους, παραπέμποντας στα σχετικά άρθρα. Πέρα από τους αλγορίθμους k-histogram και Squeezer, που ξεχωρίσαμε παραπάνω, οι He et al. (2004) έχουν προτείνει έναν ακόμα αλγόριθμο, τον **LCBCDC** (*Link Clustering Based Categorical Data Clustering*). Ο αλγόριθμος αυτός βασίζεται στη συσταδοποίηση συνδέσμων (*link clustering*). Επίσης, αναφέρουμε τον αλγόριθμο **CLICK**, που προτάθηκε από τους Peters et al. (2004) και τον **k-representatives**, μια ακόμα παραλλαγή του k-means, που προτάθηκε από τους San et al. (2004).

Τέλος, ένας άλλος αλγόριθμος που βασίζεται στη θεωρία Rough Set και προτάθηκε από τους Chen et al. (2006) είναι ο **RAHCA** (*Rough Set-Based Agglomeration Hierarchy Clustering Algorithm*).

6.2 Αλγόριθμοι συσταδοποίησης για μικτά δεδομένα

Σε αντίθεση με άλλους αλγορίθμους, ο k-means και οι παραλλαγές του προσαρμόζονται καλά στη διαδικασία ΕΔ, καθώς εμφανίζουν υψηλή αποδοτικότητα στην επεξεργασία μεγάλων συνόλων δεδομένων. Όπως έχει αποδειχθεί, οι ιεραρχικοί αλγόριθμοι αντιμετωπίζουν πρόβλημα στη συσταδοποίηση μεγάλων συνόλων δεδομένων λόγω της πολυπλοκότητάς τους. Όμως, ο αλγόριθμος k-means αποδεικνύεται πολύ χρήσιμος στη διαχείριση μεγάλων συνόλων δεδομένων.

Στην προηγούμενη ενότητα, παρουσιάσαμε τον αλγόριθμο k-modes. Ο αλγόριθμος αυτός αποτελεί μια από τις πιο γνωστές παραλλαγές του k-means, αλλά μπορεί να εφαρμοστεί σε σύνολα δεδομένων που περιέχουν μόνο κατηγορικά δεδομένα. Ένα συχνό φαινόμενο είναι να έχουμε να χειριστούμε ένα σύνολο δεδομένων που να περιλαμβάνει συνεχή και κατηγορικά χαρακτηριστικά (Huang, 1998). Αυτή είναι μια ειδική ομάδα δεδομένων, τα οποία αποκαλούνται **μικτά**.

Στη βιβλιογραφία υπάρχουν αρκετοί αλγόριθμοι και προτεινόμενες τεχνικές που επιχειρούν να συσταδοποιήσουν μικτά δεδομένα. Ο πιο γνωστός αλγόριθμος αποτελεί επίσης παραλλαγή του k-means και καλείται **k-prototypes**. Στην ουσία, ο k-prototypes είναι ένας συνδυασμός του k-means που χειρίζεται αριθμητικά δεδομένα και του k-modes, ο οποίος χειρίζεται κατηγορικά δεδομένα.

6.2.1 K-prototypes

Ο αλγόριθμος k-prototypes σχεδιάστηκε για τη διενέργεια συσταδοποίησης σε μεγάλα σύνολα δεδομένων με αριθμητικά και κατηγορικά δεδομένα (Huang, 1997). Η φιλοσοφία είναι ίδια με τον k-means, αλλά το μέτρο ανομοιότητας που ορίζεται σε αυτό τον αλγόριθμο λαμβάνει υπόψη γνωρίσματα τόσο με συνεχείς όσο και με κατηγορικές τιμές.

Έστω s^r το μέτρο της ανομοιότητας σε αριθμητικά χαρακτηριστικά και s^c το μέτρο ανομοιότητας σε κατηγορικά. Ως ανομοιότητα στα κατηγορικά δεδομένα εννοούμε τον αριθμό αταίριαστων κατηγοριών μεταξύ δύο αντικειμένων. Το μέτρο ανομοιότητας ανάμεσα στα δύο αντικείμενα ορίζεται ως (βλ. Andritsos, 2002):

$$s^r + \gamma \cdot s^c \tag{6.3}$$

όπου γ είναι ένα βάρος για τη δημιουργία ισορροπίας μεταξύ των δύο μερών και την αποφυγή εύνοιας κάποιον από τους τύπους των χαρακτηριστικών.

Το γ είναι μια παράμετρος που προσδιορίζεται από το χρήστη, αλλά αυτό συχνά θεωρείται ένα πρόβλημα για την εφαρμογή του συγκεκριμένου αλγορίθμου. Μια πρόταση για την αντιμετώπιση αυτού του προβλήματος είναι να χρησιμοποιείται σαν βάση για την επιλογή του βάρους η τυπική απόκλιση των αριθμητικών γνωρισμάτων.

Η ουσιαστική διαφορά του αλγορίθμου k-prototypes από τον k-means είναι η νέα μέθοδος που χρησιμοποιείται για την ενημέρωση των λεκτικών τιμών των προτύπων (κέντρων) των συστάδων. Τα βήματα του αλγορίθμου k-prototypes είναι όμοια με αυτά που δώσαμε παραπάνω για τον k-modes, με τη μόνη διαφορά ότι έχουμε αλλαγή στην εξίσωση μέτρησης της ανομοιότητας. Η νέα εξίσωση βασίζεται στη Σχέση 6.3.

Στα δύο σχήματα που ακολουθούν δίνουμε τον αλγόριθμο k-prototypes σε φάσεις, οι οποίες σύμφωνα με τον Huang (1997) είναι τρεις:

1. Αρχική επιλογή των prototypes
2. Πρώτος καταμερισμός
3. Επανάληψη καταμερισμού

Στην πρώτη φάση, επιλέγονται τυχαία k αντικείμενα ως τα αρχικά prototypes των συστάδων. η δεύτερη φάση απεικονίζεται μέσω του σχετικού αλγορίθμου στο σχήμα 6.6, όπου X[i] είναι το αντικείμενο i και X[i,j] η τιμή της μεταβλητής j για το i αντικείμενο. Επίσης, στα O_prototypes[] και C_prototypes[] αποθηκεύονται αντίστοιχα τα αριθμητικά και κατηγορικά μέρη των prototypes των συστάδων. Δηλαδή, τα O_prototypes[i,j] και C_prototypes[i,j] είναι δύο στοιχεία του prototype της συστάδας i, το πρώτο είναι αριθμητικό και το δεύτερο κατηγορικό.

Ακόμη, Distance() είναι η τετραγωνική ευκλείδεια απόσταση (παράγραφος 5.1.1) και Sigma() η απόσταση που δηλώθηκε στη Σχέση 6.3.

ΣΧΗΜΑ 6.6

k-protoypes: αρχικός καταμερισμός

[Πηγή: Huang, 1997]

```
FOR i=1 To NumberOfObjects
   Mindistance=Distance(X[i],O_prototypes[1])+gamma*Sigma(X[i],C_prototypes[l])
   FOR j=1 TO NumberOfClusters
     distance=Distance(X[i],O_prototypes[j])+gamma*Sigma(X[i],C_prototypes[j])
     IF(distance<Minidistance)
        Minidistance-distance
        cluster=j
     ENDIF
   ENDFOR
   Clustership[i]=cluster
   ClusterCount[cluster]+1
   FOR j=1 TO NumberOfNumericAttributes
     SumInCluster[cluster,j]+X[i,j]
     O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster]
   ENDFOR
   FOR j=1 TO NumberOfCategoricAttributes
      FrequencyInCluster[cluster,j,X[i,j]]+1
      C_prototypes[cluster,j]=HighestFreq(FrequencyInCluster,cluster,j)
   ENDFOR
ENDFOR
```

Στο Σχήμα 6.7 δίνεται ο αλγόριθμος της τρίτης φάσης, όπου γίνεται επανάληψη του καταμερισμού. Η διαδικασία μοιάζει με αυτή της προηγούμενης φάσης, με τη διαφορά όμως ότι αναβαθμίζονται τα prototypes τόσο των προηγούμενων, όσο και των νέων συστάδων του αντικειμένου.

Το πλήθος των αντικειμένων που άλλαξαν συστάδα κατά τη διαδικασία καταγράφεται από τη μεταβλητή moves.

ΣΧΗΜΑ 6.7

k-protoypes: επανάληψη καταμερισμού

[Πηγή: Huang, 1997]

```
moves=0
FOR i=1 TO NumberOfObjects
    …
    (To find the cluster whose prorotype is the nearest to object i. Same as previous)
    …
    IF(Clustership[i]<>cluster)
        moves+1
        oldcluster=Clustership[i]
        ClusterCount[cluster]+1
        ClusterCount[oldcluster]-1
        FOR j=1 TO NumberOfNumericAttributes
            SumInCluster[cluster,j]+X[i,j]
            SumInCluster[oldcluster,j]-X[i,j]
            O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster]
            O_prototypes[oldcluster,j]=SumInCluster[oldcluster,j]/ClusterCount[oldcluster]
        ENDFOR
        FOR j=1 TO NumberOfCategoricAttributes
            FrequencyInCluster[cluster,j,X[i,j]]+1
            FrequencyInCluster[oldcluster,j,X[i,j]]-1
            C_prototypes[cluster,j]=HighestFreq(cluster,j)
            C_prototypes[oldcluster,j]=HighestFreq(oldcluster,j)
        ENDFOR
    ENDIF
ENDFOR
```

Η διαδικασία που περιγράφει ο αλγόρυθμος του Σχήματος 6.7 συνεχίζεται μέχρι τη στιγμή που δεν θα καταγραφεί καμία μετακίνηση από τη μεταβλητή moves, δηλαδή όταν θα έχουμε moves=0.

6.2.2 Άλλοι αλγόριθμοι για μικτά δεδομένα

Αναζητώντας αλγορίθμους συσταδοποίησης μικτών δεδομένων, βρήκαμε και άλλες προτάσεις αλγορίθμων. Για παράδειγμα, μια άλλη πρόταση ήταν ο αλγόριθμος **SBAC** (Li and Biswas, 2002), ο οποίος υιοθετεί ένα μέτρο ομοιότητας που δίνει μεγαλύτερο βάρος σε ανόμοιους συνδυασμούς τιμών των χαρακτηριστικών.

Πρόκειται για έναν συσσωρευτικό (*agglomerative*) αλγόριθμο, με βάση τον οποίο κατασκευάζεται ένα δενδρόγραμμα. Όμως, όσο μεγαλώνει το πλήθος των καταγραφόμενων αντικειμένων στο σύνολο δεδομένων, τόσο πιο περίπλοκος γίνεται ο **SBAC**. Για την ακρίβεια, θεωρείται σχεδόν απίθανο να μπορέσει να διαχειριστεί ένα μεγάλο σύνολο δεδομένων (He et al., 2005-b).

Επιπλέον, ένας ακόμη αλγόριθμος προτάθηκε από τους Chiu et al. (2001), ο οποίος μάλιστα είναι επίσημα διαθέσιμος μέσω του **Clementine 6.0**. Το μέτρο της απόστασης παράγεται μέσω ενός μοντέλου πιθανοτήτων (*probabilistic model*), όπου η απόσταση ανάμεσα σε δύο συστάδες ισούται με τη μείωση του λογαρίθμου της συνάρτησης πιθανοφάνειας, η οποία προκύπτει ως αποτέλεσμα της συγχώνευσης. Ο αλγόριθμος αυτός βασίζεται στον BIRCH, ο οποίος είναι κατάλληλος για συνεχή δεδομένα, χρησιμοποιώντας όπως αυτή τη μετρική της απόστασης.

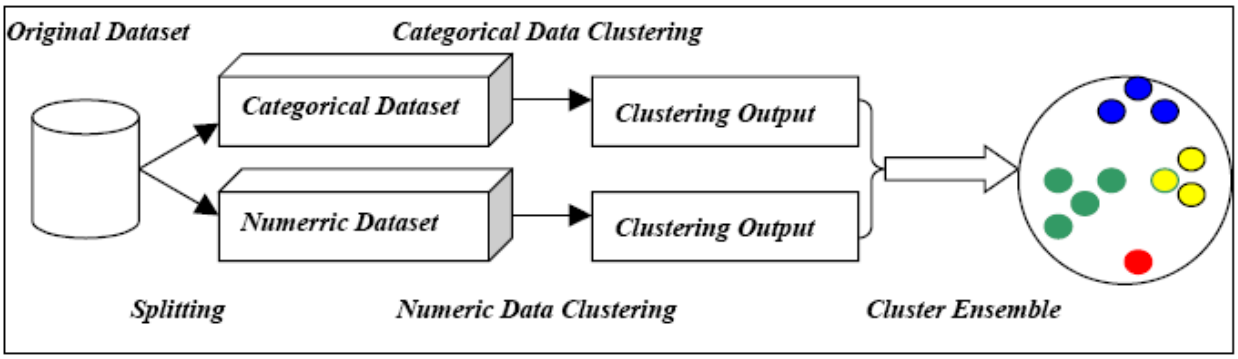
Τέλος, κλείνουμε με έναν αλγόριθμο που προτάθηκε και πάλι από τους He et al. (2005-b). Ο αλγόριθμος αυτός καλείται **algCEBMDC** και βασίζεται σε ένα πλαίσιο εργασίας που καλείται **CEBMDC** (*Cluster Ensemble Based Mixed Data Clustering*). Η πρόταση των He et al. είναι να γίνει διαχωρισμός του μικτού συνόλου δεδομένων σε δύο **σύνολα συστάδων** (*cluster ensembles*). Η μέθοδος των συνόλων συστάδων προτείνει το συνδυασμό διαφορετικών αλγορίθμων συσταδοποίησης, ώστε να υπάρξει μια διαμέριση του αρχικού συνόλου δεδομένων, με στόχο τη δημιουργία ενιαίων αποτελεσμάτων τα οποία θα προκύψουν ύστερα από μια σειρά μεμονωμένων αποτελεσμάτων συσταδοποίησης (Strehl and Ghosh, 2002).

