



ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΒΙΒΛΙΟΘΗΚΗ
ειδ.
Αρ 59593
ταξ. 005.741

χελ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Διαχείριση ασάφειας στην διαδικασία του clustering σε περιβάλλον εξόρυξης γνώσης από βάσεις δεδομένων»

Χαλκίδη Μαρία

M3970014

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΚΑΤΑΛΟΓΟΣ



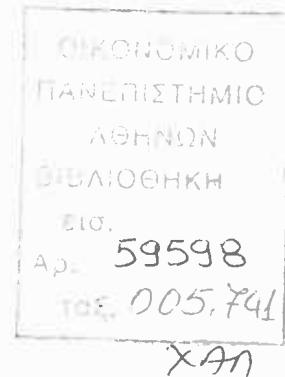
0 000000 358460

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 1999



**ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

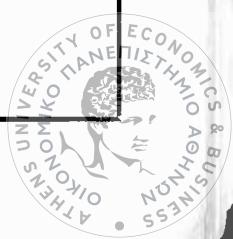


«Διαχείριση ασάφειας στην διαδικασία του clustering σε περιβάλλον εξόρυξης γνώσης από βάσεις δεδομένων»

Χαλκίδη Μαρία

M3970014

Επιβλέπων Καθηγητής: Μιχαήλ Βαζιργιάννης



ΠΕΡΙΕΧΟΜΕΝΑ

EXECUTIVE SUMMARY I

ΠΡΟΛΟΓΟΣ 1

1^ο ΚΕΦΑΛΑΙΟ

KNOWLEDGE DISCOVERY AND DATA MINING ΣΤΙΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ... 4

1.1 ΕΙΣΑΓΩΓΗ	4
1.2 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ	5
1.3 Η ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΚΑΛΥΨΗΣ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ (KDD PROCESS)	6
1.4 DATA MINING	8
1.4.1 ΑΙΓΑΙΤΗΣΕΙΣ ΤΟΥ DATA MINING	9
1.4.2 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΤΩΝ ΤΕΧΝΙΚΩΝ DATA MINING	11
1.5 ΒΑΣΙΚΕΣ ΕΡΓΑΣΙΕΣ DATA MINING	12
1.5.1 CLASSIFICATION	12
1.5.2 CLUSTERING	13
1.5.3 ΕΞΑΓΩΓΗ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ (ASSOCIATION RULES EXTRACTION)	13
1.5.4 ESTIMATION & PREDICTION	14
1.5.5 REGRESSION	14
1.5.6 SUMMARIZATION	15
1.5.7 ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΑΛΛΑΓΗΣ ΚΑΙ ΑΠΟΚΛΙΣΗΣ (CHANGE AND DEVIATION DETECTION)	15
1.6 DATA MINING ΜΕΘΟΔΟΙ	15
1.6.1 CLUSTER ANALYSIS	15
1.6.2 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ	16
1.6.3 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (NEURAL NETWORKS)	17
1.6.3.1 ΚΟΜΒΟΙ ΝΕΥΡΩΝΙΚΟΥ ΔΙΚΤΥΟΥ	18

2^ο ΚΕΦΑΛΑΙΟ

ΜΕΘΟΔΟΙ CLUSTERING 20

2.1 ΕΙΣΑΓΩΓΗ	20
2.2 ΚΑΤΗΓΟΡΙΕΣ CLUSTERING	20
2.2.1 ΙΕΡΑΡΧΙΚΟ CLUSTERING	21
2.2.2 PARTITIONAL CLUSTERING	22
2.2.3 STATISTICAL CLUSTERING	22
2.2.4 CONCEPTUAL CLUSTERING	22
2.2.5 FUZZY CLUSTERING	23
2.2.6 KOHONEN NET CLUSTERING	23
2.3 K-MEANS	26
2.3.1 ΑΛΓΟΡΙΘΜΟΣ K-MEANS	27
2.3.2 ΠΑΡΑΛΛΑΓΕΣ K-MEANS	28
2.3.3 ΕΚΛΕΙΤΥΝΣΗ ΑΡΧΙΚΩΝ ΣΗΜΕΙΩΝ ΓΙΑ ΤΟ K-MEANS CLUSTERING	29
2.3.3.1 ΑΛΓΟΡΙΘΜΟΣ ΕΚΛΕΙΤΥΝΣΗΣ	29
2.4 PAM (PARTITIONING AROUND MEDOIDS)	31
2.4.1 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ	31



2.4.2 ΑΛΓΟΡΙΘΜΟΣ RAM	33
2.5 CLARA (CLUSTERING LARGE APPLICATIONS)	33
2.5.1 ΑΛΓΟΡΙΘΜΟΣ CLARA	34
2.6 CLARANS (CLUSTERING LARGE APPLICATIONS BASED ON RANDOMIZED SEARCH)	34
2.6.1 ΑΛΓΟΡΙΘΜΟΣ CLARANS	35
2.7 CUBE: ΙΕΡΑΡΧΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ	35
2.7.1 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ	36
2.7.2 ΑΛΓΟΡΙΘΜΟΣ CUBE	36
2.7.3 ΕΠΙΕΚΤΑΣΕΙΣ ΓΙΑ ΜΕΓΑΛΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ	38
2.8 DBSCAN	39
2.8.1 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ	40
2.8.2 ΑΛΓΟΡΙΘΜΟΣ DBSCAN	40
2.8.3 INCREMENTAL DBSCAN	41
2.9 SCALING KAI WEIGHTING	42
2.10 ΑΛΓΟΡΙΘΜΟΙ CLUSTERING ΓΙΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΜΕ ΛΕΚΤΙΚΕΣ ΤΙΜΕΣ	42
2.10.1 ROCK (ROBUST CLUSTERING ALGORITHM FOR CATEGORICAL ATTRIBUTES)	43
2.10.1.1 ΣΥΝΑΡΤΗΣΗ ΚΡΙΤΗΡΙΟ (CRITERION FUNCTION)	44
2.10.1.2 ΜΕΤΡΟ ΠΟΙΟΤΗΤΑΣ	45
2.10.1.3 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ ROCK	45
2.10.2 ΑΛΓΟΡΙΘΜΟΙ CLUSTERING ΒΑΣΙΣΜΕΝΟΙ ΣΤΟΝ K-MEANS ΓΙΑ ΛΕΚΤΙΚΑ ΔΕΔΟΜΕΝΑ	47
2.10.2.1 ΑΛΓΟΡΙΘΜΟΣ K-PROTOTYPES	47
2.10.2.2 ΑΛΓΟΡΙΘΜΟΣ K-MODES	48

3^ο ΚΕΦΑΛΑΙΟ

FUZZY CLUSTER ΑΝΑΛΥΣΗ	51
3.1 ΕΙΣΑΓΩΓΗ	51
3.2 FUZZY C-MEANS ΑΛΓΟΡΙΘΜΟΣ ΚΑΙ ΟΙ ΠΑΡΑΛΛΑΓΕΣ ΤΟΥ	52
3.2.1 FUZZY C-MEANS ΑΛΓΟΡΙΘΜΟΣ ΓΙΑ OBJECT-DATA	53
3.2.2 ΠΑΡΑΛΛΑΓΕΣ ΤΟΥ FUZZY C-MEANS (FCM)	54
3.3 ΕΦΑΡΜΟΓΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ C-MEANS CLUSTERING ΣΕ ΣΧΕΣΙΑΚΑ ΔΕΔΟΜΕΝΑ	56
3.3.1 ΑΛΓΟΡΙΘΜΟΣ C-MEANS ΓΙΑ ΣΧΕΣΙΑΚΑ ΔΕΔΟΜΕΝΑ	56
3.3.2 ΠΑΡΑΤΗΡΗΣΕΙΣ ΓΙΑ ΤΟΝ FCM ΓΙΑ ΣΧΕΣΙΑΚΑ ΔΕΔΟΜΕΝΑ	58
3.4 NOISE FUZZY CLUSTERING ΑΛΓΟΡΙΘΜΟΣ	59
3.5 CONDITIONAL FUZZY C-MEANS CLUSTERING	60

4^ο ΚΕΦΑΛΑΙΟ

AΞΙΟΛΟΓΗΣΗ ΠΟΙΟΤΗΤΑΣ CLUSTERING	62
4.1 ΕΙΣΑΓΩΓΗ	62
4.2 ΜΕΤΡΑ ΠΟΙΟΤΗΤΑΣ CLUSTERING	62
4.3 ΜΕΤΡΟ ΠΟΙΟΤΗΤΑΣ ΓΙΑ CRISP CLUSTERING	64
4.3.1 ΔΕΙΚΤΗΣ ΔΙΑΧΩΡΙΣΜΟΥ (SEPARATION INDEX)	64
4.4 ΑΞΙΟΛΟΓΗΣΗ FUZZY CLUSTERING	65
4.4.1 ΚΛΑΣΙΚΕΣ ΤΕΧΝΙΚΕΣ ΕΚΤΙΜΗΣΗΣ ΠΟΙΟΤΗΤΑΣ CLUSTERING ΓΙΑ ΤΟΝ FUZZY C-MEANS	65
4.4.1.1 PARTITION COEFFICIENT (PC)	66
4.4.1.2 ΕΝΤΡΟΠΙΑ ΤΜΗΜΑΤΟΣ(PARTITION ENTROPY - PE)	66

4.4.1.3 ΣΥΝΑΡΤΗΣΗ ΑΞΙΟΛΟΓΗΣΗΣ FUZZY CLUSTERING COMPACTNESS AND SEPARATION	67
4.4.1.4 ΆΛΛΑ ΚΛΑΣΙΚΑ ΜΕΤΡΑ	68
4.4.2 ΜΕΤΡΟ ΕΚΤΙΜΗΣΗΣ ΤΗΣ ΠΥΚΝΟΤΗΤΑΣ (COMPACTNESS) ΚΑΙ ΤΟΥ ΔΙΑΧΩΡΙΣΜΟΥ (SEPARATION) ΤΩΝ FUZZY C-ΤΜΗΜΑΤΩΝ	70
4.4.2.1 Ο ΔΕΙΚΤΗΣ COMPOSE WITHIN AND BETWEEN SCATTERING	70
4.4.2.2 ΑΞΙΟΛΟΓΗΣΗ CLUSTERS ΠΟΥ ΠΡΟΚΥΠΤΟΥΝ ΑΠΟ ΤΟΝ ΑΛΓΟΡΙΘΜΟ FCM	71

5^ο ΚΕΦΑΛΑΙΟ

ΥΠΟΣΤΗΡΙΞΗ ΑΒΕΒΑΙΟΤΗΤΑΣ ΣΤΟ DATA MINING	73
5.1 ΕΙΣΑΓΩΓΗ	73
5.2 FUZZY LOGIC	73
5.3 FUZZY LOGIC AND DATA MINING	74
5.4 ΠΡΟΣΕΓΓΙΣΗ FUZZY DATA MINING	75
5.4.1 ΠΕΡΙΓΡΑΦΗ ΣΥΣΤΗΜΑΤΟΣ ΥΠΟΣΤΗΡΙΞΗΣ ΑΒΕΒΑΙΟΤΗΤΑΣ ΣΕ DATA MINING.....	76
5.4.1.1 CVS EVALUATION	79