Η βάση της ιδέας των συγγραφέων ήταν η διαμέριση του αρχικού συνόλου μικτών δεδομένων σε δύο υποσύνολα: ένα **συνεχές** και ένα **κατηγορικό**. Σε κάθε υποσύνολο, δηλαδή, θα περιέχονται οι ανάλογες μεταβλητές. Ύστερα, εφαρμόζεται ένας κατάλληλος αλγόριθμος συσταδοποίησης για κάθε τύπο δεδομένων, όπως φαίνεται και στο ακόλουθο σχήμα (βλ. Σχήμα 6.8).

ΣΧΗΜΑ 6.8

algCEBMDC: κεντρική ιδέα

[Πηγή: He et al., 2005-b]



Σαν τελευταίο βήμα έχουμε τη συνένωση των αποτελεσμάτων συσταδοποίησης των δυο υποσυνόλων δεδομένων σε ένα κατηγορικό σύνολο δεδομένων, στο οποίο εφαρμόζεται ο αλγόριθμος συσταδοποίησης κατηγορικών δεδομένων που εφαρμόστηκε και πιο πριν. Έτσι, ορίζονται και οι τελικές συστάδες.

ΣΧΗΜΑ 6.9

algCEBMDC: βήματα αλγορίθμου

[Πηγή: He et al., 2005-b]

Algorithm *algFC-CEBMDC*

Input: *D* // the data set

Output: *Each Data Object Identified with a Cluster Label*

1. Splitting the dataset *D* into categorical dataset (*CD*) and numeric dataset (*ND*)
2. Clustering *CD* using Categorical Data Clustering Algorithm (*Squeezer*, ROCK, etc.)
3. Clustering *ND* using Numeric Data Algorithm (CURE, CHAMELEON, etc.)
4. Combining the outputs of above algorithms into a categorical dataset (*CombinedCD*)
5. Clustering *CombinedCD* using *Squeezer* algorithm (or alternative effective algorithms)

Τα βήματα του αλγορίθμου algCEBMDC φαίνονται στο Σχήμα 6.9, σε μορφή μεθοδολογίας. Οι He et al. προτείνουν τον αλγόριθμο Squeezer (He et al., 2002) για τη συσταδοποίηση των κατηγορικών δεδομένων, καθώς θεωρούν ότι δίνει ικανοποιητικά αποτελέσματα και έχει δυνατότητες κλιμάκωσης.

6.3 Σχολιασμός και σύγκριση αλγορίθμων

Η συσταδοποίηση είναι ευρέως αποδεκτή ως ένα χρήσιμο εργαλείο για πολλές εφαρμογές. Πρόκειται για ένα δύκολο ζήτημα που έχει εξεταστεί από ερευνητές πολλών κλάδων, καθώς συνδυάζει έννοιες ετερόκλητων επιστημονικών πεδίων. Στην ενότητα αυτή, θα συγκρίνουμε τους σημαντικότερους από τους αλγορίθμους συσταδοποίησης που παραθέσαμε σε αυτό το κεφάλαιο και είναι ικανοί να προβούν σε συσταδοποίηση κατηγορικών ή μικτών δεδομένων.

Αρχικά, παρουσιάσαμε τον αλγόριθμο k-modes (Huang, 1998). Πρόκειται για έναν διαιρετικό αλγόριθμο, ο οποίος είναι παραλλαγή του k-means (παράρτημα Α.1), ώστε να χειρίζεται κατηγορικά δεδομένα. Στόχος του είναι η ανακάλυψη συστάδων, ενώ υιοθετεί νέες έννοιες, όπως την αντικατάσταση των κέντρων των συστάδων με τα «modes». Επίσης, εισάγει ένα νέο μέτρο ανομοιότητας για την εξέταση των κατηγορικών δεδομένων.

Δύο αδύνατα σημεία των διαιρετικών αλγορίθμων, άρα και του k-modes, είναι ότι δε μπορούν να χειριστούν το θόρυβο και τις έκτροπες παρατηρήσεις, καθώς και ότι δε μπορούν να χειριστούν συστάδες αυθαίρετου σχήματος. Ακόμη, ο k-modes και η πλειοψηφία των διαιρετικών αλγορίθμων βασίζονται σε μια συγκεκριμένη υπόθεση για να προβούν σε διαμέριση του συνόλου δεδομένων, με αποτέλεσμα να απαιτούν να προσδιορίσουν εκ των προτέρων τον αριθμό συστάδων.

Ο ROCK (Guha et al., 2000) είναι ένας αντιπροσωπευτικός ιεραρχικός αλγόριθμος συσταδοποίησης κατηγορικών δεδομένων. Η νέα έννοια που εισάγει είναι τα «links» (δεσμοί ή σύνδεσμοι), ώστε να μετρήσει την ομοιότητα / εγγύτητα ανάμεσα σε ζεύγη σημείων. Έτσι, η μέθοδος συσταδοποίησης του ROCK βασίζεται σε μη μετρικά κριτήρια ομοιότητας, τα οποία είναι εφαρμόσιμα σε κατηγορικά σύνολα δεδομένων.

Ένα προσόν του ROCK είναι ότι παρουσιάζει καλές ιδιότητες κλιμάκωσης (*scaling*) σε σχέση με τους παραδοσιακούς αλγορίθμους οι οποίοι χρησιμοποιούν τεχνικές τυχαίας δειγματοληψίας (*sampling*). Επίσης, φαίνεται να χειρίζεται επιτυχώς σύνολα δεδομένων που παρουσιάζουν σημαντικές διαφορές στο μέγεθος των συστάδων (Βαζιργιάννης και Χαλκίδη,

2005). Ας δούμε όμως, στον Πίνακα 6.1, τη σύγκριση των δύο σημαντικότερων αλγορίθμων συσταδοποίησης κατηγορικών δεδομένων.

ΠΙΝΑΚΑΣ 6.1

Σύγκριση k-modes και ROCK

Αλγόριθμος	Τύπος	Αφηρημένα σχήματα συστάδων	Outliers	Αποτελέσματα
k-modes	Διαιρετικός	Όχι	Όχι	Modes συστάδων
ROCK	Ιεραρχικός	Ναι	Ναι	Ανάθεση των δεδομένων στις ομάδες

Ένα κοινό των αλγορίθμων του Πίνακα 6.1 είναι ότι απαιτούν ως παράμετρο εισόδου τον αριθμό των συστάδων. Όμως, ο ROCK δουλεύει εντελώς διαφορετικά απ'ότι ο k-modes, όχι μόνο επειδή είναι ιεραρχικός, αλλά επειδή δουλεύει και σε δείγματα των δεδομένων (Andritsos, 2002). Βέβαια, ο k-modes έχει καλύτερες δυνατότητες κλιμάκωσης από τον ROCK. Σύμφωνα με τον Andritsos (2002), το μόνο αρνητικό του ROCK είναι ότι τα αποτελέσματα βασίζονται κατά πολύ στη δειγματοληψία που διενεργεί.

Όσον αφορά τον STIRR (Gibson et al., 1998), ένα βασικό μειονέκτημά του είναι ότι απαιτεί πολλές παραμέτρους εισόδου, ώστε να ορίσει ένα δυναμικό σύστημα (*dynamical system*). Επίσης, υπάρχει πρόβλημα στον ορισμό των τελικών συστάδων, ενώ τα τελικά αποτελέσματα επηρρεάζονται έντονα από την αρχική στάθμιση. Από την άλλη πλευρά, ο STIRR είναι ιδιαίτερα γρήγορος και δίνει ικανοποιητικά αποτελέσματα.

Ο αλγόριθμος CACTUS (Ganti et al., 1999), είναι η βελτίωση του STIRR. Ο CACTUS είναι ένας κλιμακωτός (*scalable*) αλγόριθμος, ο οποίος απαιτεί μόνο μια ανάγνωση των δεδομένων. Για τη φάση της επικύρωσης χρειάζεται άλλη μια ανάγνωση, αλλά δεν προκαλούνται επιπλοκές ή προβλήματα στην ικανότητα για κλιμάκωση (*scalability*).

Με βάση αποτελέσματα πειραματισμών, ο CACTUS είναι πιο αποτελεσματικός από τον STIRR όσον αφορά το χρόνο εκτέλεσης και τον αριθμό των μεταβλητών που μπορεί να χειριστεί. Όμως, έχει το μειονέκτημα ότι δε μπορεί να χειριστεί σύνολα δεδομένων με αυξανόμενο πλήθος διαστάσεων.

Ένας άλλος αλγόριθμος για κατηγορικά δεδομένα είναι ο COOLCAT (Barbará et al., 2002). Ο αλγόριθμος αυτός βασίζεται στην ιδέα του k-means και στην εντροπία (*entropy*), που χρησιμοποιείται για την έκφραση της ποιότητας συσταδοποίησης. Ο COOLCAT έχει αλγοριθμικές ομοιότητες με τους k-modes και k-prototypes. Όμως, δείχνει να είναι ευαίσθητος στη φάση της δειγματοληψίας, καθώς είναι πιθανή η επιλογή έκτροπων παρατηρήσεων ως αντιπρόσωποι των συστάδων (Andritsos, 2004). Ένα πλεονέκτημά του είναι ότι μπορεί να χειριστεί μεγάλα σύνολα δεδομένων.

Ο σημαντικότερος αλγόριθμος για τη συσταδοποίηση μικτών δεδομένων είναι ο k-prototypes (Huang, 1997). Προκείται για μια ακόμη επέκταση του k-means. Ένα πλεονέκτημα του k-prototypes είναι ότι έχει δυνατότητες κλιμάκωσης. Όμως, δε μπορεί να χειριστεί ικανοποιητικά σύνολα δεδομένων με έκτροπες παρατηρήσεις (βλ. Andritsos, 2002).

Στον Πίνακα 6.2 συνοψίζονται τα στοιχεία από τη σύγκριση των βασικότερων αλγορίθμων που αναφέρθηκαν στο παρών κεφάλαιο. Επίσης, η πολυπλοκότητα των αλγορίθμων του Πίνακα 6.2 σχολιάζεται στο Παράρτημα Β.

ΠΙΝΑΚΑΣ 6.2

Σύγκριση αλγορίθμων

Αλγόριθμος	**Παράμετροι εισόδου**	**Κατάλληλος για…**	**Διαχείριση έκτροπων παρατηρήσεων**
k-modes	Αριθμός συστάδων	Σύνολα δεδομένων με καλώς ορισμένες συστάδες	Όχι
ROCK	Αριθμός συστάδων	Μικρά σύνολα δεδομένων με θόρυβο	Ναι
STIRR	Αρχική διάρθρωση Τελεστής σύνδεσης Κριτήρια τερματισμού	Μεγάλα σύνολα δεδομένων με θόρυβο	Ναι
CACTUS	Όριο υποστήριξης Όριο επικύρωσης	Μεγάλα σύνολα δεδομένων με μικρή διάσταση και μικρό εύρος γνωρισμάτων (*attribute domain size*)	Ναι
COOLCAT	Μέγεθος αρχικού δείγματος Αριθμός συστάδων	Μεγάλα σύνολα δεδομένων με καλώς ορισμένες συστάδες	Όχι
k-prototypes	Αριθμός συστάδων	Μικτά σύνολα δεδομένων	Όχι

ΚΕΦΑΛΑΙΟ 7

Data Mining, κατηγορικά δεδομένα

και σχετικό λογισμικό

7.1 Διαθέσιμο λογισμικό εξόρυξης δεδομένων

Στις μέρες μας, παρατηρούμε τον όγκο των δεδομένων που αποθηκεύονται ψηφιακά να αυξάνει διαρκώς. Αυτό σημαίνει ότι είναι απαραίτητη η ύπαρξη του κατάλληλου λογισμικού και υπολογιστικών συστημάτων, τα οποία θα εξυπηρετήσουν στη διατήρηση και αξιοποίηση της χρήσιμης πληροφορίας

Στην ενότητα αυτή, θα παρουσιάσουμε τα σημαντικότερα προγράμματα που εφαρμόζουν τεχνικές της εξόρυξης δεδομένων (ΕΔ), έτσι όπως παρουσιάζονται στο διαδίκτυο, καθώς δε μπορέσαμε να έχουμε πρόσβαση στα περισσότερα από αυτά..

Μια από τις πιο αξιόλογες πλατφόρμες είναι το **Clementine**, το οποίο αποτελεί προϊόν της **SPSS** (http://www.spss.com/clementine/). Περιλαμβάνει πολλά εργαλεία ΕΔ, ενώ δίνει έμφαση στην μοντελοποίηση πρόβλεψης (βλ. παράγραφο 2.2.2), έχοντας ως στόχο τη βελτιστοποίηση των διαδικασιών λήψης αποφάσεων. Επίσης, στο Clementine περιλαμβόνονται οι δύο πιο σύγχρονες εφαρμογές της ΕΔ: η εξόρυξη κειμένου (*Text Mining*), δηλαδή η εξόρυξη γνώσης από λεκτικά δεδομένα, καθώς και η εξόρυξη από τον παγκόσμιο ιστό (*Web Mining* – βλ. ενότητα 5.3).

Βέβαια, οφείλουμε να αναφέρουμε και το **SAS** ως ένα από τα πιο σημαντικά προγράμματα πρακτικής εφαρμογής της ΕΔ. Πρόκειται για ένα λογισμικό που προσφέρει λύσεις επιχειρηματικής ευφυΐας (*Business Intelligence*) και προβλεπτικής ανάλυσης (*Predictive Analytics*). Οι δυνατότητες του SAS είναι ποικίλες, εμείς όμως θα περιοριστούμε στη δυνατότητα που προσφέρει για εξόρυξη δεδομένων και εξόρυξη κειμένου (http://www.sas.com/technologies/analytics/datamining/index.html).

Τα παραπάνω δύο προγράμματα, αν και θεωρούνται από τα πιο σημαντικά, δεν είναι τα μόνα διαθέσιμα. Ένα διαδομένο σχετικό πρόγραμμα είναι και ο **XL-MINER** (http://www.resample.com/xlminer/). Πρόκειται για μια add-in εφαρμογή του **Excel**. Τον XL-Miner τον εφαρμόσαμε, χρησιμοποιώντας μια δοκιμαστική έκδοσή του (*trial demo*). Στην παράγραφο 7.3.1 ακολουθεί εκτενέστερος σχολιασμός.

Επίσης, υπάρχουν προγράμματα που είναι ελεύθερα στο διαδίκτυο. Για παράδειγμα, στιν παράγραφο 7.3.2 παρουσιάζουμε το πρόγραμμα **Weka**, το οποίο αναπτύχθηκε από το Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία (http://www.cs.waikato.ac.nz/ml/weka/). Τέλος, η **R** (http://www.r-project.org/) είναι μια ελεύθερη γλώσσα, με πάρα πολλές δυνατότητες. Χρησιμοποιώντας το ανάλογο πακέτο (*package*), μπορεί κανείς να τρέξει κάποιον αλγόριθμο και να βγάλει χρήσιμα συμπεράσματα (βλ. παράγραφο 7.3.3).

Τα αποτελέσματα της χρήσης από τα προγράμματα που παρουσιάζουμε στην ενότητα 7.3 δεν ήταν απόλυτα ικανοποιητικά. Εναλλακτικά, παραθέτουμε στην ενότητα που ακολουθεί τα αποτελέσματα από την εφαρμογή μεθοδολογιών ΕΔ σε ένα κατηγορικό σύνολο δεδομένων, μέσω του **SQL Server 2005**, τον οποίο χρησιμοποιήσαμε στα πλαίσια της συνεργασίας μας με τραπεζικό όμιλο.

Ο SQL Server 2005 (http://www.microsoft.com/sql/prodinfo/overview/default.mspx) είναι ένα ολοκληρωμένο λογισμικό διαχείρισης και ανάλυσης δεδομένων, το οποίο είναι περιεκτικό και κατανοητό στη χρήση. Οι δυνατότητές του διευκολύνουν τους οργανισμούς και επιχειρήσεις να έχουν πιο άμεσα αποτελέσματα από τις διαθέσιμες πληροφορίες τους, ώστε να γίνονται πιο ανταγωνιστικοί.

Η ΕΔ στον SQL Server 2005 είναι ένα προϊόν που αναπτύχθηκε από κοινού από την ομάδα της Microsoft SQL Server και από την Microsoft Research, κυρίως από όσους απάρτιζαν την ομάδα μηχανικής εκμάθησης και εφαρμοσμένης στατιστικής (*Machine Learning and Applied Statistics Group – MLAS*). Στόχος ήταν η ολοκλήρωση των τεχνολογιών ΕΔ και βάσεων δεδομένων (βλ. Tang and MacLennan, 2005), καθώς η προσπάθεια είχε ξεκινήσει από την εισαγωγή αλγορίθμων ΕΔ και εργαλείων απεικόνισης (*visualization tools*) στον SQL Server 2000.

7.2 Εφαρμογή εξόρυξης δεδομένων σε στεγαστικά δάνεια

Στόχος αυτής της ενότητας είναι να συνειδητοποιήσουμε τη χρησιμότητα της πρακτικής εφαρμογής των αλγορίθμων συσταδοποίησης. Το ενδιαφέρον μας επικεντρώνεται στα κατηγορικά δεδομένα. Για το λόγο αυτό, ως μέρος της εργασίας σε τραπεζικό όμιλο, επιχειρήσαμε την εφαρμογή των μεθοδολογιών της ΕΔ σε ένα σύνολο δεδομένων με κατηγορικές μεταβλητές.

Το σύνολο δεδομένων που επιχειρήσαμε να επεξεργαστούμε περιείχε 120.000 κατόχους στεγαστικού δανείου από την επιχειρηματική μονάδα στεγαστικής πίστης της τράπεζας. Για αυτά τα άτομα (αντικείμενα), συλλέχθηκαν ορισμένα στοιχεία στις 31/12/2006. Για την ακρίβεια, καταγράφηκαν δεδομένα σε συνολικά 46 μεταβλητές ανά άτομο. Δηλαδή, το αρχείο δεδομένων περιείχε αρχικά 120.000 γραμμές και 46 στήλες.

Σε μια εκ των μεταβλητών αυτών καταγραφόταν εάν έφυγαν οι πελάτες οποιαδήποτε στιγμή μέσα στο 2007. Η μεταβλητή αυτή ονομάστηκε «Outcome» και είναι δίτιμη (ναι / όχι), καθώς δήλωνε εάν αποχώρησε ο πελάτης από την τράπεζα ή όχι. Στόχος της ανάλυσης ήταν η δημιουργία ενός μοντέλου με βάση το οποίο θα προβλέπαμε τι μπορεί να γίνει κατά το 2008.

Δηλαδή, συγκεντρώσαμε τα στοιχεία των πελατών την τελευταία μέρα του 2006, είδαμε τι κίνηση έκαναν μέσα στο 2007 και με βάση το μοντέλο που κατασκευάστηκε, θα διενεργήσουμε πρόβλεψη για τη συμπεριφορά κατά το 2008 των «νέων» πελατών, των οποίων τα χαρακτηριστικά συλλέχθηκαν στις 31/12/2007. Φυσικά, ως αποκριτική μεταβλητή θεωρήσαμε τη μεταβλητή Outcome.

Αναμφισβήτητα πρόκειται για μια μεγάλη βάση δεδομένων. Όμως, για τις ανάγκες της εργασίας μας, δημιουργήσαμε ένα «υποθετικό» αρχείο 3.000 γραμμών και 46 στηλών (τυχαίο δείγμα από το αρχικό), καθώς δε θα μπορούσαμε να αποκαλύψουμε ευαίσθητα προσωπικά δεδομένα. Ας δούμε, όμως, ποιες είναι οι διαδικασίες που ακολουθήσαμε ώστε να ανακαλύψουμε τις ομάδες πελατών που αναμένεται να αποχωρήσουν από την επιχειρηματική μονάδα στεγαστικής πίστης, αποπληρώνοντας το στεγαστικό τους δάνειο.

7.2.1 Ανάπτυξη μοντέλου διακράτησης πελατών

Για να εντοπίσουμε τις ομάδες των «επικίνδυνων» πελατών, δηλαδή αυτών που μπορεί να αποπληρώσουν το δάνειό τους και να αποχωρήσουν, διαμορφώσαμε ένα μοντέλο ΕΔ στον Microsoft SQL Server 2005 (βλ. Tang and MacLennan, 2005). Παράλληλα, εκτός από τα

χαρακτηριστικά της κάθε ομάδας, μέσα από το μοντέλο αυτό βλέπουμε και την πιθανότητα προς αποχώρηση της κάθε ομάδας,

Να σημειώσουμε ότι η ανάπτυξη αυτού του μοντέλου από τον Server δεν υποδεικνύει ότι χρησιμοποιήθηκε μόνο μια μέθοδος περιγραφής ή πρόβλεψης, αλλά ένας συνδυασμός τους. Άλλωστε, όπως σχολιάσαμε και σε προηγούμενα κεφάλαια, η συσταδοποίηση μπορεί, για παράδειγμα, να προηγηθεί της ταξινόμησης. Στην επόμενη υποενότητα θα επιχειρήσουμε να εφαρμόσουμε μόνο συσταδοποίηση και θα δούμε τα αποτελέσματα που προέκυψαν.

Πριν εισάγουμε τα δεδομένα στον Server, εφαρμόσαμε τεχνικές αξιολόγησης πιστοληπτικής ικανότητας (*credit scoring*) ανά μεταβλητή (βλ. Siddiqi, 2006), ώστε να επιλέξουμε τις μεταβλητές που προσφέρουν μονοδιάστατα τη μεγαλύτερη πληροφορία και να τις εισάγουμε εν συνεχεία στο μοντέλο ΕΔ. Η μεθοδολογία αυτή εφαρμόστηκε στο Excel, όπου διαχωρίσαμε τους πελάτες σε «καλούς» και «κακούς», ανάλογα με τα εάν έμειναν ή αποχώρησαν μέσα στο 2007 από τη στεγαστική πίστη, αντίστοιχα (βλ Πίνακα 7.1).

Στους Πίνακες 7.1 και 7.2 δίνουμε ενδεικτικά τα αποτελέσματα από την επεξεργασία μίας μεταβλητής, ενώ πρέπει να σημειώσουμε ότι η ίδια διαδικασία πραγματοποιήθηκε 45 φορές, δηλαδή, μια φορά για κάθε μεταβλητή, εκτός της μεταβλητής Outcome. Εξάλλου, τα αποτελέσματα των πράξεων στις υπόλοιπες 45 μεταβλητές προκύπτουν με βάση τη μεταβλητή Outcome.

ΠΙΝΑΚΑΣ 7.1

Κατασκευή Credit Risk Scoreboard

	«Καλοί» πελάτες			«Κακοί» πελάτες		
Staff	Συχνότητα	Ποσοστό % γραμμής	Ποσοστό % στήλης	Συχνότητα	Ποσοστό % γραμμής	Ποσοστό % στήλης
Σύνολο	**2672**	**89,07%**	**100,00%**	**328**	**10,93%**	**100,00%**
ΟΧΙ	2541	88,63%	95,10%	326	11,37%	99,39%
ΝΑΙ	131	98,50%	4,60%	2	1,50%	0,61%

Η μεταβλητή που μελετήσαμε στον Πίνακα 7.1 είναι η Staff, η οποία είναι δίτιμη (ναι / όχι) και δηλώνει εάν ο κάτοχος του στεγαστικού δανείου είναι υπάλληλος του τραπεζικού ομίλου ή όχι.

Στη συνέχεια, στον Πίνακα 7.2 βρίσκουμε το odds «καλοί προς κακοί πελάτες» για κάθε κατηγορία της μεταβλητής Staff, ενώ ως Information Odds εννοούμε το αντίστοιχο odds, υπολογισμένο με βάση τις στήλες «Ποσοστό % στήλης».

Το WOE (Πίνακας 7.2) είναι ένας συντελεστής βαρύτητας που υπολογίζεται ανά κατηγορία της εξεταζόμενης μεταβλητής και συνολικά. Είναι συντομογραφία της φράσης «Weight of Evidence» και υπολογίζεται από τον τύπο (βλ. Siddiqi, 2006):

$$\text{Ln}(\frac{\text{DistGood}}{\text{DistBad}})*100 = \text{Ln}(\text{InformationOdds})*100$$

όπου DistGood το «Ποσοστό % στήλης» των «καλών» πελατών, δηλαδή η κατανομή (*distribution*) των καλών. Ομοίως για το DistBad.

ΠΙΝΑΚΑΣ 7.2

Ανάλυση χαρακτηριστικών

	Ανάλυση χαρακτηριστικών			
Staff	Overall Odds	Information Odds	WOE	Information Value
Σύνολο	**8,15**	**1,00**	**0,00%**	**0,09**
OXI	7,79	0,96	-4,42%	0,00
NAI	65,50	8,04	208,45%	0,09

Να σημειώσουμε ότι το συνολικό WOE θα είναι πάντα 0,00% όπως και στον Πίνακα 7.2, αφού ln(1)=0, όπου 1 το συνολικό Information Odds. Η πληροφορία (*information value*) που εκφράζει η μεταβλητή Staff δίνεται από τον τύπο (βλ. Siddiqi, 2006):

$$\text{InformationValue} = \sum_{i=1}^{n}(\text{DistGood}_i - \text{DistBad}_i)*\text{Ln}(\frac{\text{DistGood}_i}{\text{DistBad}_i})$$

$$\text{ή}\ \ \text{InformationValue} = \sum_{i=1}^{n}(\text{DistGood}_i - \text{DistBad}_i)*\text{WOE}_i$$

Από τον Πίνακα 7.2 βλέπουμε ότι η μεταβλητή Staff προσφέρει μονοδιάστατα 9% πληροφορία. Ακολουθώντας τη μεθοδολογία δημιουργίας ενός Credit Risk Scoreboard (βλ.