6^ο ΚΕΦΑΛΑΙΟ

ΥΠΟΣΤΗΡΙΞΗ ΤΗΣ ΑΣΑΦΕΙΑΣ ΣΤΟ CLUSTERING (FUZZY CLUSTERING)....	83
6.1 ΕΙΣΑΓΩΓΗ	83
6.2 ΠΡΟΣΕΓΓΙΣΗ FUZZY CLUSTERING	83
6.2.1 ΥΠΑΡΧΟΥΣΣΕΣ ΕΡΓΑΣΙΕΣ ΣΤΟ ΧΩΡΟ TOY FUZZY CLUSTERING	84
6.2.2 ΠΡΟΣΕΓΓΙΣΗ ΠΑΡΟΥΣΑΣ ΕΡΓΑΣΙΑΣ	84
6.2.2.1 ΣΥΝΔΕΣΗ ΤΩΝ ΠΑΡΑΠΑΝΩ ΒΗΜΑΤΩΝ ΜΕ ΤΗΝ ΔΙΑΔΙΚΑΣΙΑ CLASSIFICATION.....	85
6.3 ΤΕΧΝΙΚΗ ΠΡΟΣΕΓΓΙΣΗΣ ΕΡΓΑΣΙΑΣ	87
6.3.1 ΑΛΓΟΡΙΘΜΟΣ CLUSTERING.....	87
6.3.2 SCALING ΔΕΛΟΜΕΝΩΝ	87
6.3.3 ΕΠΛΟΓΗ ΚΑΛΥΤΕΡΟΥ ΣΧΗΜΑΤΟΣ CLUSTERING.....	88
6.3.3.1 ΟΡΙΣΜΟΣ ΣΥΝΑΡΤΗΣΗΣ ΕΚΤΙΜΗΣΗΣ ΠΟΙΟΤΗΤΑΣ CRISP CLUSTERING.....	88
6.3.3.2 ΕΥΡΕΣΗ ΒΕΛΤΙΣΤΟΥ ΑΡΙΘΜΟΥ CLUSTERS ΠΟΥ ΠΡΟΚΥΠΤΟΥΝ ΑΠΟ ΑΛΓΟΡΙΘΜΟΥΣ CRISP CLUSTERING.....	98
6.3.3.4 ΣΥΝΑΡΤΗΣΕΙΣ ΣΥΜΜΕΤΟΧΗΣ	98
6.3.3.4.1 HYPERTRAPEZOIDAL FUZZY MEMBERSHIP FUNCTIONS	99
6.3.3.5 ΕΚΤΙΜΗΣΗ ΠΛΗΡΟΦΟΡΙΑΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ ΣΧΗΜΑΤΩΝ CLUSTERING.....	102

7^ο ΚΕΦΑΛΑΙΟ

ΕΦΑΡΜΟΓΗ FUZZY CLUSTERING	105
7.1 ΕΙΣΑΓΩΓΗ	105
7.2 ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ	105
7.3 ΥΛΟΠΟΙΗΣΗ	106
7.3.1 CLUSTERING.....	106
7.3.2 ΑΞΙΟΛΟΓΗΣΗ CLUSTERING (CLUSTERING VALIDATION).....	109
7.3.3 ΣΥΝΑΡΤΗΣΕΙΣ ΣΥΜΜΕΤΟΧΗΣ	111
7.3.4 INTERFACE CLASSES	112



7.4 ΠΟΛΥΠΛΟΚΟΤΗΤΑ ΕΦΑΡΜΟΓΗΣ	113
7.5 ΠΑΡΟΥΣΙΑΣΗ ΕΦΑΡΜΟΓΗΣ	115
7.5.1 ΟΘΟΝΗ ΕΚΚΙΝΗΣΗΣ ΤΗΣ ΕΦΑΡΜΟΓΗΣ.....	115
7.5.2 ΟΘΟΝΗ ΕΠΛΟΓΗΣ ΠΙΝΑΚΑ.....	117
7.5.3 ΚΕΝΤΡΙΚΗ ΟΘΟΝΗ CLUSTERING.....	118
ΣΥΜΠΕΡΑΣΜΑΤΑ	121
ΠΑΡΑΡΤΗΜΑ Α'	123
ΠΑΡΑΡΤΗΜΑ Β'	133
Α. ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ-ΕΙΚΟΝΩΝ	135
Β. ΕΥΡΕΤΗΡΙΟ ΔΙΑΓΡΑΜΜΑΤΩΝ	135
Γ. ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ.....	136
ΒΙΒΛΙΟΓΡΑΦΙΑ	137



EXECUTIVE SUMMARY

1. The volume of stored data that are available and the need for analyzing, summarize and extract "knowledge" from this data, lead us to a new research area known as Knowledge Discovery & Data Mining.

The last decade has brought an explosive growth in our capabilities to both generate and collect data. Advances in database technology have provided us with the basic tools and methods for efficient data collection, storage and lookup of datasets. The result is that a flood of data has been generated and a growing data glut problem has been brought to the worlds of science, business and government. Also our ability to analyze, interpret large bodies of data and extract "useful" knowledge has outpaced and the need for new generation of tools and techniques for intelligent database analysis has been created. This need has been recognized by researchers in different areas (artificial intelligence, statistics, data warehousing, on-line analysis processing, expert systems and data visualization) and a new research area is emerged, known as *Data Mining*.

Data Mining is a step in the KDD process that is mainly concerned with methodologies for knowledge extraction from large data repositories. There are many data mining methods that are described and are available in literature. The most common of these are: *Cluster Analysis*, *Decision Trees*, *Neural Networks*. These methods accomplishing a limited set of tasks produces a particular enumeration of patterns over datasets. The main tasks according to well established data mining process are:

- *Clustering*
- *Classification*
- *Rule Extraction*
- *Estimation & Prediction*
- *Regression*
- *Summarization*

2. Clustering is one of the most useful tasks in Data Mining process.

Clustering is a common data mining task for discovering groups and identifying interesting distributions and patterns in the underlying data. The fundamental clustering problem is to partition a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters. For example, consider the database records containing the set of items purchased by customers. A clustering procedure could group the customers such that customers with similar buying patterns are in a single cluster.

In the clustering process, there are no predefined classes and no examples which would show what kind of desirable relations should be valid among the data. In this thesis, we present an overview of clustering methods as well as the main issues that we have to take in account in clustering process in order to have the optimum results in data mining.



More specifically, we referred to various clustering methods that are available and we present some of the most representative clustering algorithms that are proposed in literature. These algorithms can be classified into two basic types:

- *Hierarchical clustering* proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The result of the algorithm is a tree of clusters, called dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level a clustering of the data items into disjoint groups is obtained. In recent years, a number of hierarchical clustering algorithms has been proposed, such as ROCK, CUBE, PAM, CLARA, CLARANS e.t.c. The main difference of these algorithms is the rule according to which they decide which of smaller clusters will be merged or which of the larger clusters will be split.
- *Partitional clustering*, attempts to directly decompose the data set into a set of disjoint clusters. More specifically, they attempt to determine c partitions that optimize a certain criterion function. The criterion function may emphasize the local or global structure of the data and its optimization is an iterative procedure. The most commonly used partitional clustering algorithm is the K-Means.

For each of above types exists a wealth of subtypes and different algorithms for finding the clusters. Thus, according to the type of variables allowed in the data set or the technique they use to cluster data, the clustering methods can be categorized into[Huang97][GRK99][RR98]:

- *Statistical*, which are based on statistical analysis concepts. They use similarity measures to partition objects and they are limited to numeric data.
- *Conceptual*, which are used to cluster categorical data. They cluster objects according to the concepts objects carry.
- *Fuzzy*, which use fuzzy techniques to cluster data and they consider that an object can be classified to more than one clusters. This type of algorithms leads to clustering schemes that are compatible with everyday life experience as they handle the uncertainty of real data. A common fuzzy clustering algorithm is *Fuzzy C-Means*.
- *Kohonen net clustering*, which is based on the concepts of neural networks. The Kohonen network has input and output nodes. The input layer (input nodes) has a node for each attribute of the record, each one connected to every output node (output layer). Each connection is associated with a weight, which determines the position of the corresponding output node. Thus, according to an algorithm which changes the weights properly, output nodes move to form clusters.

3. The evaluation of clusters quality can lead to the optimum clustering schemes for our application.

Since clustering algorithms define clusters that are not known a priori, irrespective of the clustering methods (crisp, fuzzy), the final partition of data requires some kind of evaluation in most applications[RR98]. Another important issue in clustering is to find out the number of clusters that give the optimum partitioning. But what does it mean to say that a partitioning of data set is good ?



In general terms, there are two criteria proposed for clustering evaluation and selection of an optimal clustering scheme. These criteria are [Berry97]:

1. *Compactness*, the members of each cluster should be as close to each other as possible. A common measure of the first criterion is the variance, which should be minimized.
2. *Separation*, the clusters themselves to be widely spaced. There are three common approaches which measure the distance between two different clusters:
 - *Single linkage*: It measures the distance between the closest members of the clusters.
 - *Complete linkage*: It measures the distance between the most distant members.
 - *Comparison of centroids*: It measures the distance between the centroids of the clusters.

A reliable validity index must consider both compactness and separation of partitioning as well as the geometry of clusters. Based on the above criteria, a number of cluster validity indices are described in literature. A cluster validity index is proposed, D_1 , which attempts to identify "compact and separate clusters". The implementation of this measure is very expensive, especially when the number of clusters and number of objects in the dataset grows very large [XB91]. For fuzzy clustering, Bezdek proposed the partition coefficient (1974) and the classification entropy (1984). The limitations of these measures are its monotonic tendency with number of clusters and the lack of direct connection to the geometry of the data [Dave96]. Some other fuzzy validity measures are proposed in [GG89], [XB91], [RR98]. In this point, we have to mention that every validity index may fail in some cases since all of them are based on some parameters that may influence indices values and that may lead to unreliable results.

4. The objective of Data Mining process is the extraction of "useful" and comprehensible patterns.

A widely recognized requirement is that the patterns discovered must be valid and ultimately comprehensible. Another requirement addressed in KDD process is the reveal and usage of uncertainty in the data mining tasks, i.e. in the clustering and classification processes and association rules extraction. In this thesis we concentrate in the definition of a clustering scheme so as to support uncertainty. More specifically, we present an approach for the definition of optimum initial categories for a dataset based on well established clustering methods and quality while we propose a procedure based on fuzzy logic concepts in order to handle uncertainty.

5. Handling uncertainty in clustering process

5.1 Related work in fuzzy clustering

According to a well-established data mining process we can define/extract clusters which give the initial categories of a dataset based on widely known clustering methodologies that are available in literature. Then the database values can be classified into the categories defined and we are able to extract rules and other knowledge artifacts.

Most of the clustering algorithms result in crisp clusters, meaning that a data point either belongs to a class or not. The clusters are non-overlapping and this kind of partitioning is further called *crisp clustering*.

The issue of uncertainty support in clustering task leads to the introduction of algorithms that use fuzzy logic concepts in their procedure. A common fuzzy clustering algorithm is the Fuzzy C-Means(FCM), an extension of classical C-Means algorithm for fuzzy applications. FCM attempts to find the most characteristic point in each cluster which can be considered as the “center” of the cluster and, then, the grade of membership for each object in the clusters.

It is important, however, to handle uncertainty in clustering and classification process when we implement a crisp clustering algorithm. For this purpose, we need a procedure that maps discrete values to the fuzzy domain and defines degrees of belief in the classification process.

5.1 Proposed methodology

The main problem that we have to solve is as follows: Given a data set of n objects containing non-categorical data, we aim at

- *definition of a clustering scheme*, that represents the best partitioning of the specific data set based on a well defined quality measure,
- *definition of a mapping function for crisp clusters* defined in the previous step, based on fuzzy logic. This function will assign the values of non-categorical attributes to the clusters and produces degrees of belief in classification process.

The basic idea for our approach is to define a clustering and classification framework that supports uncertainty, irrespective of the clustering methods used (crisp or fuzzy). Moreover, we introduce a quality measure for clustering schemes. The basic steps of our approach can be described as follows:

1. *Clustering scheme extraction*. In this step we define/extract clusters that give the initial categories for a particular data set. We can use any of the well established clustering methods that are available.
2. *Evaluation of the clustering scheme*. The clustering methods can find a partition of our data set, assuming a-priori specified number of clusters. Our purpose is to define a number of clusters that is optimum for our data set. Thus, the extraction of clustering schemes is repeated for different number of clusters, and each one is evaluated using a set of quality clustering measures that we have defined.
3. *Definition of membership functions*. Fuzzy clustering algorithms define clusters and compute the grade of membership of each data value to the clusters. However, most of the clustering methods are crisp, i.e. all values in a cluster belong totally to it. As mentioned in previous section we aim at assigning uncertainty features in this case. This is achieved by assignment of appropriate mapping functions to the clusters.

The resulting clustering scheme is the input of another module that aims at classification and extraction of interesting knowledge artifacts based on the uncertainty of the data values [Vaz98]. The functions defined in step three(3) are used to classify the values of the whole data set to the clusters, defining a degree of belief

for this classification. Thus, we have a mapping of the data set values to the fuzzy domain. According to the classification framework described in [Vaz98], we can transform the whole data set into a Classification Value Space (CVS), using all the above information (clusters and mapping functions). The CVS is represented by a cube which cells hold the degrees of belief for the classification of the attributes values. Important decisions can be made during the evaluation of the information included in a CVS.

5.2 Technical approach

Based on above described methodology we implement a clustering system for non-categorical data so as to handle uncertainty. It is implemented in Java and uses ODBC to connect to a database (dataset).

The main characteristics of this system are as follows:

- We implement the crisp clustering algorithm ***K-means*** for the clustering process.
- Our system is implemented only to cluster **non-categorical data**.
- The clustering process supports ***multiple dimensions***, so as we have the chance to define categories taking in account more than one attributes simultaneously (e.g height, weight or height, weight, age etc).
- In the case of multidimensional data, we usually have to deal with the problem that different variables are measured in different units[Berry97]. It is clear that the values of data must be converted into a common scale before the clustering process takes place. For this purpose, we adapt a ***scaling method*** to our clustering procedure
- The clustering schemes that are produced by clustering algorithms, are evaluated using a ***clustering quality measure***. The objective of this measure is the definition of more compact and well-separated clusters. More specifically, in our approach we implement clustering for different numbers of clusters and we evaluate the clustering schemes using a well defined quality measure in order to select the optimum one for our application.

It is obvious that the selection of a good clustering measure (i.e. a reliable measure) is very important for our system. There are many validation indices available in literature. However, the issue of crisp cluster validity is under-addressed. Also, the evaluation of the proposed measures and the analysis of their reliability are limited. Thus, we define two quality measures for our approach based on concepts and other validity indices proposed in the literature. Then, we analyze their reliability in order to select the measure that give the most reliable results for our application.

- We define membership functions for clusters which give the degrees of belief in classification process based on ***Hypertrapezoidal Fuzzy Membership Functions (HFMFs)***. The main reasons of HFMFs selection are that they are proposed as a convenient mechanism for representing and calculating multidimensional fuzzy sets.

Using our system, we experiment with real-life datasets in order to measure the time complexity of the implemented system. More specifically, we use datasets of size 250-1000 tuples each one having 1-3 attributes, while the number of clusters

ranges from 2 to 10. The study shows that the execution time is almost linear to the number of tuples and is nearly quadratic with respect to the number of clusters.

5.3 Future Work

Further work will be concentrated in the following issues:

- implementation of the fuzzy clustering algorithm (i.e. Fuzzy C-means) in order to compare the results with those of the approach introduced in this paper.
- connection of the clustering process with the classification system described in [Vaz98] that supports uncertainty. Thus, the output of proposed clustering scheme (i.e. clusters and membership functions) will be used in the classification process in order to produce classification beliefs. The overall objective of this connection is the production of a data mining system that will overall handle uncertainty.
- Finally, it is important to extend our approach so as to support incremental clustering. The databases are frequently updated and, thus, the patterns derived from datasets by data mining methods have to be updated as well.

ΠΡΟΛΟΓΟΣ

Τις τελευταίες δεκαετίες οι δυνατότητες που παρέχονται για δημιουργία και συλλογή δεδομένων αυξάνονται με ταχύτατους ρυθμούς με αποτέλεσμα να οδηγούμαστε σε μεγάλους όγκους συσσωρευμένης πληροφορίας. Παράλληλα όμως η δυνατότητα για ανάλυση και εξαγωγή χρήσιμων προτύπων γνώσης περιορίστηκαν καθώς οι παραδοσιακές μέθοδοι δεν μπορούσαν να ανταποκριθούν στις απαιτήσεις μεγάλων συνόλων δεδομένων. Η ανάγκη για την ανάπτυξη νέων μεθόδων ανάλυσης των δεδομένων και εξαγωγής γνώσης έγινε σύντομα κατανοητή και οι προσπάθειες των ερευνητών από διάφορα επιστημονικά πεδία στράφηκαν προς την κατεύθυνση αυτή. Έτσι ένα νέο πεδίο έρευνας γνωστό ως Data Mining άρχισε να συγκεντρώνει το ενδιαφέρον τόσο του επιστημονικού όσο και του επιχειρηματικού κόσμου.

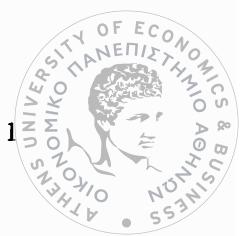
Το Data Mining έχει σαν βασικό σκοπό την εξαγωγή γνώσης από μεγάλα σύνολα δεδομένων. Η μορφή της γνώσης μπορεί να έχει διάφορες μορφές όπως classifications, association rules, decision trees κλπ. Σύμφωνα με καλά τεκμηριωμένες διαδικασίες data mining, εφαρμόζοντας κάποια διαδικασία clustering μπορούμε να εξάγουμε από το σύνολο των δεδομένων τις αρχικές κατηγορίες. Στην συνέχεια μπορούμε να ταξινομήσουμε τα δεδομένα μας στις κατηγορίες αυτές και να εξάγουμε κανόνες ή άλλες μορφές γνώσης χρήσιμες για την λήψη αποφάσεων.

Το αντικείμενο της παρούσας διπλωματικής εργασίας εντάσσεται στο χώρο του Data Mining και ειδικότερα στην μελέτη του Clustering. Το Clustering αφορά την εξαγωγή ομάδων ομοίων αντικειμένων (clusters) από μεγάλα σύνολα δεδομένων και αποτελεί μία από τις βασικές τεχνικές στην διαδικασία του Data Mining.

Οι περισσότερες μέθοδοι clustering που έχουν αναπτυχθεί και αναφέρονται στην βιβλιογραφία οδηγούν σε clusters με συγκεκριμένα όρια, δηλαδή κάθε στοιχείο ανήκει σε ένα και μόνο ένα cluster. Στην πραγματικότητα τα δεδομένα μας εμπεριέχουν κάποια ασάφεια και δεν είναι δυνατό να τα κατανείμουμε σε συγκεκριμένες ομάδες ομοίων αντικειμένων με απόλυτη βεβαιότητα. Για το λόγο αυτό θα ήταν χρήσιμο να αναπτυχθεί ένα σύστημα clustering που θα λάμβανε υπόψη του την ασάφεια και θα κατένειμε τα δεδομένα στα clusters με κάποιο βαθμό βεβαιότητας.

Στα πλαίσια της εργασίας αυτής έγινε μία προσπάθεια προσέγγισης του συγκεκριμένου προβλήματος με την μελέτη ενός συστήματος το οποίο ενσωματώνει στην διαδικασία του clustering βαθμούς ασάφειας. Ειδικότερα, ο στόχος της εργασίας είναι:

- Η μελέτη των διαφόρων τεχνικών clustering
- Η μελέτη και ανάπτυξη ενός συστήματος clustering το οποίο θα υποστηρίξει την αβεβαιότητα.



ΔΟΜΗ ΕΡΓΑΣΙΑΣ

Η εργασία αποτελείται ουσιαστικά από δύο μέρη:

- Το πρώτο αφορά στην μελέτη της παρούσας βιβλιογραφίας σε θέματα data mining και ειδικότερα clustering. Περιγράφονται διάφορες τεχνικές clustering και θέματα που σχετίζονται με την αξιολόγηση της ποιότητας του clustering.
- Το δεύτερο μέρος αφορά στην μελέτη ενός συστήματος το οποίο θα υποστηρίζει την αβεβαιότητα στην διαδικασία εξόρυξης γνώσης. Στο κομμάτι αυτό της εργασίας περιγράφεται και μία πρώτη προσέγγιση ανάπτυξης ενός συστήματος clustering το οποίον λαμβάνει υπόψη την ασάφεια.

Ειδικότερα, η δομή της εργασίας είναι η εξής:

Στο 1^ο Κεφάλαιο γίνεται μία εισαγωγή στις βασικές έννοιες του Data Mining καθώς και στις βασικές μεθόδους και τεχνικές εξαγωγής γνώσης.

Στο 2^ο Κεφάλαιο εξετάζονται οι διάφοροι μέθοδοι clustering και παρουσιάζονται οι κυριότεροι από τους κλασικούς αλγορίθμους clustering. Έχει επιλεχθεί η παρουσίαση αντιπροσωπευτικών αλγορίθμων από κάθε κατηγορία clustering (*partitionnal, hierarchical, conceptual*) ώστε να έχουμε μία όσο το δυνατόν καλύτερη και πιο ολοκληρωμένη εικόνα.

Στο 3^ο Κεφάλαιο είναι μία εισαγωγή στην έννοια του *fuzzy clustering* και τον ρόλο που μπορεί να παίξει στην διαδικασία του data mining. Επίσης παρουσιάζεται ο αντιπροσωπευτικότερος fuzzy clustering αλγόριθμος, ο *Fuzzy C-Means* καθώς και παραλλαγές αυτού.

Στο 4^ο Κεφάλαιο γίνεται αναφορά στην αξιολόγηση των clustering σχημάτων που προκύπτουν από την εφαρμογή των διαφόρων αλγορίθμων. Παρουσιάζονται τα βασικότερα κριτήρια ποιότητας clustering και περιγράφονται μερικά από τα προτεινόμενα μέτρα αξιολόγησης που αναφέρονται στην βιβλιογραφία.