Siddiqi, 2006), δηλαδή ενός πίνακα αξιολόγησης του πιστωτικού κινδύνου (Πίνακες 7.1 και 7.2), επιλέξαμε τις μεταβλητές που έχουν Information Value πάνω από 5% για να εισαχθούν στο μοντέλο, καθώς για αυτές τις μεταβλητές θεωρείται ότι προσφέρουν μονοδιάστατα επαρκή πληροφορία.

Άρα, η μεταβλητή Staff εισάγεται στο μοντέλο. Εκτός από τη μεταβλητή αυτή, εισάγονται άλλες επτά, οι οποίες καταγράφονται στον Πίνακα 7.3. Στον ίδιο πίνακα δίνουμε και την επεξήγηση των κατηγοριών κάθε επιλεχθείσας μεταβλητής. Να σημειώσουμε όμως, ότι δεν θεωρούνταν εξαρχής κατηγορικές όλες οι μεταβλητές.

Για παράδειγμα, μια μεταβλητή όπως η Interest (βλ. Πίνακα 7.3) που υποδηλώνει το επιτόκιο του δανείου και σε κάθε ένα από τα 3.000 αντικείμενα μπορεί να έχει ως τιμή ένα άλλο ποσοστό, δεν έχει νόημα να αντιμετωπίζεται ως συνεχής.

Η δημιουργία των κατηγοριών ανά μεταβλητή (βλ. Πίνακα 7.3) έγινε μέσω ομαδοποιήσεων στο Excel, ανάλογα με τις τιμές της στήλης «Ποσοστό % γραμμής» των «καλών» πελατών (βλ. Πίνακα 7.1). Οι τιμές της κάθε μεταβλητής που ενώνονταν σε συγκεκριμένες ομάδες ήταν αυτές που είχαν κοντινά ποσοστά στη συγκεκριμένη στήλη.

Στον Πίνακα 7.3 βλέπουμε τις κατηγορίες που δημιουργήθηκαν. Έτσι, όλες οι μεταβλητές που εισάχθηκαν στο μοντέλο, αντιμετωπίστηκαν ως κατηγορικές.

ΠΙΝΑΚΑΣ 7.3

Κατηγορίες των υπό μελέτη μεταβλητών

Μεταβλητή	Κατηγορίες	Επεξήγηση
Staff	no	Όχι υπάλληλος της τράπεζας
	yes	Υπάλληλος της τράπεζας
Rate_Type	fixed at this time	Σταθερό επιτόκιο κατά την τρέχουσα περίοδο
	floating at this time	Κυμαινόμενο επιτόκιο κατά την τρέχουσα περίοδο (ήταν σταθερό πριν)
	fixed or floating until end of duration	Σταθερό ή κυμαινόμενο ως το τέλος της διάρκειας
REM_FixedPeriod	0-24 months	(*Remaining Fixed Period*) Πλήθος εναπομείνοντων μηνών με σταθερό επιτόκιο
	25+ months	
Interest	Null - 4%	Επιτόκιο δανείου (Null: καταγγελμένα δάνεια ή δάνεια που έχουν περάσει το maturity, δηλ. έχουν αρνητικό interest rate)
	4 - 6%	
	6+ %	
NIM	0 - 0.5% or Null	(*Net Interest Margin*) Καθαρό περιθώριο κέρδους (Null: αρνητικό NIM)
	0.5 - 2.5%	
	2.5 - 6%	
Installment	0 - 150	Ποσό δόσης σε €
	150 - 300	
	300 - 700	
	700 - 2500	
	2500+	
MaxState_Ever	-1 or 0	Μέγιστη περίοδος καθυστέρησης πληρωμής δόσης -1: δε χρωστάει τίποτα 0: χρωστάει πυρασφαλιστήρια (αλλά όχι δόση) 1: χρωστάει έως 30 μέρες (1 μήνα) 2 to 7: χρωστάει από 2 έως 7 μήνες
	1	
	2 to 7	
Purpose	Buy / Buy Land / Construction / Refinance	Αγορά / Αγορά Οικοπέδου / Κατασκευή / Επαναπροσδιορισμός δανείου
	Repair / Home Equity	Επισκευή / Ιδιοκτησία ακινήτου
Outcome	no	Δεν αποχώρησε από τράπεζα μέσα στο 2007
	yes	Αποχώρησε από τράπεζα μέσα στο 2007

Πέρα από την εισαγωγή των επιλεγμένων μεταβλητών με τις διαμορφωμένες κατηγορίες, ο SQL Server ζητά τον ορισμό της μεταβλητής απόκρισης. Αυτή, όπως αναφέραμε και προηγουμένως, ήταν η μεταβλητή Outcome, αφού μας ενδιαφέρει να προβλέψουμε το

αποτέλεσμα της κίνησης των πελατών, τα στοιχεία των οποίων συλλέχθηκαν στις 31/12/2007 (ένα χρόνο μετά από την πρώτη καταγραφή).

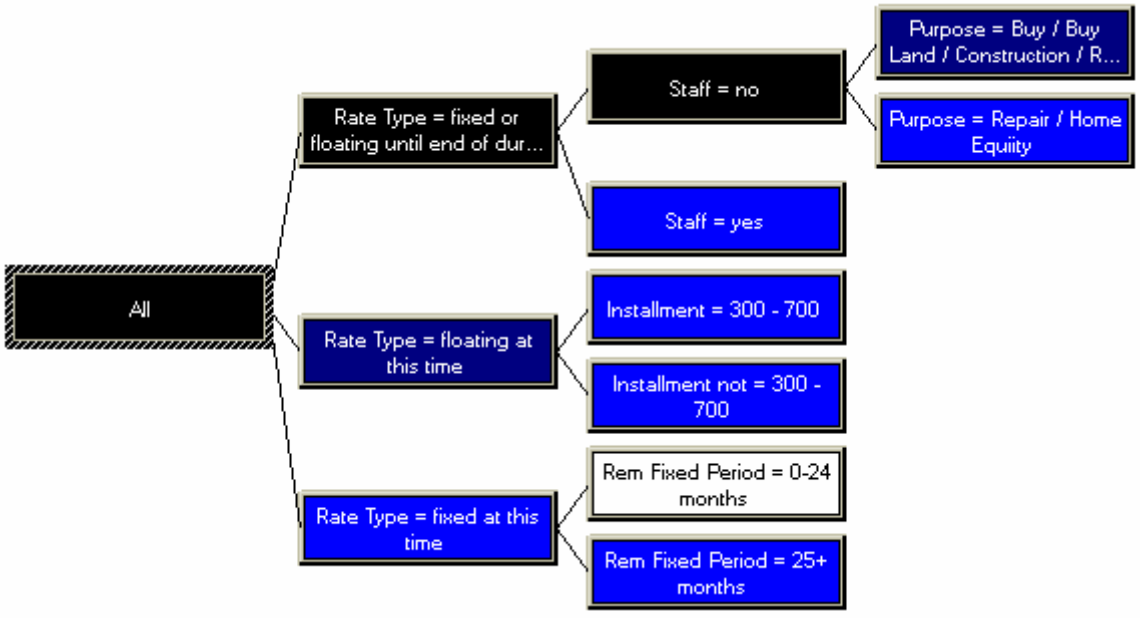
Η πρόβλεψη που θέλουμε να κάνουμε είναι εάν αυτοί οι πελάτες θα αποχωρήσουν ή όχι μέσα στο 2008 και με ποια πιθανότητα. Εκτός αυτού, επειδή το δείγμα που εισάγαμε θεωρείται μικρό, ορίσαμε στον Server ότι κατά την εφαρμογή του δέντρου απόφασης, κάθε «κουτί» θα περιέχει το λιγότερο 25 περιπτώσεις (*cases*), αριθμός που θεωρείται αρκετά μικρός.

Επίσης, ως complexity ορίσαμε την τιμή 0,05. Το complexity παίρνει τιμές από 0 έως 1 και όσο πιο κοντά στο 0 είναι η τιμή που βάζουμε, τόσο πιο «αυστηρός» γίνεται ο Server κατά τη δημιουργία του δέντρου. Αυτό που ζητήσαμε από τον Server είναι η δημιουργία ενός «Data Mining Model», ενώ στην επόμενη υποενότητα θα δώσουμε τα αποτελέσματα από την επιλογή «Clustering Model».

Το αποτέλεσμα που πήραμε ήταν το παρακάτω δέντρο απόφασης (βλ. Σχήμα 7.1). Τα σκούρα χρώματα υποδεικνύουν ότι το συγκεκριμένο μονοπάτι (*path*) είναι πιο ισχυρό. Δηλαδή, εκεί βρίσκουμε τις πιο πολλές παρατηρήσεις. Στον Πίνακα 7.4 αναφέρουμε τα αποτελέσματα της πιο σημαντικής διαδρομής.

ΣΧΗΜΑ 7.1

Δέντρο απόφασης



ΠΙΝΑΚΑΣ 7.4

Αποτελέσματα Microsoft SQL Server 2005

Node Path		
Staff=no and RateType=fixed or floating until end of duration and Purpose=Buy/BuyLand/Construction/Refinance		
(Node Total)	**1380**	**100,00%**
No	1250	90,46%
Yes	130	9,47%
Missing	0	0,07%

Αυτό σημαίνει ότι, από το συγκεκριμένο μονοπάτι (Node Path), αναμένεται να αποχωρήσει το 9,47% των πελατών. Αυτοί είναι πελάτες που δεν αποτελούν προσωπικό της τράπεζας, έχουν σταθερό ή κυμαινόμενο επιτόκιο ως το τέλος της διάρκειας του δανείου τους, ενώ ζήτησαν το δάνειο για αγορά, αγορά οικοπέδου, κατασκευή ή επαναπροσδιορισμό δανείου.

Να σημειώσουμε ότι με βάση το σύνολο των ατόμων, ο Server δηλώνει ότι αναμένεται να αποχωρήσει το 10,93%, ποσοστό που φαίνεται και στον Πίνακα 7.1 (ποσοστό % γραμμής / «κακοί» πελάτες). Αυτός είναι ο λόγος που η συγκεκριμένη διαδρομή έχει σκούρο χρώμα στο Σχήμα 7.1, αφού τα ποσοστά αποχώρησης είναι πολύ κοντά.

7.2.2 Συσταδοποίηση στον Microsoft SQL Server 2005

Πέρα από την επιλογή του «Data Mining Model», ο Server δίνει και τη δυνατότητα δημιουργίας ενός «Clustering Model». Κάνοντας αυτή την επιλογή, επιχειρήσαμε να προβούμε σε συσταδοποίηση των δεδομένων των στεγαστικών δανείων.

Ο επιθυμητός αριθμός συστάδων είναι το μόνο ζητούμενο του Server στην περίπτωση δημιουργίας «Microsoft Clustering», για το οποίο δεν καταφέραμε να συλλέξουμε πληροφορίες για τον αλγόριθμο συσταδοποίησης που χρησιμοποιείται κ.λπ. Να αναφέρουμε, επίσης, ότι στο βιβλίο των Tang και MacLennan (2005) δεν γίνεται ιδιαίτερα εκτενής αναφορά στο κομμάτι της συσταδοποίησης. Όμως, από το συγκεκριμένο βίβλιο πληροφορηθήκαμε ότι οι αλγόριθμοι πάνω στους οποίου βασίζεται ο Server είναι ο k-means (βλ. Παράρτημα Α) και ο EM (βλ. παράγραφο 6.1.4).

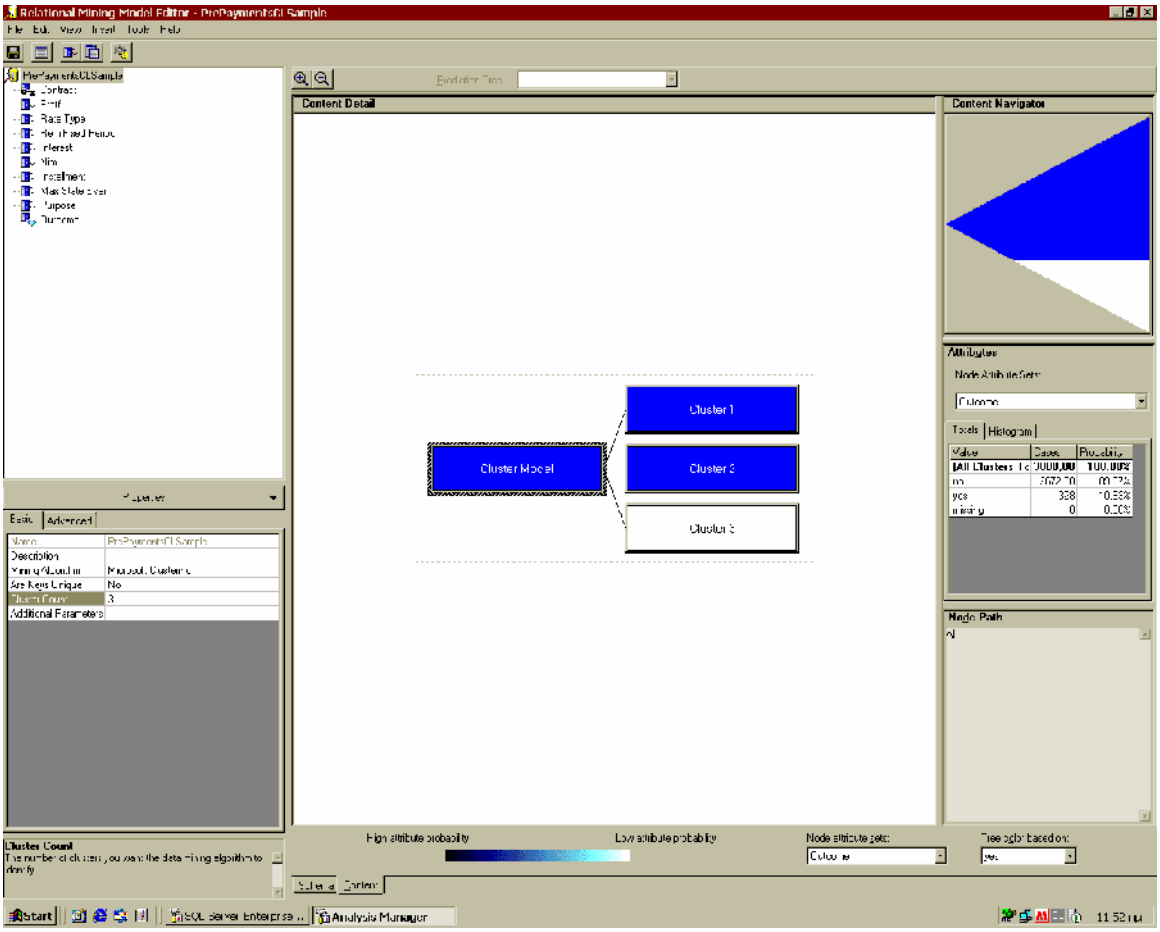
Όσον αφορά στη συσταδοποίηση κατηγορικών δεδομένων, οι Tang και MacLennan (2005) δηλώνουν ότι χρησιμοποιείται ο αλγόριθμος EM, ενώ μια μέθοδος του «Microsoft Clustering» για τη μέτρηση της απόστασης κατηγορικών δεδομένων από μια συστάδα είναι

ότι η απόσταση αυτή ισούται με 1 μείον την πιθανότητα να ανήκει η τιμή ενός συγκεκριμένου αντικειμένου σε αυτή τη συστάδα.