Το 5^ο Κεφάλαιο αφορά στην ανάπτυξη ενός γενικότερου συστήματος data mining το οποίο θα ενσωματώνει στοιχεία της **Ασφαρούς Λογικής**. Συγκεκριμένα παρουσιάζεται η λογική ενός τέτοιου συστήματος καθώς και τα βασικά βήματα μιας μεθοδολογίας ανάπτυξης ενός συστήματος *fuzzy data mining*.

Το 6^ο Κεφάλαιο αποτελεί το βασικό κεφάλαιο του δεύτερου μέρους της εργασίας. Στο κεφάλαιο αυτό παρουσιάζεται η μεθοδολογία ανάπτυξης ενός συστήματος *fuzzy clustering*. Συγκεκριμένα, περιγράφεται η διαδικασία ενσωμάτωσης της ασάφειας σε μία κλασική διαδικασία clustering η οποία οδηγεί σε crisp clusters.

Στο 7^ο Κεφάλαιο παρουσιάζεται το σύστημα *fuzzy clustering* που αναπτύχθηκε σύμφωνα με την προσέγγιση του 6^{ου} κεφαλαίου ενώ περιγράφονται και οι βασικές δομές υλοποίησης του συστήματος.

Τέλος, στο Παράρτημα A' γίνεται μία σύντομη παρουσίαση της μορφής των αποτελεσμάτων που προκύπτουν από την εφαρμογή του συστήματος *fuzzy clustering*, το οποίο αναπτύχθηκε στα πλαίσια της διπλωματικής εργασίας. Για την παραγωγή των αποτελεσμάτων χρησιμοποιήθηκαν ενδεικτικά σύνολα πραγματικών δεδομένων.

Πριν προχωρήσουμε στην λεπτομερή παρουσίαση της διπλωματικής εργασίας θα ήθελα να εκφράσω τις ευχαριστίες μου στον καθηγητή μου κ. Μ. Βαζηργιάννη για την πολύ καλή συνεργασία που είχαμε, τις χρήσιμες παρατηρήσεις του και το υλικό που

μου παρείχε κατά την διάρκεια της εργασίας. Επίσης θα ήθελα να ευχαριστήσω τους συναδέλφους Χ. Αμανατίδη και Μ. Τζούρη, φοιτητές του τμήματος Πληροφορικής του Ο.Π.Α, και Β. Παπαπαναγιώτου, φοιτητή του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του ΕΜΠ, για την συνεργασία και την χρήσιμη ανταλλαγή απόψεων που είχαμε όλο αυτό το χρονικό διάστημα.

1^ο ΚΕΦΑΛΑΙΟ

KNOWLEDGE DISCOVERY AND DATA MINING ΣΤΙΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

1.1 ΕΙΣΑΓΩΓΗ

Η εξόρυξη πληροφορίας και γνώσης από μεγάλες βάσεις δεδομένων έχει αναγνωριστεί ως ένα θέμα κλειδί για την έρευνα στις βάσεις δεδομένων και στην μηχανική μάθηση (machine learning) καθώς και σαν μία πολύ σημαντική ευκαιρία για καινοτομίες στο χώρο της βιομηχανίας. Το *data mining* έχει γενικά προκαλέσει το ενδιαφέρον σε διάφορα πεδία έρευνας. Επίσης διάφορες εφαρμογές που εμφανίζονται στο χώρο της παροχής πληροφορίας, όπως είναι το *data warehousing* και οι *on-line* υπηρεσίες μέσω *Internet*, επικαλούνται διάφορες *data mining* τεχνικές με σκοπό να βοηθηθούν στην καλύτερη κατανόηση της συμπεριφοράς των πελατών και έτσι να βελτιώσουν τις παρεχόμενες υπηρεσίες και να επιτύχουν επιχειρηματικά πλεονεκτήματα.

Τα τελευταία χρόνια, οι δυνατότητες μας να παράγουμε και να συλλέγουμε δεδομένα έχουν αυξηθεί σημαντικά. Η ευρεία χρήση των υπολογιστών στις συναλλαγές μας σε όλους τους τομείς της σύγχρονης κοινωνίας (στο χώρο των επιχειρήσεων, της βιομηχανίας, των επιστημών) καθώς και τα πολλαπλά πλεονεκτήματα που παρέχουν τα διάφορα εργαλεία συλλογής δεδομένων έχουν οδηγήσει στην συγκέντρωση μεγάλου όγκου πληροφορίας. Ο αριθμός των βάσεων δεδομένων που χρησιμοποιούνται στην διοίκηση επιχειρήσεων, στην διαχείριση επιστημονικών δεδομένων και δεδομένων από τον χώρο της βιομηχανίας αυξάνεται με ταχύτατους ρυθμούς καθώς εμφανίζονται συστήματα βάσεων δεδομένων με περισσότερες δυνατότητες. Αυτή η μεγάλη αύξηση στον όγκο της πληροφορίας και των διαθέσιμων συστημάτων βάσεων δεδομένων έχει προκαλέσει επιτακτική ανάγκη για την εύρεση νέων τεχνικών και εργαλείων τα οποία θα υποστηρίζουν την αυτόματη μετατροπή των υπό επεξεργασία δεδομένων σε χρήσιμη πληροφορία και γνώση. Για το λόγο αυτό ένα νέο πεδίο έρευνας το οποίον αφορά στην διαδικασία εξόρυξης γνώσης και πληροφορίας από μεγάλα συστήματα βάσεων δεδομένων (KDD and data mining) άρχισε να κάνει την εμφάνιση του, ενώ έχει προκαλέσει το έντονο ενδιαφέρον πολλών και από διαφορετικά επιστημονικά πεδία ερευνητών, όπως συστημάτων βάσεων δεδομένων, συστημάτων βάσης γνώσης, τεχνητής νοημοσύνης, στατιστικής, χωρικών βάσεων δεδομένων, παρουσίασης δεδομένων (data visualization).

Στο κεφάλαιο αυτό θα περιγράψουμε τις βασικές έννοιες και μεθόδους της νέας ερευνητικής περιοχής που αποσκοπεί στην εξαγωγή χρήσιμης πληροφορίας από μεγάλους όγκους δεδομένων.

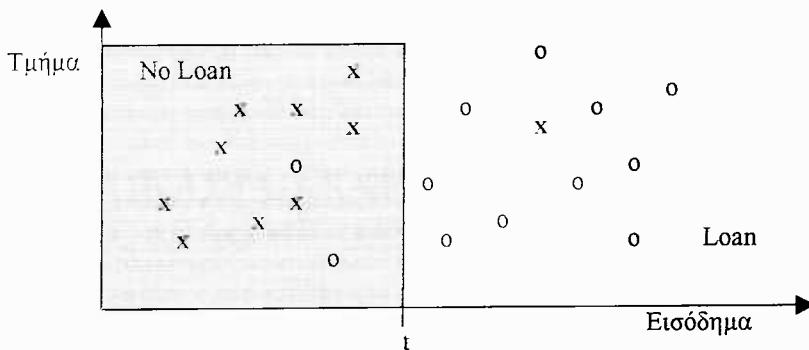
1.2 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Ο όρος *KDD* (*Knowledge Discovery in Database*) αναφέρεται στην συνολική διαδικασία εύρεσης χρήσιμης πληροφορίας από σύνολα δεδομένων. Μία γενικότερη έκφραση που παρουσιάζει με μεγαλύτερη σαφήνεια την έννοια του όρου KDD δόθηκε από τους Frawley, Piatesky-Shapiro & Matheus (1991)[FPSU96], σύμφωνα με τους:

"KDD είναι μία μη τετριμμένη διαδικασία εύρεσης έγκυρων, νέων, χρήσιμων και πλήρως κατανοητών προτύπων από τα δεδομένα".

Για να κατανοήσουμε καλύτερα τον ορισμό, θα εξετάσουμε με λεπτομέρεια τις βασικότερες από τις έννοιες που χρησιμοποιεί[FPSU96]:

- **Δεδομένα** είναι το σύνολο των περιπτώσεων που εμφανίζονται στην βάση δεδομένων. Για παράδειγμα θα μπορούσε να είναι μία συλλογή εγγραφών από την βάση δεδομένων μίας τράπεζας, οι οποίες θα περιέχαν τιμές τριών πεδίων (π.χ. για το τμήμα, το εισόδημα, την κατάσταση του δανείου).
- Τα **Πρότυπα** είναι εκφράσεις σε μία συγκεκριμένη γλώσσα οι οποίες περιγράφουν ένα υποσύνολο των δεδομένων. Για παράδειγμα ένα τέτοιο πρότυπο είναι η έκφραση: "Εάν το εισόδημα είναι $< \$t$, τότε ο υπάλληλος δεν μπορεί να λάβει δάνειο". Αυτό το πρότυπο περιγράφεται διαγραμματικά στο σχήμα 1.1.



Σχήμα 1.1. Χρήση απλού ορίου για την μεταβλητή "εισόδημα" προκειμένου να κατηγοριοποιήσουμε το σύνολο δεδομένων για τα δάνεια.

- Η *KDD* διεργασία είναι μία πολλαπλών σταδίων διαδικασία, η οποία περιλαμβάνει προετοιμασία των δεδομένων, αναζήτηση για πρότυπα, αξιολόγηση της γνώσης που ανακτάται από τα δεδομένα.
- *Έγκυρότητα*. Τα πρότυπα που προκύπτουν από την διαδικασία εξόρυξης γνώσης θα πρέπει να ισχύουν και σε νέα δεδομένα με κάποιο βεβαιότητας. Ένα

μέτρο βεβαιότητας είναι μία συνάρτηση αντιστοίχησης εκφράσεων σε κάποια γλώσσα και σ' έναν εν μέρει ή πλήρως διαβαθμισμένο χώρο μέτρησης. Για παράδειγμα εάν στο παραπάνω σχήμα το όριο για τα πρότυπα μετακινηθεί προς τα δεξιά, τότε το μέτρο βεβαιότητας θα μειωθεί καθώς περισσότερα αποδεκτά δάνεια θα περιληφθούν στην περιοχή μη αποδεκτών δανείων.

- Η *πιθανή χρησιμότητα* αφορά τα πρότυπα τα οποία θα πρέπει πιθανόν να οδηγούν σε κάποιες χρήσιμες ενέργειες, όπως η μέτρηση με κάποια συνάρτηση χρησιμότητας. Για παράδειγμα, στην περίπτωση των δεδομένων σχετικά με δάνεια θα μπορούσαμε να θεωρήσουμε μία συνάρτηση η οποία θα έδινε την αναμενόμενη αύξηση των κερδών μίας τράπεζας με βάση κάποιο κριτήριο απόφασης που μας δίνει τα πρότυπα (π.χ. στο σχήμα 1.1 το όριο του εισοδήματος για τους υπαλλήλους).
- Ο στόχος της ανακάλυψης γνώσης (knowledge discovery) από τις βάσεις δεδομένων είναι να δημιουργήσουμε πρότυπα (patterns) κατανοητά στους ανθρώπους προκειμένου τα επικείμενα δεδομένα να είναι πλήρως κατανοητά και να βοηθούν ακόμα και μη ειδικούς στην εξαγωγή χρήσιμων συμπερασμάτων.

Το *Data Mining* είναι ένα συγκεκριμένο βήμα στην επεξεργασία της ανακάλυψης γνώσης από βάσεις δεδομένων (KDD process) η οποία αποτελείται από συγκεκριμένους *algorίθμους data mining* και οι οποίοι κάτω από κάποιους υπολογιστικά αποδεκτούς περιορισμούς αποδοτικότητας παράγουν ένα συγκεκριμένο σύνολο προτύπων.

Το *data mining* ως στοιχείο της διαδικασίας ανακάλυψης γνώσης από σύνολα δεδομένων αφορά κυρίως τις διαδικασίες και τα μέσα με τα οποία θα εξάγονται τα πρότυπα από τα σύνολα των δεδομένων. Ενώ η ανακάλυψη γνώσης περιλαμβάνει την εκτίμηση και πιθανή διερμηνεία των προτύπων ώστε να προσδιοριστεί τι αποτελεί γνώση και τι όχι. Επίσης περιλαμβάνει την επιλογή κωδικοποίησης των σχημάτων, της κατάλληλης επεξεργασίας των δεδομένων πριν αυτά οδηγηθούν στο στάδιο του *data mining*.

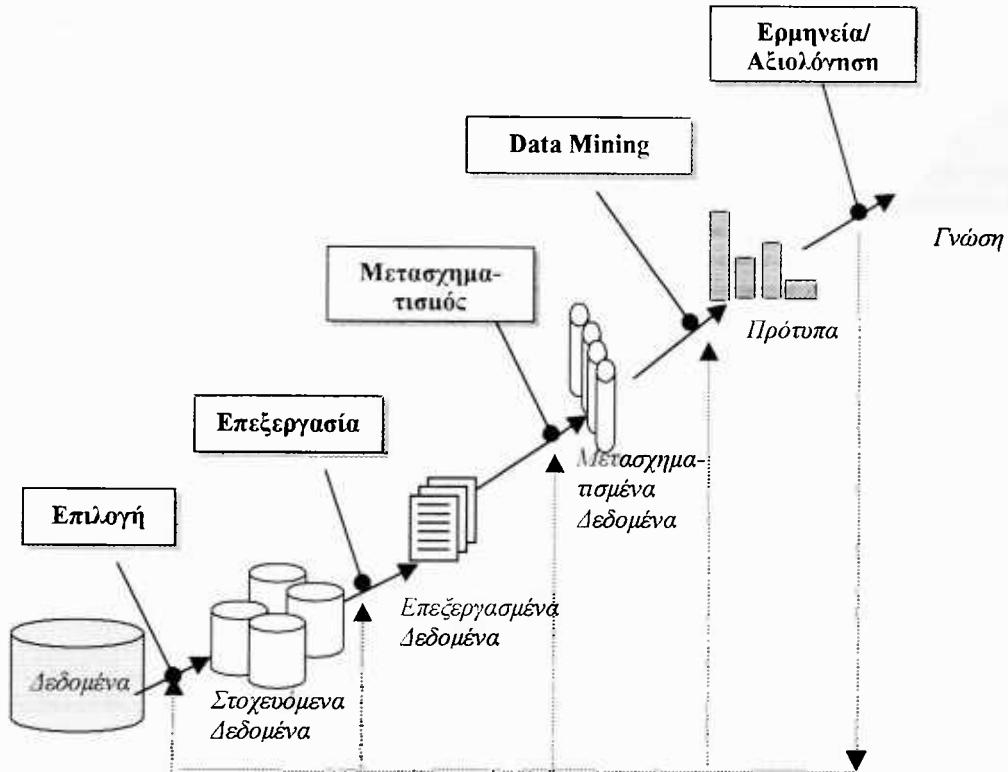
1.3 Η ΔΙΑΛΙΚΑΣΙΑ ΑΝΑΚΑΛΥΨΗΣ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ (KDD PROCESS)

Η KDD διαδικασία είναι μία αλληλεπιδραστική και επαναληπτική διαδικασία, η οποία περιλαμβάνει πλήθος βημάτων στα οποία θα πρέπει να ληφθούν αποφάσεις από τον χρήστη. Τα βασικότερα βήματα της διαδικασίας αυτής συνοψίζονται στα εξής [FPSU96](σχήμα 1.2):

- *Ανάπτυξη και κατανόηση των πεδίων της εφαρμογής περιλαμβανομένης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα και των στόχων των τελικών χρηστών.*
- *Δημιουργία των στοχευόμενον συνόλου δεδομένων.* Επιλέγουμε το σύνολο των δεδομένων ή επικεντρώνουμε το ενδιαφέρον μας στις μεταβλητές ή τα δείγματα δεδομένων από τα οποία πρόκειται να εξαχθεί η γνώση.

- *Καθάρισμα και επεξεργασία δεδομένων.* Το στάδιο αυτό περιλαμβάνει κάποιες βασικές λειτουργίες όπως για την απομάκρυνση του θορύβου από τα δεδομένα εάν θεωρείται απαραίτητο, για την συλλογή της απαιτούμενης πληροφορίας, ώστε να δημιουργήσουμε ένα μοντέλο ή να μετρήσουμε τον θόρυβο που υπεισέρχεται στα δεδομένα, για να αποφασίσουμε τις στρατηγικές που θα χρησιμοποιήσουμε για να χειριστούμε δεδομένα που τυχόν έχουν χαθεί, καθώς και για την καταμέτρηση της χρονικής ακολουθίας της πληροφορίας και των αλλαγών που τυγχάνει να συμβούν στα δεδομένα.
- *Εύρεση χρήσιμων χαρακτηριστικών* για να αναπαραστήσουμε τα δεδομένα ανάλογα με τους στόχους της εφαρμογής. Χρησιμοποιώντας μείωση διαστάσεων ή μεθόδους μετασχηματισμού γίνεται προσπάθεια μειώσεως του αριθμού των μεταβλητών που λαμβάνονται υπόψη ή επιτυγχάνεται η αναπαράσταση των δεδομένων ανεξάρτητα από μεταβλητές.
- *Επιλογή εργασιών data mining.* Στο στάδιο αυτό αποφασίζεται ποιες εργασίες data mining (π.χ. clustering, classification, regression κλπ.) θα εκτελεστούν κατά την διαδικασία KDD.
- *Επιλογή αλγορίθμου data mining.* Επιλέγουμε τις μεθόδους που πρόκειται να χρησιμοποιηθούν για την αναζήτηση προτύπων στα δεδομένα. Αυτό περιλαμβάνει απόφαση για το ποία μοντέλα και παράμετροι είναι οι κατάλληλοι να χρησιμοποιηθούν (π.χ. μοντέλα για δεδομένα με λεκτικές τιμές είναι διαφορετικά από τα μοντέλα για δεδομένα με αριθμητικές τιμές), καθώς και αντιστοίχηση μίας δεδομένης μεθόδου data mining με τα συνολικά κριτήρια της διαδικασίας KDD (π.χ. ο τελικός χρήστης μπορεί να ενδιαφέρεται περισσότερο να κατανοήσει το μοντέλο απ' ότι τις μεθόδους πρόβλεψης).
- *Data Mining.* Αναζήτηση των προτύπων που μας ενδιαφέρουν σε μία συγκεκριμένη μορφή αναπαράστασης ή σ' ένα σύνολο τέτοιων αναπαραστάσεων, όπως classification rules, trees, regression, clustering κλπ. Ο χρήστης μπορεί να βοηθήσει την μέθοδο εξόρυξης γνώσης με την σωστή εκτέλεση των προηγούμενων βημάτων.
- *Ερμηνεία των προτύπων* που έχουμε εξάγει από ένα σύνολο δεδομένων, επιστρέφοντας σε οποιοδήποτε από τα παραπάνω βήματα εάν θεωρηθεί απαραίτητο.
- *Ενοποίηση της γνώσης που έχει εξαχθεί.* Ενσωματώνουμε την γνώση αυτή στο σύστημά μας ή απλά παρουσιάζουμε την γνώση αυτή με την κατάλληλη τεκμηρίωση στα ενδιαφερόμενα μέλη. Επίσης ελέγχουμε και επιλύουμε τυχόν συγκρούσεις προηγούμενης γνώσης που υπήρχε ή είχε εξαχθεί.

Η *KDD διαδικασία* μπορεί να περιλαμβάνει επανάληψη μεταξύ οποιονδήποτε βημάτων. Η βασική ροή των βημάτων της διαδικασίας περιγράφεται διαγραμματικά στο σχήμα 1.2. Η περισσότερη εργασία κατά την διαδικασία της εξαγωγή γνώσης από τις βάσεις δεδομένων επικεντρώνεται στο data mining, ωστόσο και τα άλλα βήματα είναι ιδιαίτερης σημασίας για επιτυχή εφαρμογή της διαδικασίας.



Σχήμα 1.2. Τα στάδια που αποτελούν την διαδικασία KDD.

Στην συνέχεια θα επικεντρωθούμε στο κομμάτι της διαδικασίας εξαγωγής γνώσης που αφορά το data mining, το οποίο έχει συγκεντρώσει και το μεγαλύτερο ενδιαφέρον.

1.4 DATA MINING

Το *Data Mining* περιλαμβάνει την προσαρμογή μοντέλων στα εξεταζόμενα δεδομένα ή τον καθορισμό προτύπων από αυτά. Τα μοντέλα παίζουν τον ρόλο της γνώσης που εξάγεται από το σύνολο των δεδομένων. Η απόφαση για το αν τα μοντέλα αντανακλούν ή όχι χρήσιμη γνώση είναι μέρος της συνολικής διαδικασίας KDD για την οποία συνήθως απαιτείται η υποκειμενική ανθρώπινη κρίση.

Σήμερα υπάρχει πλήθος αλγορίθμων data mining οι οποίοι μπορεί να προέρχονται από τα διάφορα πεδία όπως από το χώρο της στατιστικής, της αναγνώρισης προτύπων, της μηχανικής γνώσης και των βάσεων δεδομένων. Οι περισσότεροι αλγόριθμοι μπορούμε να θεωρήσουμε ότι αποτελούνται από κάποιες βασικές τεχνικές και θεμελιώδεις αρχές. Συγκεκριμένα, οι αλγόριθμοι data mining αποτελούνται από τον συνδυασμό των εξής στοιχείων [FU96] [FPSU96]:

Το μοντέλο. Υπάρχουν δύο παράγοντες που σχετίζονται με το μοντέλο:

- *λειτουργία του μοντέλου (function of the model)* η οποία καθορίζει τις βασικές εργασίες κατά την διαδικασία data mining (π.χ. classification and clustering).
- *Ο τύπος αναπαράστασης του μοντέλου(representational form of the model)*. Η αναπαράσταση του μοντέλου καθορίζει τόσο την προσαρμοστικότητα του μοντέλου στην αναπαράσταση των δεδομένων όσο και την δυνατότητα ερμηνείας του μοντέλου με όρους κατανοητούς από τους ανθρώπους. Τυπικά, τα πιο πολύπλοκα μοντέλα προσαρμόζονται καλύτερα στα δεδομένα αλλά μπορεί να είναι περισσότερο δύσκολο να γίνουν κατανοητά και να προσαρμοστούν στην πραγματικότητα. Οι πιο γνωστές αναπαραστάσεις μοντέλων είναι τα δέντρα αποφάσεων και κανόνες, τα γραμμικά μοντέλα, τα μη γραμμικά μοντέλα(π.χ. νευρωνικά δίκτυα), τα μοντέλα που βασίζονται σε παραδείγματα(exampled-based) (π.χ. μέθοδοι βασισμένοι στις περιπτώσεις), τα γραφικά μοντέλα βασισμένα σε πιθανότητες (π.χ. Bayesian networks) και σχεσιακά μοντέλα.