Ζητώντας από τον Server να δημιουργήσει 3 συστάδες για τα δεδομένα μας, πήραμε τα αποτελέσματα του Σχήματος 7.2.

ΣΧΗΜΑ 7.2

Συσταδοποίηση στον Microsoft SQL Server 2005



Με μπλε χρώμα συμβολίζονται οι συστάδες που περιλαμβάνουν άτομα με ποσοστό αποχώρηση κοντά στο 10,93%, (βλ. Πίνακα 7.1, ποσοστό % γραμμής / «κακοί» πελάτες). Επομένως, το ενδιαφέρον μας θα στραφεί στις δύο πρώτες συστάδες, καθώς τα άτομά τους εμφανίζουν μεγαλύτερο ποσοστό αποχώρησης, σε σχέση με αυτά της τρίτης συστάδας.

Δηλαδή, τα άτομα της τρίτης συστάδας αναμένουμε να μη μας αποσχολήσουν με πιθανή αποπληρωμή δανείου και αποχώρησή τους από την τράπεζα. Ας δούμε όμως και συνοπτικά τα αποτελέσματα των τριών συστάδων, στον Πίνακα 7.5.

ΠΙΝΑΚΑΣ 7.5

Χαρακτηριστικά συστάδων και ποσοστά αποχώρησης

	Συστάδα 1		Συστάδα 2		Συστάδα 3	
	Cases	**Prob.**	**Cases**	**Prob.**	**Cases**	**Prob.**
Total	**1747,8**	**100,00%**	**808,09**	**100,00%**	**444,03**	**100,00%**
no	1559,93%	89,25%	682,44	84,45%	429,63	96,76%
yes	187,96%	10,75%	125,65	15,55%	14,39	3,24%
missing	0	0,00%	0	0,00%	0	0,00%
Node Path	Staff=yes, RateType=fixed or floating until end of duration, NIM=0.5-2.5%, Interest=4-6%, REM_FixedPeriod=0-24 months, Installment=700-2500, Purpose= Buy / Buy Land / Construction / Refinance, Interest=Null-4%		Interest=6+ %, NIM=2.5-6%, RateType=floating at this time, Purpose=Repair / Home Equity, Installment=150-300, REM_FixedPeriod=0-24 months, MaxStEver=1, Staff=no, Installment=2500+, Installment=0-150, Installment=300-700		REM_FixedPeriod=25+ months, RateType= fixed at this time, NIM=0-0.5% or Null, Interest=4-6%, Purpose= Buy / Buy Land / Construction / Refinance, Installment=300-700, Staff=no, Installment=700-2500, Interest=Null-4%	

Από τον Πίνακα 7.5, βλέπουμε ότι η δεύτερη συστάδα περιέχει άτομα που παρουσιάζουν το μεγαλύτερο ποσοστό αποχώρησης (15,5%). Τα χαρακτηριστικά αυτών των ατόμων καταγράφονται στο Node Path. Τα άτομα αυτά είχαν επιτόκιο πάνω από 6% και καθαρό περιθώριο κέρδους από 2,5 έως 6%. Επίσης, είχαν κυμαινόμενο επιτόκιο, δηλαδή έχουν περάσει την περίοδο που είχαν σταθερό, ή απέμεναν 0 έως 24 μήνες για να τελιώσει η περίοδος με σταθερό επιτόκιο.

Επιπλεόν, τα άτομα αυτά ζήτησαν το δάνειο για επισκευή ή απόκτηση ιδιοκτησίας ενός ακινήτου, πλήρωναν δόση από 0 έως 700 ευρώ (διάφορες κατηγορίες της μεταβλητής εισήχθηκαν στη συστάδα) ή πάνω από 2500 ευρώ και έχουν καθυστερήσει το πολύ μέχρι ένα

μήνα την πληρωμή της δόσης τους. Τέλος, τα άτομα αυτά δεν αποτελούσαν προσωπικό της τράπεζας.

Η ανάλυση των αποτελεσμάτων έγινε με βάση τη θεωρούμενη ως πιο «επικίνδυνη» ομάδα. Όπως βλέπουμε και από τα χαρακτηριστικά αυτής της ομάδας, είναι λογικό να είναι αυτοί οι πιο επικύνδυνοι προς αποχώρηση πελάτες. Στηριζόμενοι στα αποτελέσματα αυτά, θα μπορούμε να επικεντρωθούμε την επόμενη χρονιά σε άτομα με παρόμοια χαρακτηριστικά, ώστε να κάνουμε κάποια μετατροπή ή ρύθμιση στο δάνειό τους, με στόχο να μην αποχωρήσουν από την τράπεζά μας.

7.3 Αποτελέσματα χρήσης άλλων προγραμμάτων

Αξίζει να αναφέρουμε ότι οι δυνατότητας της ΕΔ και του σχετικού λογισμικού είναι πολύ μεγάλες. Είναι σίγουρο ότι θα μπορούσαμε να έχουμε πολύ πιο ικανοποιητικά αποτελέσματα εάν χρησιμοποιούσαμε πλατφόρμες όπως το Clementine ή το SAS.

Όμως, στα πλαίσια των παροχών και των δυνατοτήτων μας, θεωρούμε ότι τα πιο ικανοποιητικά αποτελέσματα χρήσης είναι αυτά που παρουσιάσαμε παραπάνω, μέσω του SQL Server. Παρ'όλα αυτά, δοκιμάσαμε και άλλα προγράμματα, τα οποία μπορέσαμε να βρούμε ελεύθερα στο διαδίκτυο. Αυτά ήταν το Weka, ο XL-Miner και η γλώσσα R.

Ας δούμε ποια ήταν τα αποτελέσματα που προέκυψαν από τη χρήση κάθε προγράμματος. Αυτό που μας απασχολεί είναι εάν μπορούν να χειριστούν ικανοποιητικά τα κατηγορικά δεδομένα σε γενικές γραμμές, αλλά και αν μπορούν να δώσουν έγκυρα αποτελέσματα συσταδοποίησης. Ακολουθεί μια υποενότητα για κάθε πρόγραμμα που χρησιμοποιήσαμε.

7.3.1 XL-Miner

Ο XL-Miner παρέχει ένα σύνολο σαφών τεχνικών ανάλυσης, οι οποίες στηρίζονται σε στατιστικές μεθόδους και μεθόδους μηχανικής εκμάθησης. Το πρόγραμμα αυτό μπορεί να χειριστεί πολύ μεγάλα σύνολα δεδομένων, τα οποία μπορεί να ξεπερνούν και τη χωρητικότητα του Excel (βλ. http://www.resample.com/xlminer/capabilities.shtml).

Μια συνήθης διαδικασία που προτείνει ο XL-Miner είναι η επιλογή ενός δείγματος από μια μεγαλύτερη βάση δεδομένων, η μεταφορά του στο Excel με στόχο τη δημιουργία ενός κατάλληλου μοντέλου και, στην περίπτωση που θέλουμε να πραγματοποιήσουμε μια

διαδικασία επιβλεπόμενης μάθησης, η επιστροφή στη βάση δεδομένων με στόχο την αξιολόγηση του αποτελέσματος (*output*).

Στην τυπική έκδοση του XL-Miner, αυτή η διαδικασία υποστηρίζεται από βάσεις δεδομένων τύπου Oracle, SQL Server και Access. Όμως, εμείς χρησιμοποιήσαμε ένα δοκιμαστικό πράγραμμα (*trial demo*), στο οποίο είχαμε τη δυνατότητα να εισάγουμε μόλις 200 αντικείμενα. Επιπλέον, αυτό που παρατηρήσαμε είναι ότι ο XL-Miner είναι πιο χρήσιμος σαν πρόγραμμα όταν ενδιαφερόμαστε να εκτελέσουμε μια διαδικασία επιβλεπόμενης εκμάθησης. Το σύνολο των λειτουργιών του XL-Miner συγκεντρώνεται στις παρακάτω ομάδες:

- Διαμέριση (*partitioning*)
- Ταξινόμηση (*classification*)
- Πρόβλεψη (*prediction*)
- Ανάλυση συσχετίσεων (*association analysis*)
- Πρόβλεψη χρονοσειρών (*time series forecasting*)
- Μείωση και εξερεύνηση δεδομένων (*data reduction and exploration*)

Στο κομμάτι της διαμέρισης προτείνεται μια έγκυρη μεθοδολογία, με βάση την οποία προτείνεται ο προσδιορισμός της μεταβλητής απόκρισης (όπως το outcome που ορίσαμε στον Server) και στη συνέχεια η διαμέριση του συνόλου δεδομένων σε μέρη, ώστε να διαμορφώσουμε ένα μοντέλο (*training partition*), να επικυρώσουμε τα αποτελέσματα (*validation partition*) και να ελέγξουμε τα αποτελέσματα του μοντέλου σε ένα μέρος δεδομένων που δεν έχει χρησιμοποιηθεί πριν (*test partition*).

Όσον αφορά τη διαχείριση συνόλων δεδομένων με κατηγορικές μεταβλητές, συμπεράναμε ότι ο XL-Miner δεν μπορεί να προσφέρει ικανοποιητικά αποτελέσματα. Όμως, προσφέρει τη δυνατότητα άμεσου προσδιορισμού βουβών μεταβλητών (*dummy variables*) ή την ανάθεση σκορ στις κατηγορίες των μεταβλητών.

Το κομμάτι της συσταδοποίησης στον XL-Miner περιλαμβάνεται, στην ομάδα της μείωσης και εξερεύνησης δεδομένων, που αναφέρθηκε προηγουμένως. Όμως, οι δυνατότητες που παρέχει είναι περιορισμένες. Στην ουσία, μέσω του XL-Miner μπορούμε να κάνουμε

εφαρμογή του k-means και όχι των παραλλαγών του, ή ιεραρχική συσταδοποίηση, χρησιμοποιώντας την ευκλείδεια απόσταση (παράγραφος 5.1.1).

Στο βιβλίο των Shmueli et al. (2005) γίνεται μια σύντομη αναφορά στη συσταδοποίηση, ενώ αναφέρονται απλά κάποια μέτρα ομοιότητας, χωρίς να αποδεικνύεται ότι υπάρχει δυνατότητα σχετικής επιλογής στον XL-Miner. Επομένως, ο XL-Miner πιθανότατα δε μπορεί να εξυπηρετήσει στη συσταδοποίηση κατηγορικών δεδομένων.

Τέλος, μια από τις δυνατότητες του συγκεκριμένου προγράμματος είναι η ικανοποιητική γραφική απόδοση των δεδομένων, με στόχο την εμπέδωση της φύσης των δεδομένων και των αποτελεσμάτων.

7.3.2 Weka

Το Weka είναι μια ελεύθερη πλατφόρμα, η οποία φαίνεται να έχει αρκετές δυνατότητες. Όμως, συναντήσαμε αρκετές δυσκολίες κατά την απόπειρα ανάλυσης του συνόλου δεδομένων με τους πελάτες της στεγαστικής πίστης.

Η μεγαλύτερη δυσκολία είχε σχέση με τον τύπο του αρχείου δεδομένων, καθώς το Weka μπορεί να ανοίξει κατά βάση αρχεία τύπου arff (βλ. Witten and Frank, 2005). Βέβαια, αποδέχεται και αρχεία τύπου csv (*comma separated*), ανάλογα με τον τρόπο που ορίζεται το «comma». Εμείς προτείνουμε την εισαγωγή των δεδομένων μέσω αρχείου τύπου txt (*text*).