Αξιολόγηση Μοντέλου. Με βάση κάποια κριτηρίων αξιολόγησης(π.χ. maximum likelihood) καθορίζεται πόσο καλά ένα συγκεκριμένο μοντέλο και οι παράμετροι του προσαρμόζεται στα κριτήρια της διαδικασίας KDD. Γενικά, η αξιολόγηση των μοντέλων αφορά τόσο στην εκτίμηση της εγκυρότητας των προτύπων όσο και στην εκτίμηση της ακρίβειας, της χρησιμότητας και της εύκολης κατανόησης του μοντέλου.

Αλγόριθμος αναζήτησης. Αφορά στον καθορισμό ενός αλγορίθμου για την εύρεση συγκεκριμένων μοντέλων και παραμέτρων, με βάση ένα συγκεκριμένο σύνολο δεδομένων, μία οικογένεια μοντέλων και ένα κριτήριο αξιολόγησης. Οι αλγόριθμοι αναζήτησης είναι δύο τύπων:

- *Αναζήτησης παραμέτρων*, οι οποίοι αναζητούν παραμέτρους που θα βελτιστοποιούν το κριτήριο αξιολόγησης του μοντέλου. Οι αλγόριθμοι εκτελούν την αναζήτηση λαμβάνοντας ως είσοδο ένα σύνολο δεδομένων και μία αναπαράστασης μοντέλου.
- *Αναζήτησης μοντέλου*, οι οποίοι εκτελούν μία επαναληπτική διαδικασία αναζήτησης μοντέλου για την αναπαράσταση των δεδομένων μας. Για μία συγκεκριμένη αναπαράσταση μοντέλου εκτελείται η μέθοδος αναζήτησης παραμέτρων και εκτιμάται η ποιότητα του συγκεκριμένου μοντέλου.

1.4.1 ΑΠΑΙΤΗΣΕΙΣ ΤΟΥ DATA MINING

Προκειμένου να επιτύχουμε ένα αποτελεσματικό data mining, θα πρέπει πρώτα να εξετάσουμε τι είδους χαρακτηριστικά αναμένεται να έχει ένα σύστημα εξόρυξης γνώσης καθώς και τις απαιτήσεις που θα πρέπει να λάβουμε υπόψη μας στην ανάπτυξη data mining τεχνικών.

Οι βασικότερες από τις απαιτήσεις είναι [CHY96] [AGGR98]:

- **Διαχείριση διαφορετικών τύπων δεδομένων.**

Καθώς διαφορετικοί τύποι δεδομένων και βάσεων δεδομένων χρησιμοποιούνται σε διαφορετικές εφαρμογές, είναι αναμενόμενο ότι το σύστημα εξόρυξης γνώσης θα πρέπει να έχει την δυνατότητα εκτέλεσης data mining με αποτελεσματικό

τρόπο πάνω σε διαφορετικά είδη δεδομένων. Οι περισσότερες βάσεις δεδομένων που είναι σήμερα διαθέσιμες είναι σχεσιακές. Έτσι είναι σημαντικό ένα σύστημα data mining να εκτελεί αποδοτική και αποτελεσματική εξόρυξη γνώσης σε σχεσιακά δεδομένα. Επιπρόσθετα, πολλές από τις βάσεις δεδομένων που χρησιμοποιούνται σήμερα περιέχουν πολυπλοκούς τόπους δεδομένων, όπως δομημένα δεδομένα και σύνθετα αντικείμενα, hypertext και δεδομένα πολυμέσων, χωρικά και χρονικά δεδομένα, κλπ. Ένα δυνατό σύστημα data mining θα πρέπει να μπορεί να εκτελέσει αποτελεσματικό data mining σε τέτοιους σύνθετους τύπους δεδομένων. Ωστόσο, η διαφοροποίηση των τύπων δεδομένων και οι διαφορετικοί στόχοι του data mining κάνουν μη ρεαλιστική την ύπαρξη ενός συστήματος data mining που θα χειρίζεται όλες τις περιπτώσεις. Αντίθετα θα πρέπει να αναπτύσσονται συγκεκριμένα συστήματα για συγκεκριμένα είδη δεδομένων, όπως συστήματα που θα εξάγουν γνώση από σχεσιακές ΒΔ, χωρικές ΒΔ, χρονικές ΒΔ, ΒΔ πολυμέσων κλπ.

- **Αποδοτικότητα και κλιμάκωση αλγορίθμων data mining**

Για την αποτελεσματική εξαγωγή πληροφορίας από ένα μεγάλο όγκο δεδομένων θα πρέπει οι αλγόριθμοι για την εξαγωγή γνώσης να είναι αποδοτικοί και προσαρμόσιμοι σε μεγάλες βάσεις δεδομένων. Αυτό σημαίνει ότι ο χρόνος εκτέλεσης των data mining αλγορίθμων θα πρέπει να είναι αναμενόμενος και αποδεκτός σε μεγάλες βάσεις δεδομένων. Αλγόριθμοι εκθετικής ή ακόμα πολυωνυμικής πολυπλοκότητας μέσης τάξης δεν θα ήταν κατάλληλοι.

- **Χρησιμότητα, βεβαιότητα και εκφραστικότητα των data mining αποτελεσμάτων.**

Η εξαγόμενη γνώση θα πρέπει να παρουσιάζει με ακρίβεια τα περιεχόμενα της βάσης δεδομένων. Η μη καταλληλότητα θα πρέπει να εκφράζεται με μέτρα αβεβαιότητας. Ο θόρυβος και τα δεδομένα που αποτελούν εξαιρέσεις θα πρέπει να χειρίζονται αποτελεσματικά από τα συστήματα data mining. Αυτό δίνει κίνητρα για μία συστηματική μελέτη μέτρησης της ποιότητας της εξαγόμενης γνώσης, κατασκευάζοντας στατιστικά, αναλυτικά μοντέλα, μοντέλα προσομοίωσης και εργαλεία.

- **Εκφραση διαφόρων ειδών data mining ερωτήσεων και αποτελεσμάτων.**

Διάφορα είδη γνώσης μπορούν να εξαχθούν από ένα μεγάλο σύνολο δεδομένων. Επίσης, μπορεί να θέλουμε να εξετάσουμε την γνώση που έχει εξαχθεί από διαφορετικές όψεις και να τις παρουσιάσουμε σε διαφορετικές μορφές. Αυτό δημιουργεί την ανάγκη να εκφράσουμε τόσο τις data mining ερωτήσεις όσο και την εξαγόμενη γνώση σε γλώσσες υψηλού επιπέδου ή μέσω γραφικών συστημάτων διεπαφής, έτσι ώστε η εργασία του data mining να μπορεί να εκτελεστεί από μη ειδικούς και η εξαγόμενη γνώση να μπορεί να χρησιμοποιηθεί άμεσα από τους χρήστες. Μία ακόμα απαίτηση για την αποτελεσματική παρουσίαση της γνώσης είναι το σύστημα να νιοθετεί εκφραστικές τεχνικές αναπαράστασης γνώσης.

- **Αλληλεπιδραστική εξόρυξη γνώσης σε πολλαπλά αφηρημένα επίπεδα.**

Η αλληλεπιδραστική εξαγωγή γνώσης δίνει την δυνατότητα σε έναν χρήστη να αλληλεπιδράσει με το σύστημα και να εκλεπτύνει την ερώτηση data mining, να αλλάξει δυναμικά το επίκεντρο των δεδομένων, να προωθήσει την διαδικασία data mining σε λεπτομερέστερο επίπεδο και να δει τα δεδομένα και τα αποτελέσματα

του data mining σε πολλαπλά αφαιρετικά επίπεδα και από πολλές διαφορετικές γωνίες.

- **Εξόρυξη πληροφορίας από διαφορετικές πηγές δεδομένων.**

Η ευρεία δικτύωση των υπολογιστών σε τοπικό αλλά και σε ευρύτερο επίπεδο, περιλαμβανομένου και του Internet, έχει συνδέσει πολλές πηγές δεδομένων δημιουργώντας μεγάλες κατανεμημένες και ετερογενείς βάσεις δεδομένων. Η εξόρυξη γνώσης από διαφορετικές πηγές δεδομένων με διαφορετική σημειολογία θέτει νέες απαιτήσεις στο data mining. Το μεγάλο μέγεθος των βάσεων δεδομένων, η ευρεία κατανομή των δεδομένων και υπολογιστική πολυπλοκότητα κάποιων μεθόδων data mining δίνουν το κίνητρο για την ανάπτυξη παράλληλων και κατανεμημένων data mining αλγορίθμων.

- **Προστασία των ιδιωτικών στοιχείων και ασφάλεια των δεδομένων.**

Όταν μπορούμε να δούμε τα δεδομένα από πολλές διαφορετικές γωνίες και από διαφορετικά επίπεδα αφαιρεσης, ο στόχος προστασίας των δεδομένων καθώς και της πληροφορίας ιδιωτικής φύσης απειλείται. Είναι σημαντικό να μελετήσουμε τέτοια θέματα και να δούμε τι μέτρα προστασίας μπορούν να αναπτυχθούν για την προστασία της ευαίσθητης πληροφορίας.

1.4.2 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΤΩΝ ΤΕΧΝΙΚΩΝ DATA MINING

Τα τελευταία χρόνια έχουν αναπτυχθεί πολλές τεχνικές data mining καθώς και συστήματα. Διαφορετικά σχήματα κατηγοριοποίησης μπορούν να χρησιμοποιηθούν για να κατηγοριοποιήσουν τις μεθόδους data mining και τα συστήματα, βασιζόμενοι στα είδη των βάσεων δεδομένων που πρόκειται να χρησιμοποιηθούν, στα είδη της γνώσης που εξάγονται και στα είδη των τεχνικών που χρησιμοποιούνται. Η κατηγοριοποίηση των συστημάτων εξόρυξης γνώσης βασίζεται στα εξής κριτήρια[CHY96]:

- **Τι είδος βάσης δεδομένων εξετάζουμε**

Ένα σύστημα data mining μπορεί να ταξινομηθεί σύμφωνα με τα είδη των βάσεων δεδομένων στα οποία εκτελείται το data mining. Για παράδειγμα, ένα σύστημα που χρησιμοποιείται για εξόρυξη γνώσης από σχεσιακά δεδομένα καλείται σχεσιακό σύστημα γνώσης. Εάν εξάγει γνώση από αντικειμενοστρεφή βάση δεδομένων καλείται αντικειμενοστρεφές σύστημα εξαγωγής γνώσης. Γενικά, ένα σύστημα εξόρυξης γνώσης μπορεί να κατηγοριοποιηθεί ανάλογα με τα διαφορετικά είδη των βάσεων δεδομένων που χρησιμοποιούνται, όπως σχεσιακές βάσεις δεδομένων, αντικειμενοστρεφείς βάσεις δεδομένων, χωρικές βάσεις δεδομένων, χρονικές βάσεις δεδομένων, βάσεις δεδομένων πολυμέσων κλπ.

- **Τι είδος γνώσης εξάγουμε**

Από ένα σύστημα εξόρυξης γνώσης μπορούν να εξαχθούν διάφορα είδη γνώσης, περιλαμβανομένων *association rules*, *classification rules*, *characteristic rules*, *clustering*.

Επίσης το σύστημα εξόρυξης γνώσης μπορεί να κατηγοριοποιηθεί σύμφωνα με το αφαιρετικό επίπεδο της εξαγόμενης γνώσης η οποία μπορεί να κατηγοριοποιηθεί σε γενική γνώση, γνώση πρώτου-επιπέδου και πολλαπλών επιπέδων γνώση.

- **Tι είδος τεχνική χρησιμοποιείται**

Τα συστήματα εξόρυξης γνώσης μπορούν να κατηγοριοποιηθούν σύμφωνα με το είδος των επικείμενων τεχνικών data mining. Για παράδειγμα, μπορούν να κατηγοριοποιηθούν σύμφωνα με την μέθοδο σε αυτόνομα συστήματα εξόρυξης γνώσης, σε οδηγούμενα από τα δεδομένα συστήματα, οδηγούμενα από τις ερωτήσεις και σε αλληλεπιδραστικά συστήματα δεδομένων. Επίσης ανάλογα με την προσέγγιση data mining που χρησιμοποιείται μπορούν να κατηγοριοποιηθούν σε γενικευμένη εξόρυξη, βασισμένη σε πρότυπα, εξόρυξη βασισμένη στην στατιστική ή μαθηματική θεωρία κλπ.

1.5 ΒΑΣΙΚΕΣ ΕΡΓΑΣΙΕΣ DATA MINING

Οι δύο βασικοί στόχοι του data mining πρακτικά είναι η *πρόβλεψη (prediction)* και η *περιγραφή (description)*. Η πρόβλεψη περιλαμβάνει την χρήση κάποιων μεταβλητών ή πεδίων στις βάσεις δεδομένων για να προβλέψουμε άγνωστες ή μελλοντικές τιμές άλλων μεταβλητών που έχουν ενδιαφέρον. Η περιγραφή επικεντρώνεται στην εύρεση προτύπων που περιγράφουν τα δεδομένα και τα οποία μπορούν να ερμηνευτούν από τον άνθρωπο. Η σχετική σημαντικότητα της πρόβλεψης και περιγραφής για συγκεκριμένες data mining εφαρμογές μπορούν να διαφέρουν σημαντικά. Ωστόσο, σε ότι αφορά την ανακάλυψη γνώσης (KDD), η περιγραφή τείνει να είναι περισσότερο σημαντική σε σχέση με την πρόβλεψη σε αντίθεση με τις εφαρμογές αναγνώρισης προτύπων και μηχανικής μάθησης που βασικός σκοπός είναι η πρόβλεψη.

Οι μέθοδοι data mining προκειμένου να επιτύχουν τους στόχους για την εξαγωγή και περιγραφή γνώσης από ένα σύνολο δεδομένων, χρησιμοποιούν ή εκτελούν κατά την εφαρμογή τους ένα σύνολο από εργασίες (tasks). Οι βασικότερες από αυτές τις εργασίες περιγράφονται στην συνέχεια της παραγράφου [Berry97][FPSU96].

1.5.1 CLASSIFICATION

Το *classification* αποτελεί μία από τις βασικές εργασίες (tasks) data mining. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νεοεμφανιζόμενου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να ταξινομηθούν αναπαριστάνονται γενικά από τις εγγραφές της βάσης δεδομένων και η διαδικασία του classification αποτελείται από την κατηγοριοποίηση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κλάσεις.

Η εργασία του classification χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κλάσεων και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προταξινομημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να ταξινομήσει δεδομένα που δεν έχουν ακόμα ταξινομηθεί (τοποθετηθεί σε κάποια από τις κλάσεις).

Στις περισσότερες περιπτώσεις, υπάρχει ένας περιορισμένος αριθμός κλάσεων και εμείς θα πρέπει να τοποθετήσουμε κάθε εγγραφή στην κατάλληλη κλάση. Για το σκοπό αυτό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο κατηγορίες. Η πρώτη χρησιμοποιεί *Δέντρα Αποφάσεων (Decision Trees)* και η

δεύτερη Νευρωνικά Δίκτυα (*Neural Networks*). Και οι δύο στηρίζονται στην ιδέα της “εκπαίδευσης” (training) με τη βοήθεια ενός υποσυνόλου δεδομένων που ονομάζεται *training set*. Το υποσύνολο αυτό επιλέγεται σαν αντιπροσωπευτικό δείγμα του συνολικού όγκου δεδομένων. Με την εφαρμογή της διαδικασίας εκπαίδευσης καθορίζονται κάποια πρότυπα για τις κατηγορίες δεδομένων. Έτσι, όταν προκύψει ένα νέο δεδομένο τότε μπορεί εύκολα να κατηγοριοποιηθεί. Για τη διαδικασία αυτή χρησιμοποιούνται είτε τεχνικές βασισμένες στα νευρωνικά δίκτυα είτε συμβολικές τεχνικές. Στις πρώτες υπάρχει το φαινόμενο της αμφίδρομης αναμετάδοσης και επεξεργασίας δεδομένων ενώ στη δεύτερη υπάρχουν μοντέλα δένδρων αποφάσεων ή μοντέλα για IF...THEN...ELSE ανάλυση.

1.5.2 CLUSTERING

Το *clustering* είναι η εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών υποομάδων, *clusters*. Αυτό που διαφοροποιεί το clustering από το classification είναι ότι το clustering δεν βασίζεται σε προκαθορισμένες κλάσεις. Στο classification, ο πληθυσμός διαιρείται σε κλάσεις αναθέτοντας κάθε στοιχείο ή εγγραφή σε μία προκαθορισμένη κλάση με βάση ένα μοντέλο που αναπτύσσεται μέσω της εκπαίδευσης του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων.

Στο clustering δεν υπάρχουν προκαθορισμένες κλάσεις. Οι εγγραφές ομαδοποιούνται σε σύνολα με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Επαφίεται σε εμάς να καθορίσουμε την σημασία που θα έχει κάθε ένα από τα clusters που προκύπτουν. Για παράδειγμα, τα clusters συμπτωμάτων μπορεί να υποδεικνύουν διαφορετικές ασθένειες, clusters που περιλαμβάνουν τα χαρακτηριστικά που σχετίζονται με τα φύλλα και τον καρπό φυτών μπορεί να υποδεικνύουν διαφορετικές ποικιλίες ενός φυτού.

Το clustering μπορεί να χρησιμοποιηθεί και σαν εισαγωγή σε κάποια άλλη μορφή data mining ή μοντελοποίησης. Για παράδειγμα, το clustering μπορεί να χρησιμοποιηθεί σαν πρώτο βήμα στην προσπάθεια μερισμού της αγοράς. Αντί δηλαδή να προσπαθούμε να προσδιορίσουμε τι είδος promotion θα ταιριάζει καλύτερα σε κάθε πελάτη, μπορούμε να κατηγοριοποιήσουμε τους πελάτες αρχικά σε ομάδες (clusters) απόμων που παρουσιάζουν τις ίδιες συνήθειες σχετικά με την αγορά προϊόντων και στην συνέχεια να προσδιορίσουμε το είδος του promotion που ταιριάζει σε κάθε ομάδα.

1.5.3 ΕΞΑΓΩΓΗ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ (ASSOCIATION RULES EXTRACTION)

Στην περίπτωση αυτή έχουμε σύνολα από αντικείμενα ή εγγραφές, κάθε ένα από τα οποία περιέχει έναν αριθμό από αντικείμενα τα οποία ανήκουν σε μία δεδομένη συλλογή. Μία συνάρτηση συσχέτισης είναι μία συνάρτηση που εφαρμόζεται σε ένα σύνολο εγγραφών η οποία επιστρέφει σχέσεις ή πρότυπα τα οποία υπάρχουν στην συλλογή αυτή των αντικειμένων. Τα πρότυπα αυτά μπορεί να εκφραστούν με κανόνες, των οποίων η γενική μορφή είναι “**If X then Y**”.

Η εξαγωγή των κανόνων γίνεται με την βοήθεια κάποιων αλγορίθμων, οι οποίοι αποδεικνύονται αρκετά αποδοτικοί. Έπειτα από την ανάλυση και εύρεση των κανόνων θα πρέπει να διαπιστωθεί κατά πόσο είναι έγκυροι και σημαντικοί για την εφαρμογή μας. Υπάρχουν δύο συντελεστές οι οποίοι αναφέρονται σε αυτό το θέμα: είναι ο *support factor* και ο *confidence factor*. Έτσι για τον κανόνα $X \rightarrow Y$, ο πρώτος υποδεικνύει το ποσοστό των εγγραφών που ισχύει ο συνδυασμός X και Y , ενώ ο δεύτερος αναφέρεται στο ποσοστό των εγγραφών που όταν ισχύει το X ισχύει και το Y . Για παράδειγμα, στην έκφραση "72% των εγγραφών που περιέχουν τα αντικείμενα A, B και C επίσης περιέχει και τα αντικείμενα D και E", το ποσοστό των συμβάντων (72) καλείται *confidence factor* του κανόνα.

Ένα παράδειγμα χρήσης των συσχετίσεων είναι στην ανάλυση των αιτήσεων που υποβάλλονται από τους ασθενείς στις ασφαλιστικές εταιρίες. Κάθε αίτηση περιέχει ένα σύνολο από ιατρικές διαδικασίες που εκτελέστηκαν σε ένα συγκεκριμένο ασθενή κατά την διάρκεια μίας επίσκεψης. Ορίζοντας το σύνολο των αντικείμενων που αποτελούν όλες τις ιατρικές διαδικασίες που μπορούν να εκτελεστούν σε κάθε ασθενή καθώς και τις εγγραφές που αντιστοιχούν σε κάθε αίτηση, η εφαρμογή μπορεί να βρει με την βοήθεια της συνάρτησης συσχέτισης την σχέση που υπάρχει ανάμεσα στις ιατρικές διαδικασίες που εμφανίζονται πιο συχνά μαζί.

1.5.4 ESTIMATION & PREDICTION

Σε αυτή την κατηγορία χρησιμοποιούνται δύο ειδών τεχνικές: η **γραμμική και η μη γραμμική παλινδρόμηση**. Στην πρώτη περίπτωση ο αλγόριθμος προσπαθεί να βρει μία γραμμή η οποία να προσεγγίζει με την μεγαλύτερη δυνατή πιθανότητα τις τιμές από ένα σύνολο σημείων του επιπέδου. Στην δεύτερη περίπτωση χρησιμοποιούνται κάποιοι μη γραμμικοί όροι για να μπορέσει το μοντέλο να πλησιάσει ακόμη περισσότερο το σύνολο των δεδομένων. Παρόλα αυτά, όμως, δεν είναι σίγουρο ότι μία τέτοια προσέγγιση μπορεί να καλύψει όλο το σύνολο των δεδομένων με σχετική ασφάλεια.