Πέρα από αυτά, το Weka είναι ένα εύκολο στη χρήση πρόγραμμα. Από την επιλογή Applications, χρησιμοποιήσαμε τον Explorer (βλ. Kirkby et al., 2007), ώστε να εισάγουμε και να κάνουμε μία πρώτη μελέτη των δεδομένων. Ύστερα από την εισαγωγή των δεδομένων, ο Explorer δίνει τη δυνατότητα εργασίας μέσα από τις παρακάτω καρτέλες:

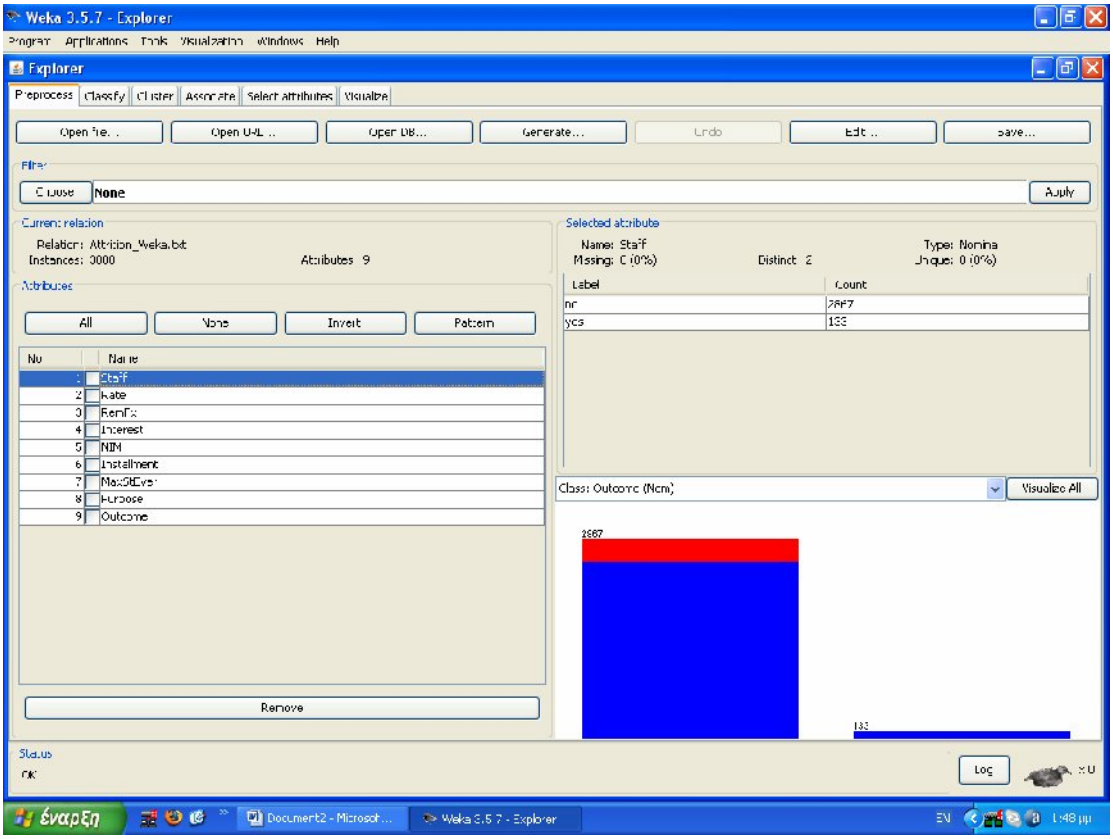
- Προεπεξεργασία (*preprocess*)
- Ταξινόμηση (*classify*)
- Συσταδοποίηση (*cluster*)
- Ανάλυση συσχετίσεων (*associate*)
- Επιλογή γνωρισμάτων (*select attributes*)
- Νοερή απεικόνιση (*visualize*)

Οι καρτέλες της προεπεξεργασίας και απεικόνισης δίνουν ενδιαφέροντα αποτελέσματα. Για παράδειγμα, από την καρτέλα «preprocess» μπορούμε να δούμε τη γραφική απόδοση των

τιμών κάθε μεταβλητής, ως προς τη μεταβλητή Outcome. Στο Σχήμα 7.3, παραθέτουμε ενδεικτικά την απεικόνιση της μεταβλητής Staff μέσω ιστογραμμάτων.

ΣΧΗΜΑ 7.3

Προεπεξεργασία δεδομένων στο Weka



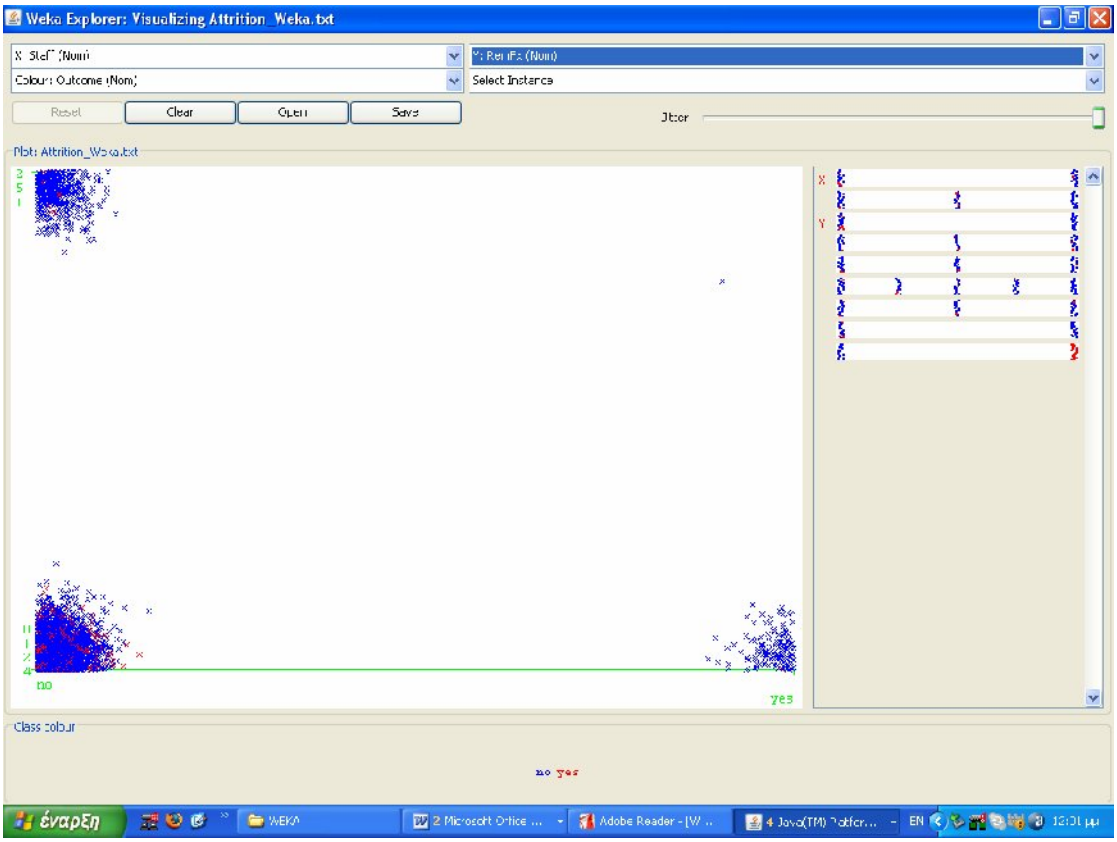
Παρατηρούμε ότι, από τα 3.000 άτομα του δείγματος, οι 2.867 (αριστερό ιστόγραμμα) δεν είναι υπάλληλοι της τράπεζας, ενώ οι 133 (δεξί ιστόργαμμα) είναι. Επίσης, με κόκκινο χρώμα συμβολίζονται οι «κακοί» πελάτες, δηλαδή αυτοί που είχαν Outcome = yes, άρα αποπλήρωσαν το δάνειό τους. Όπως βλέπουμε από το Σχήμα 7.3, αυτοί που αποπλήρωσαν το δάνειό τους και απομάκρυναν το χαρτοφυλάκιό τους από την τράπεζα ήταν όλοι άτομα που δεν αποτελούσαν προσωπικό της τράπεζας.

Ανάλογα χρήσιμα περιγραφικά στοιχεία μπορούν να προκύψουν και για τις υπόλοιπες υπό μελέτη μεταβλητές. Μια άλλη χρήσιμη καρτέλα είναι η «visualize», μέσα από την οποία μπορούμε να έχουμε μια νοερή απεικόνιση ανά ζεύγη μεταβλητών. Ενδεικτικά, παραθέτουμε

ένα διάγραμμα (βλ. Σχήμα 7.4) που απεικονίζει τη σχέση μεταξύ των μεταβλητών Staff και REM_FixedPeriod (βλ. Πίνακα 7.3).

ΣΧΗΜΑ 7.4

Νοερή απεικόνιση δεδομένων στο Weka



Στον οριζόντιο άξονα βλέπουμε τη μεταβλητή Staff που έχει τις κατηγορίες no (κάτω αριστερά) και yes (κάτω δεξιά), ενώ στον κάθετο άξονα έχουμε τη μεταβλητή REM_FixedPeriod, με τις κατηγορίες 0-24 (κάτω αριστερά) και 25+ (πάνω αριστερά). Παρατηρώντας τα σχηματιζόμενα νέφη σημειών (Σχήμα 7.4), έχουμε να πούμε ότι οι περισσότεροι πελάτες με κόκκινο χρώμα, άρα οι «επικίνδυνοι» πελάτες, βρίσκονται κάτω αριστερά.

Αυτό σημαίνει ότι πρέπει να στρέψουμε το ενδιαφέρον μας, όσον αφορά αυτές τις δύο μεταβλητές, σε πελάτες που δεν αποτελούν προσωπικό της τράπεζας και αναμένεται να λήξει σύντομα η περίοδος σταθερού επιτοκίου, πιο συγκεκριμένα σε 0-24 μήνες. Το συγκεκριμένο

συμπέρασμα φαίνεται αρκετά λογικό, αν αναλογιστούμε ότι ένας πελάτης που έχει περιορισμένο υπόλοιπο μηνών με σταθερό επιτόκιο, θα ψάξει λύση για το επόμενο στάδιο του τρόπου πληρωμής του στεγαστικού του δανείου, ενώ είναι πιθανό να επιθυμήσει να μεταφέρει το δάνειό του σε άλλη τράπεζα.

Όσον αφορά την καρτέλα «cluster», οφείλουμε να αναφέρουμε ότι επιχειρήσαμε να την αξιοποιήσουμε, χωρίς όμως να βγουν ενδιαφέροντα αποτελέσματα. Οι αλγόριθμοι συσταδοποίησης κατηγορικών δεδομένων που είχαμε τη δυνατότητα να εκτελέσουμε μέσω του Weka ήταν οι COBWEB και EM (βλ. παράγραφο 6.1.4).

Η εκτίμησή μας είναι ότι το Weka δεν δίνει τη δυνατότητα διενέργειας συσταδοποίησης κατηγορικών δεδομένων, ενώ πιθανότατα αντιμετώπισε πρόβλημα με το μέγεθος του αρχείου στο οποίο ζητήσαμε να πραγματοποιήσει μια διαδικασία συσταδοποίησης. Όμως, σύμφωνα με τους Maimon και Rokah (2005), το Weka είναι ένα πρόγραμμα που παρουσιάζει έντονη εξελιξιμότητα.

7.3.3 R

Η R είναι μια ελεύθερη στατιστική γλώσσα προγραμματισμού, η οποία γνωρίζει μεγάλη εξέλιξη τα τελευταία χρόνια. Είναι ελεύθερα διαθέσιμη, μέσω της ιστοσελίδας http://www.r-project.org/ και στηρίζεται στην ανάπτυξη προγραμμάτων μέσω πακέτων (*packages*), τα οποία διατίθενται επίσης ελεύθερα από χρήστες όλου του κόσμου (βλ. Φωκιανός και Χαραλάμπους, 2008).

Ο λόγος που επιχειρήσαμε την εφαρμογή μεθόδων ΕΔ στην R είναι γιατί σε αυτή τη γλώσσα μπορούμε να τρέξουμε έναν αλγόριθμο, δεν θεωρούμε όμως ότι εφαρμόζουμε όλη τη διαδικασία της ΕΔ. Για παράδειγμα, στα πλαίσια αναζήτησής μας στο διαδίκτυο, εντοπίσαμε το πακέτο **cba** (*Clustering for Business Analytics*), όπου περιλαμβάνεται ο αλγόριθμος ROCK (βλ. παράγραφο 6.1.2).

Η απόπειρά μας να εκτελέσουμε τον αλγόριθμο ROCK από το πακέτο cba δεν έβγαλε άμεσα συμπεράσματα. Ενδεικτικά αναφέρουμε ότι, σύμφωνα με τα αποτελέσματα κατά την εκτέλεση του αλγορίθμου, ο ROCK συγχώνευσε τα δεδομένα σε 531 συστάδες, εκ των οποίων οι περισσότερες περιελάμβαναν πολύ μικρό αριθμό αντικειμένων. Επίσης, σε επόμενο βήμα απομακρύνθηκαν 333 συστάδες.

Σχετικά με το Output που προέκυψε, θεωρούμε ότι δεν ήταν ιδιαίτερα άμεσο ώστε να εξυπηρετήσει στην αξιοποίηση των πιο σημαντικών εκ των αποτελεσμάτων. Κλείνοντας,

προτείνουμε και άλλα πακέτα που υπάρχουν στην R και διαθέτουν τεχνικές ΕΔ για την επεξεργασία δεδομένων. Για παράδειγμα, το πακέτο **ElemStatLearn** δίνει τη δυνατότητα εκτέλεσης των παραδειγμάτων του βιβλίου των Hastie et al. (2001), ενώ το πακέτο **FactorMineR** εξυπηρετεί στην περίπτωση που θέλουμε να παραγματοποιήσουμε διερευνητική ανάλυση δεδομένων (*Exploratory Data Analysis* – βλ. παράγραφο 2.2.1).

Τέλος, μέσα από το πακέτο **rattle** μπορούμε να προβούμε σε μοντελοποίηση πρόβλεψης (βλ. παράγραφο 2.2.2), ενώ στο **dprep** δίνονται δυνατότητες προεπεξεργασίας δεδομένων (βλ. κεφάλαιο 4). Για εκτενέστερη ανάλυση των μεθοδολογιών της R και προτάσεις χρήσης της συγκεκριμένης γλώσσας, παραπέμπουμε στο βιβλίο του Crawley (2007).

7.4 Άλλα προγράμματα και πλατφόρμες

Στην ενότητα αυτή, θα αναφέρουμε ενδεικτικά άλλα προγράμματα και τεχνικές που προτείνονται για την αξιοποίηση εφαρμογών ΕΔ σε πολύ μεγάλα σύνολα δεδομένων. Για παράδειγμα, αναφέρουμε το πρόγραμμα **Statistica** (http://www.statsoft.com/) το οποίο ενδείκνυται για την αποτελεσματική μετατροπή των δεδομένων σε χρήσιμη πληροφορία, ενώ σχετίζεται με την ΕΔ, την επιχειρηματική ευφυΐα και τον ποιοτικό έλεγχο.

Επίσης, η χρήση της **Oracle** (http://www.oracle.com/index.html) μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη. Μάλιστα, υπάρχει δυνατότητα δωρεάν χρήσης μιας βάσης δεδομένων της Oracle, η οποία καλείται **Oracle10g** και περιλαμβάνει εφαρμογές ΕΔ. Η χρήση της Oracle10g μπορεί να διευκολυνθεί στην περίπτωση που κάποιος γνωρίζει προγραμματισμό SQL, καθώς ο συνδυασμός της Oracle10g με την **PostgreSQL**, η οποία είναι επίσης δωρεάν διαθέσιμη μέσω του διαδικτύου (http://www.postgresql.org/), μπορεί να εξάγει σημαντικά αποτελέσματα.

Ακόμη, αναφέρουμε τη δράση των Andritsos και Tzerpos, οι οποίοι επιχείρησαν την εφαρμογή του αλγορίθμου LIMBO (βλ. παράγραφο 6.1.4 και Andritsos, 2003) στα λειτουργικά συστήματα **TOBEY**, **Linux**, και **Mozilla** (βλ. Andritsos and Tzerpos, 2005).

Τέλος, δύο ενδιαφέρουσες ιστοσελίδες με χρήσιμες πληροφορίες και ενημέρωση για το διαθέσιμο λογισμικό ΕΔ είναι οι http://www.salford-systems.com/landing.php και http://www.kdnuggets.com/news/2007/index.html

7.5 Γενικά συμπεράσματα

Είναι γεγονός ότι αντιμετωπίσαμε αρκετές δυσκολίες κατά την εφαρμογή των μεθόδων και αλγορίθμων ΕΔ στον υπολογιστή και τα σχετικά προγράμματα. Όμως, σε καμμία περίπτωση δεν θα αμφέβαλλε κανείς ότι το πρακτικό κομμάτι κατά την εκτέλεση μια σεφαρμογής ΕΔ είναι και το πιο ουσιαστικό.

Για την επιτυχημένη εκτέλεση ενός αλγορίθμου ή τεχνικής απαιτείται και η ύπαρξη καλά θεμελιωμένου θεωρητικού υπόβαθρου. Για το λόγο αυτό, στα προηγούμενα κεφάλαια, επικεντρωθήκαμε στη θεωρητική προσέγγιση της ΕΔ, με στόχο την εμπέδωση όλων των σχετικών διαδικασιών.

Το ενδιαφέρον μας στράφηκε στην παρουσίαση των μεθοδολογιών των ΕΔ και KDD, κάνοντας και διάκριση των σχετικών μεθόδων, ανάλογα με το αποτέλεσμα που θέλουμε να έχουμε από τη μελέτη ενός συνόλου δεδομένων. Επιπλέον, αναφερθήκαμε στη διαδικασία προεπεξεργασίας δεδομένων, ένα πολύ σημαντικό μέρος της αλυσίδας KDD.

Όσον αφορά τα κεφάλαια που ακολούθησαν, στόχος μας ήταν η παρουσίαση της πιο χαρακτηριστικής μεθόδου μη επιβλέπουσας εκμάθησης, που ανήκει στην περιγραφική μοντελοποίηση. Αυτή η μέθοδος ήταν η συσταδοποίηση, αλγορίθμους της οποίας συλλέξαμε, με κριτήριο τη δυνατότητά τους να συσταδοποιήσουν κατηγορικά ή μικτά δεδομένα.

Στο μέρος της πρακτικής εφαρμογής, στο παρόν κεφάλαιο, θεωρούμε ότι τα πιο χρήσιμα αποτελέσματα προέκυψαν από τον SQL Server 2005, αλλά παραθέσαμε και τις απόπειρες που κάναμε σε άλλα προγράμματα. Το συμπέρασμά μας δε θα μπορούσε να είναι άλλο από το γεγονός ότι η ΕΔ είναι σε θέση να προσφέρει πολλά και χρήσιμα αποτελέσματα σε ετερόκλητους κλάδους, αρκεί να υπάρχει καλή γνώση της φιλοσοφίας και των στόχων της, καθώς και αξιόλογο τεχνολογικό υλικό.

ΠΑΡΑΡΤΗΜΑ Α

Αλγόριθμος k-means

Α.1 Περιγραφή αλγορίθμου

Η μέθοδος k-means αποτελεί μια από τις πιο συχνά χρησιμοποιούμενες μεθόδους συσταδοποίησης (MacQueen, 1967 – Anderberg, 1973). Πρόκειται για έναν αλγόριθμο που ανήκει στην κατηγορία της **διαιρετικής ή μη ιεραρχικής συσταδοποίησης** (Jain and Dubes, 1988).

Ο k-means είναι ένας αλγόριθμος που είναι κατάλληλος για αριθμητικά (συνεχή) δεδομένα. Στόχος του είναι η άμεση αποσύνθεση του συνόλου των δεδομένων σε ένα σύνολο ασυσχέτιστων συστάδων.

Η συνάρτηση που επιχειρεί να ελαχιστοποιήσει ο k-means είναι η μέση τετραγωνική απόσταση των δεδομένων από τα πλησιέστερα κέντρα των συστάδων και εκφράζεται από την παρακάτω εξίσωση:

$$E = \sum_{i=1}^{k} \sum_{x \in Ci} d(x, m_i)$$

όπου m_i το κέντρο της συστάδας C_i και $d(x, m_i)$ η ευκλείδεια απόσταση (βλ. παράγραφο 5.1.1) μεταξύ ενός στοιχείου x και του κέντρου m_i.

Στην ουσία, ο k-means επιχειρεί μέσω της συνάρτησης E να ελαχιστοποιήσει την απόσταση κάθε σημείου από το κέντρο της συστάδας όπου ανήκει το σημείο. Ύστερα, αναθέτει κάθε στοιχείο του συνόλου δεδομένων στη συστάδα της οποίας το κέντρο είναι πιο κοντά και ξανά-υπολογίζει τα κέντρα. Η διαδικασία συνεχίζεται μέχρι τη στιγμή που θα σταματήσουν να αλλάζουν τα κέντρα των συστάδων.

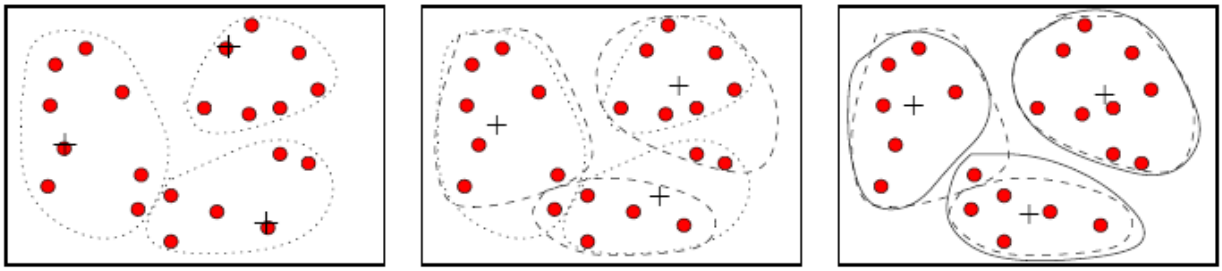
Για να ελαχιστοποιηθεί η E, θεωρούνται από τον αλγόριθμο k σημεία ως τα κέντρα k συστάδων. Στην περίπτωση που η σειρά των δεδομένων δεν έχει σημασία, παίρνουμε τις πρώτες k εγγραφές. Διαφορετικά, επιλέγουμε αντιπροσωπευτικά σημεία για τις θεωρούμενες συστάδες.

Στη συνέχεια, κάθε σημείο αντισοιχίζεται στη συστάδα της οποίας το κέντρο βρίσκεται πιο κοντά και υπολογίζονται τα νέα κέντρα των συστάδων με χρήση του μέσου όρου των σημείων τους. Το Σχήμα Α.1 εξηγεί αυτή τη διαδικασία.

ΣΧΗΜΑ Α.1

Αλγόριθμος k-means

[Πηγή: Han and Kamber, 2001]



Η διαδικασία αυτή επαναλαμβάνεται μέχρι τη στιγμή όπου τα όρια των συστάδων σταματούν να μεταβάλλονται, ή όταν η συνάρτηση Ε δε μεταβάλλεται σημαντικά. Ο αριθμός συστάδων που χρησιμοποιείται θεωρείται σταθερός και δεδομένος εξαρχής.

Επομένως, η τελική λύση που προτείνεται από τον k-means έχει σχέση με τον ορισμό των αρχικών κέντρων, καθώς και με τον τρόπο που είναι διατεταγμένα τα αντικείμενα στο σύνολο δεδομένων. Ένας περιορισμός για το k είναι ότι πρέπει $k \leq n$, όπου n το πλήθος των αντικειμένων του συνόλου δεδομένων.

Α.2 Βασικά βήματα

Ο αλγόριθμος k-means θεωρείται ιδιαίτερα γρήγορος και συνήθως τερματίζεται ύστερα από λίγες επαναλήψεις. Αυτό τον κάνει ιδιαίτερα χρήσιμο στις περιπτώσεις όπου απαιτείται ομαδοποίηση σε μεγάλα σύνολα δεδομένων (βλ. Huang, 1998).

Επίσης, δεν χρειάζεται να κρατά στη μνήμη πολλά στοιχεία. Έτσι, δεν απαιτεί τεράστιες χωρητικότητες ούτε μεγάλη υπολογιστική ισχύ (Κούτρας, 2007) για να λειτουργήσει, πράγμα που αποτελεί ένα ακόμη πλεονέκτημά του. Βέβαια, στην ενότητα 6.3 αναφέραμε και μειονεκτήματα των διαιρετικών αλγορίθμων, άρα και του k-means.

Τα βασικά βήματα του αλγορίθμου σε μορφή ψευδοκώδικα, έτσι όπως καταγράφονται από τους Βαζιργιάννη και Χαλκίδη (2005), είναι τα εξής:

1. Εύρεση των αρχικών κέντρων m_i με $i = 1, 2, ..., k$ για τις k συστάδες.

   Για κάθε επανάληψη h = 1, …, h_{max} :

2. Υπολογισμός της απόστασης κάθε στοιχείου του συνόλου δεδομένων από το κέντρο κάθε συστάδας:

$$d_{ri} = (x_r - m_i^{(h)})^2$$

$$\text{με } r = 1, 2, ..., n \text{ , } i=1,2,...,k$$

3. Κάθε στοιχείο x_r αντιστοιχίζεται στη συστάδα για την οποία ισχύει:

$$\min_{r,i}(d_{ri}) \text{ για κάθε r, i}$$

4. Υπολογισμός των νέων κέντρων των συστάδων:

$$m_i^{(h+1)} = \frac{\sum_{i=1}^{n_i} x_r}{n_i}$$

5. If $|| m_i^{(h)} - m_i^{(h+1)} || < \varepsilon$ then

   stop

   else

   h = h + 1, goto2

ΠΑΡΑΡΤΗΜΑ Β

Πολυπλοκότητα αλγορίθμων

Β.1 Αλγόριθμοι συσταδοποίησης κατηγορικών και μικτών δεδομένων

Στο παράρτημα αυτό, παρατίθεται ο Πίνακας Β.1 όπου καταγράφεται η πολυπλοκότητα κάθε αλγορίθμου συσταδοποίησης, καθώς δεν αφορά μέρος της εργασίας μας. Πρόκειται για τους πιο βασικούς από τους αλγορίθμους συσταδοποίησης κατηγορικών και μικτών δεδομένων, που αναφέραμε στο έκτο κεφάλαιο.

ΠΙΝΑΚΑΣ Β.1

Πολυπλοκότητα αλγορίθμων

Αλγόριθμος	**Πολυπλοκότητα**
k-modes	$O(n)$
ROCK	$O(n^2 + n \cdot m_m \cdot m_\alpha + n^2 \cdot \log n)$
STIRR	$O(n)$
CACTUS	$O(n)$
COOLCAT	$O(n)$
k-prototypes	$O(n)$

Να αναφέρουμε ότι ως n εννοούμε το πλήθος των αντικειμένων, ενώ με k συμβολίζεται ο αριθμός των συστάδων. Επίσης, στον αλγόριθμο ROCK, με m_m συμβολίζεται ο μέγιστος αριθμός γειτόνων για ένα αντικείμενο και m_α είναι ο μέσος αριθμός γειτόνων για ένα αντικείμενο.

Παρατηρούμε ότι (βλ. Πίνακα Β.1), όλοι οι προαναφερθέντες αλγόριθμοι έχουν όμοια πολυπλοκότητα, εκτός του ROCK, ο οποίος έχει μεγαλύτερη.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Βαζιργιάννης, Μ. και Χαλκίδη, Μ. (2005). *Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό*, 2η εκδ., τυπωθήτω – Γιώργος Δαρδάνος, Αθήνα.

Κατέρη, Μ. (2006). Εφαρμοσμένη Ανάλυση Δεδομένων, (Σημειώσεις), Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς.

Κούτρας, Μ. (2007). Εφαρμοσμένη Πολυμεταβλητή Ανάλυση: Ανάλυση κατά συστάδες, (Σημειώσεις), Π.Μ.Σ. στην Εφαρμοσμένη Στατιστική, Πανεπιστήμιο Πειραιώς.

Φωκιανός, Κ. και Χαραλάμπους, Χ. (2008). Εισαγωγή στην R, (Σημειώσεις), Τμήμα Μαθηματικών και Στατιστικής, Πανεπιστήμιο Κύπρου.

Ξένη

Abonyi, J. and Feil, B. (2007). *Cluster Analysis for Data Mining and System Identification*, Birkhäuser Verlag AG.

Aczel, A.D. (1989). *Complete Business Statistics*, Irwin Series in Quantitative Analysis for Business, Irwin.

Agrawal, R., Gehrke, J., Gunopoulos, D. and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications, In *Proceedings of 1998 ACM-SIGMOD International Conference on Management of Data*, **27** (2), 94-105.

Agresti, A. (2002). *Categorical Data Analysis,* 2nd ed., Wiley.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley.

Akaike, H. (1974). A new look at statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716-723.

Alpaydin, E. (2004). *Introduction to Machine Learning*, MIT Press.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*, Academic Press.

Andersen, E.B. (2001). *Introduction to the Statistical Analysis of Categorical Data*, Springer.

Andritsos, P. (2002). Data Clustering Techniques (Qualifying Oral Examination Paper), *Tech. Report CSRG-443, U. of Toronto, Dep. Of Computer Science*.

Andritsos, P., Tsaparas, P., Miller, R.J. and Sevcik, K. C. (2003). LIMBO: A Scalable Algorithm to Cluster Categorical Data, *Tech. Report CSRG-467, U. of Toronto, Dep. Of Computer Science*.

Andritsos, P. (2004). Scalable Clustering of Categorical Data and Applications, PhD Thesis, University of Toronto, Department of Computer Science.

Andritsos, P. and Tzerpos, V. (2005). Information-Theoretic Software Clustering, *IEEE Transactions on Software Engineering*, **31** (2).

Barbará, D., Couto, J. and Li, Y. (2002). COOLCAT: An entropy-based Algorithm for Categorical Clustering, In *Proceedings of the 11th International Conference on Information and Knowledge Management*, 582-589.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*, John Wiley and sons.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivarate Analysis: Theory nad Practise*, MIT Press.

Bramer, M. (2007). *Principles of Data Mining: Undergraduate Topics in Computer Science*, Springer.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*, 2nd ed., Springer-Verlag.

Casella, G. and Berger, R.L. (2002). *Statistical Inference*, 2nd ed., Duxbury Advanced Series.

Cheeseman, P. and Stutz, J. (1996). Bayesian classification (AutoClass): Theory and results, *Advances in Knowledge Discovery and Data Mining*, 153-180, AAAI/MIT Press.

Chen, D., Cui, D.W., Wang, C.X. and Wang, Z.R. (2006). A Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data, *International Journal of Information Technology*, **12** (3), 149-159.

Chen, M., Han, J. and Yu, P.S. (1996). Data Mining: An Overview from Database Perspective, *IEEE Transactions on Knowledge and Data Engineering,* **8**, 866-883.

Chiu, T., Fang, D., Chen, J. and Wang, Y. (2001). A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment, In *Proceedings of 2001 International Conference on Knowledge Discovery and Data Mining*, 263-268.

Cox, R.D. (2006). *Principles of Statistical Inference*, Cambridge University Press.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*, Chapman and Hall.

Crawley, M.J. (2007). *The R Book*, Wiley.

Davidson, R. and MacKinnon, J. (2004). *Econometric Theory and Methods*, Oxford University Press.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, **39** (1), 1-38.

Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*, Prentice Hall.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall.

Ester, M., Kriegel, H.P., Sander, J. And Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases, In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Minng (KDD'96)*, 226-231.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996-a). *Advances in Knowledge Discovery and Data Mining*, AAAI Press.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996-b). From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, 37-54.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996-c). The KDD Process for extracting useful knowledge from volumes of data, *Journal of the ACM*, **39** (11), 27-34.

Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering, *Machine Learning*, **2**, 139-172.

Gamerman, D. and Lopez, H.F. (2006). *Markov Chain Monte Carlo: Stichastic Simulation for Bayesian Inference*, 2nd ed., Chapman and Hall / CRC.

Ganti, V., Gehrke, J. and Ramakrishnan, R. (1999). CACTUS: Clustering Categorical Data Using Summaries, In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA*, 73-83, ACM Press.

Gentle, J.E. (2002). *Elements of Computational Statistics*, Springer-Verlag.

Gibson, D., Kleinberg, J. and Raghavan, P. (1998). Clustering Categorical Data: An Approach Based on Dynamical Systems, In *Proceedings of the 24th International Conference on Very Lerge Data Bases*, 311-322, Morgan Kaufmann.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in practice*, Chapman & Hall.

Glymour, C., Madigan, D., Pregibon, D. and Smyth, P. (1997). Statistical Themes and Lessons for Data Mining, *Data Mining and Knowledge Discovery*, **1**, 11-28.

Grabmeier, J. & Rudolph, A. (2002). Techniques of Cluster Algorithms in Data Mining, *Data Mining and Knowledge Discovery*, **6**, 303-360.

Grünwald, P.D. and Rissanen, J. (2007). *The Minimum Descreption Length Principle*, MIT Press.

Guha, S., Rastogi, R. and Shim K. (1998). CURE: An efficient clustering algorithm for large databases, In *Proceedings of 1998 ACM-SIGMOD International Conference on Management of Data*, 73-84.

Guha, S., Rastogi, R. and Shim K. (2000). ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems*, **25**, 345-366.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, **17:2/3**, 107-145.

Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques*, Morgan Kaufman Publishers.

Hand, D.J. (1981). *Discrimination and Classification*, Wiley.

Hand, D., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining*, MIT Press, Cambridge.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.

He, Z., Xu, X., Deng, S. and Dong, B. (2002). Squeezer: An Efficient Algorithm for Clustering Categorical Data, *Journal of Computer Science and Technology*, **17**, (5), 611-624.

He, Z., Xu, X. and Deng, S. (2004). A link clustering based approach for clustering categorical data, In *Proceedings of the WAIM conference*, available at: http://xxx.sf.nchc.org.tw/ftp/cs/papers/0412/0412019.pdf

He, Z., Xu, X., Deng, S. and Dong, B. (2005-a). K-Histograms: An Efficient Clustering Algorithm for Categorical Dataset, available: http://arxiv.org/ftp/cs/papers/0509/0509033.pdf

He, Z., Xu, X. and Deng, S. (2005-b). Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach, available at: http://arxiv.org/ftp/cs/papers/0509/0509011.pdf

Hinneburg, A. and Keim, D.A. (1998). An efficient approach to clustering in large multimedia databases with noise, In *Proceedings of 1998 International Conference on Knowledge Discovery and Data Mining (KDD'98)*, 58-65.

Huang, Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values, In *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 21-34, Springer.

Huang, Z. (1998). Extentions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery*, **2** (3), 283-304.

Huber, P.J. (1981). *Robust Statistics*, Wiley.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data*, Prentice Hall.

Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs – Statistics for Engineering and Inormation Science*, Springer.

Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*, Prentice Hall.

Karypis, G., Han, E.H. and Kumar, V. (1999). CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, *Computer*, **32**, 68-75.

Kateri, M. (2008). Categorical Data, *Encyclopedia of Statistical Sciences*, Wiley.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley and Sons.

Kim, D., Lee, K. and Lee, D. (2004). Fuzzy clustering of categorical data using fuzzy centroids, *Pattern Recognition Letters*, **25** (11), 1263-1271.

Kirkby, R., Frank, E. and Reutemann, P. (2007). WEKA Explorer User Guide for Version 3-5-5, University of Waikato.

Koch, K.R. (1999). *Parameter Estimation and Hypothesis Testing in Linear Models*, Springer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, **43**, 59-69.

Kudová, P., Řezanková, H., Hůsek, D. and Snášel, V. (2006). Categorical Data Clustering Using Statistical Methods and Neural Networks. In *Proceedings of the Spring Young Researcher's Cooloquium on Database and information Systems, Moscow, Russia*.

Li, C. and Biswas, G. (2002). Unsupervized Learning with Mixed Numeric and Nominal Data, *IEEE Transactions on Knowledge and Data Engineering*, **14** (4).

Lutz, R. W. and Bühlmann, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression, *Statistica Sinica*, **16**, 471-494.

MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations, In *Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability*, Vol.I: Statistics, 281-297.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window, *Journal of the American Statistical Association*, **89**, 1335-1346.

Maimon, O. and Rokah, L. (2005). *Data Mining and Knowledge Discovery Handbook*, Springer.

Marques de Sá, J. P. (2001). *Pattern Recognition: Concepts, Methods and Applications*, Springer.

Miller, R.G. (1981). *Simultaneous statistical inference*, 2nd ed., Springer-Verlag.

Ng, R. and Han, J. (1994). Efficient and effective clustering method for spatial data mining, In *Proceedings of 1994 International Conference on Very Large Data Bases*, 144-155.

Parmar, D., Wu, T. and Blackhurst, J. (2007). MMR: An Algorithm for Clustering Categorical Data Using Rough Set Theory, *Data Knowledge Engineering*, **63** (3), 879-893.

Peters, M. and Zaki, M. (2004). CLICK: Clustering Categorical Data Using K-partite Maximal Cliques, available at: http://www.cs.rpi.edu/research/pdf/04-11.pdf

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop, *AI Magazine*, **11** (5), 68-70.

Ross, S. (1997). *Simulation*, 2nd ed., Academic Press.

San, O.M., Huyhn, V.N. and Nakamori, Y. (2004). An alternative extension of the k-means algorithm for clustering categorical data, *International Journal of Applied Mathematics and Computer Science*, **14** (2), 241-247.

Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, Wiley.

Simonoff, J.S. (2003). *Analyzing Categorical Data*, Springer.

Scheffé, H. (1959). *The Analysis of Variance*, Wiley.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **63**, 461-464.

Sheikholeslami, G., Chatterjee, S. and Zhang, A. (1998). WaveCluster: A multi-resolution clustering approach for very large spatial databases, In *Proceedings of 1998 International Conference on Very Large Data Bases*, 428-439.

Shmueli, G., Patel, N.R. and Bruce, P.C. (2005). *Data Mining in Excel: Lecture Notes and Cases*, Resampling Stats Inc (www.xlminer.com).

Strehl, A. and Ghosh, J. (2002). Cluster Ensembles – A Knowledge Reuse Framework for Combining Partitions, In *Proceedings of the 8th National Conference on Artificial Intelligence and 4th Conference on Innovative Applications of Artificial Intelligence*, 93-99.

Tan, Steinbach and Kumar (2005). *Introduction to Data Mining*, Addison Wesley.

Tang, Z. and MacLennan, J. (2005). *Data Mining with SQL Server 2005*, Wiley.

Theodoridis, S. and Koutroumbas (2003). *Pattern Recognition*, 2nd ed., Elsevier.

Tishby, N., Pereira, F. C. and Bialek W. (1999). The Information Bottleneck Method, In *37th Annual Allerton Conference on Communication, Control and Computing, Urban-Champaign, IL*.

Wang, W., Yang, J. and Muntz, R. (1997). STING: A statistical information grid approach to spatial data mining, In *Proceedings of 1997 International Conference on Very Large Data Bases*, 186-195.

Weiss, S.I. and Kulikowski, C. (1991). *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems*, Morgan Kaufmann.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Elsevier.

Young, G.A. and Smith, R.L. (2005). *Essentials of Statistical Inference*, Cambridge University Press.

Zhang, T., Ramakrishnan, R. and Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases, In *Proceedings of 1996 ACM-SIGMOD International Conference on Management of Data*, 103-114.