Η **RBF (Radial Basis Function)** είναι μία τεχνική για πρόβλεψη τιμών που παρουσιάζει μεγαλύτερη ευστάθεια και ευελιξία σε σχέση με τις παραδοσιακές τεχνικές. Η τεχνική αυτή βασίζεται στην επιλογή όχι μίας αλλά πολλών μη γραμμικών συναρτήσεων οι οποίες έχουν διαφορετικά βάρη στον τρόπο με τον οποίο επηρεάζουν τα δεδομένα. Τα RBFs μπορούν να χρησιμοποιηθούν για διαφορετικές περιοχές δεδομένων εισόδου. Με αυτόν τον τρόπο προσπαθεί κανείς να πλησιάσει όσο το δυνατόν με μεγαλύτερη ακρίβεια τα δεδομένα της εξόδου.

1.5.5 REGRESSION

To **regression** αναφέρεται στην εκμάθηση μίας συνάρτησης η οποία αντιστοιχεί τα δεδομένα σε μία μεταβλητή πρόβλεψης (prediction variable) πραγματικής τιμής. Οι εφαρμογές του regression είναι πάρα πολλές π.χ. εκτίμηση της πιθανότητας ένας ασθενής να έχει κάποια ασθένεια δεδομένων των αποτελεσμάτων ενός συνόλου διαγνωστικών tests, πρόβλεψη της ζήτησης ενός νέου προϊόντος από τους πελάτες σαν συνάρτηση των εξόδων για διαφήμιση.

1.5.6 SUMMARIZATION

Περιλαμβάνει μεθόδους για την εύρεση μίας περιγραφής για ένα υποσύνολο δεδομένων. Ένα απλό παράδειγμα θα μπορούσε να είναι η εκτίμηση της μέσης και της τυπικής απόκλισης για όλα τα πεδία. Πιο εξεζητημένες λειτουργίες περιλαμβάνουν την παραγωγή συνοπτικών κανόνων, τεχνικές παρουσίασης πολλαπλών μεταβλητών και την ανακάλυψη λειτουργικών σχέσεων μεταξύ των μεταβλητών. Οι εργασίες του *summarization* χρησιμοποιούνται συχνά στην αλληλεπιδραστική ανάλυση δεδομένων και στην αυτοματοποιημένη παραγωγή αναφορών.

1.5.7 ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΑΛΛΑΓΗΣ ΚΑΙ ΑΠΟΚΛΙΣΗΣ (CHANGE AND DEVIATION DETECTION)

Η λειτουργία αυτή επικεντρώνεται στην εύρεση των σημαντικότερων αλλαγών στα δεδομένα λαμβάνοντας υπόψη προηγούμενες μετρήσεις.

1.6 DATA MINING ΜΕΘΟΔΟΙ

Υπάρχει μία μεγάλη γκάμα από διαφορετικές μεθόδους data mining. Στην παράγραφο όμως αυτή θα επικεντρωθούμε σε ένα υποσύνολο μόνο των μεθόδων αυτών, οι οποίες αποτελούν και τις κυριότερες μεθόδους data mining [FPSU96][Berry97].

1.6.1 CLUSTER ANALYSIS

Στο περιβάλλον της μη εποπτευόμενης μάθησης το σύστημα θα πρέπει να προσδιορίσει τις κλάσεις του και ένας τρόπος για να το επιτύχει αυτό είναι να γίνει clustering στα δεδομένα της βάσης. Το πρώτο βήμα είναι να προσδιοριστούν τα υποσύνολα των σχετικών αντικειμένων και στην συνέχεια να καθορισθούν οι περιγραφές τους, οι οποίες και περιγράφουν κάθε μία από τις κλάσεις αυτές.

Το clustering τημηματοποιεί την βάση δεδομένων έτσι ώστε κάθε τμήμα ή ομάδα να περιέχει στοιχεία παρόμοια σύμφωνα με κάποιο κριτήριο ή μέτρο. Το clustering σύμφωνα με την ομοιότητα μπορεί να εμφανιστεί σε πολλούς επιστημονικούς κλάδους. Εάν είναι διαθέσιμο ένα μέτρο ομοιότητας υπάρχει ένας μεγάλος αριθμός τεχνικών για την διαμόρφωση των clusters. Η συμμετοχή των ομάδων μπορεί να βασίζεται στο επίπεδο ομοιότητας ανάμεσα στα μέλη και με βάση αυτό μπορεί να καθορισθούν και οι κανόνες συμμετοχής. Μία άλλη προσέγγιση είναι η δημιουργία ενός συνόλου συναρτήσεων οι οποίες μετρούν κάποιες ιδιότητες τμημάτων δηλαδή τα τμήματα σαν συνάρτηση κάποιων παραμέτρων του τμήματος.

Πολλές data mining εφαρμογές χρησιμοποιούν το clustering σύμφωνα με την ομοιότητα για να τμηματοποιήσουν για παράδειγμα μία βάση πελάτη / πωλητή. Το clustering σύμφωνα με την βελτιστοποίηση του συνόλου συναρτήσεων χρησιμοποιείται στην ανάλυση δεδομένων π.χ. όταν καθορίζουμε τιμολόγια

ασφάλειας οι πελάτες μπορούν να τημηματοποιηθούν σύμφωνα με έναν αριθμό παραμέτρων ώστε να επιτευχθεί η βέλτιστη τημηματοποίηση τιμολογίων.

To clustering στις βάσεις δεδομένων είναι οι επεξεργασίες διαχωρισμού ενός συνόλου δεδομένων σε στοιχεία τα οποία αντανακλούν ένα συνεπές πρότυπο συμπεριφοράς. Μόλις καθορισθούν τα πρότυπα μπορούν να χρησιμοποιηθούν για να διασπαστούν τα δεδομένα σε πιο κατανοητά υποσύνολα, ενώ μπορούν να παρέχουν και υπο-ομάδες του πληθυσμού για παραπέρα ανάλυση η οποία είναι σημαντική όταν έχουμε να κάνουμε με μεγάλες βάσεις δεδομένων.

1.6.2 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ

Τα Δένδρα Αποφάσεων (Decision Trees) είναι πολύ ισχυρά και δημοφιλή εργαλεία για classification και prediction. Τα Δένδρα Αποφάσεων αντιπροσωπεύουν κανόνες, οι οποίοι μπορούν εύκολα να διατυπωθούν σε φυσική γλώσσα ώστε να είναι εύκολα κατανοητοί από τους ανθρώπους ή να διατυπωθούν σε μία γλώσσα προσπέλασης βάσεων δεδομένων π.χ. σε SQL. Υπάρχει μια πληθώρα αλγορίθμων που αναλαμβάνουν να φτιάξουν Δένδρα Αποφάσεων, όπως : *CART* (Classification and Regression Trees), *CHAID* (CHi-squared Automation Interaction Detection), ένας πιο πρόσφατος πολλά υποσχόμενος αλγόριθμος είναι ο *C4.5*.

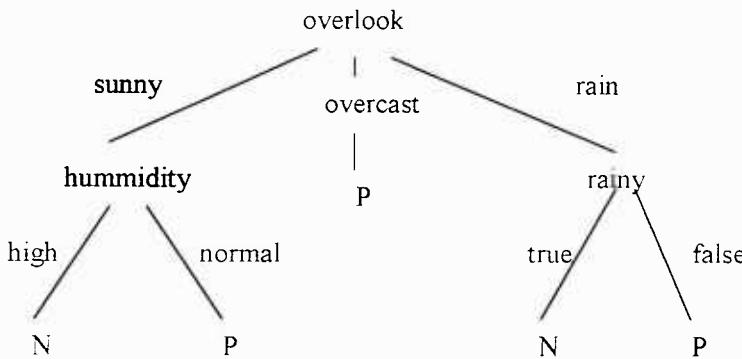
Γενικά ένα Δένδρο Απόφασης αντιπροσωπεύει μια σειρά από **IF THEN** κανόνες που συνδυάζονται μεταξύ τους από τη ρίζα του δένδρου προς τα φύλλα. Οι κόμβοι του δέντρου χαρακτηρίζονται με τα ονόματα των χαρακτηριστικών, οι ακμές ονομάζονται με τις δινατές τιμές που μπορεί να πάρει ένα χαρακτηριστικό και τα φύλλα με τις διάφορες κλάσεις. Τα αντικείμενα ταξινομούνται ακολουθώντας ένα μονοπάτι που οδηγεί προς τα κάτω στο δέντρο, λαμβάνοντας τις ακμές που αντιστοιχούν στις τιμές των χαρακτηριστικών ενός αντικειμένου.

Μία εγγραφή εισέρχεται στο δέντρο από τον κόμβο της κορυφής. Στην ρίζα, εφαρμόζεται έλεγχος για να καθορισθεί ποιο κόμβο παιδί θα ακολουθήσει στην συνέχεια η εγγραφή. Υπάρχουν διάφοροι αλγόριθμοι για την επιλογή του αρχικού ελέγχου, αλλά ο στόχος είναι πάντα ο ίδιος, δηλαδή, να επιλέξουμε τον έλεγχο ο οποίος διαχωρίζει καλύτερα τις τελικές κλάσεις. Η επεξεργασία αυτή επαναλαμβάνεται μέχρι η εγγραφή να φτάσει στο κόμβο φύλλο. Όλες οι εγγραφές οι οποίες καταλήγουν σε ένα συγκεκριμένο φύλλο ταξινομούνται με τον ίδιο τρόπο. Υπάρχει ένα μοναδικό μονοπάτι που οδηγεί από την ρίζα σε κάθε φύλλο. Το μονοπάτι αυτό είναι μία έκφραση του κανόνα που χρησιμοποιείται για να ταξινομήσουμε τις εγγραφές.

Πολλά διαφορετικά φύλλα μπορούν να οδηγούν στην ίδια ταξινόμηση, αλλά κάθε φύλλο κάνει την ταξινόμηση αυτή για διαφορετικό λόγο. Για παράδειγμα, σε ένα δέντρο το οποίο ταξινομεί φρούτα και λαχανικά με βάση το χρώμα, οι τελικοί κόμβοι του δέντρου απόφασης για τα μήλα, ντομάτες και κεράσια θα πρέπει όλα να προβλέπουν "κόκκινο", παρά τον διαφορετικό βαθμό πίστης καθώς υπάρχουν πράσινα μήλα και μαύρα κεράσια.

Στο σχήμα 1.3 παρουσιάζεται ένα παράδειγμα αντικειμένων το οποίο περιγράφει τον καιρό σε μία δεδομένη στιγμή. Κάποια αντικείμενα τα οποία είναι θετικά

παραδείγματα δηλώνονται ως P και άλλα τα οποία είναι αρνητικά δηλώνονται ως N. Το classification στην περίπτωση αυτή είναι η κατασκευή ενός δέντρου το οποίο μπορεί να χρησιμοποιηθεί για να ταξινομήσει τα αντικείμενα με σωστό τρόπο.



Σχήμα 1.3. Δέντρα Αποφάσεων

1.6.3 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (NEURAL NETWORKS)

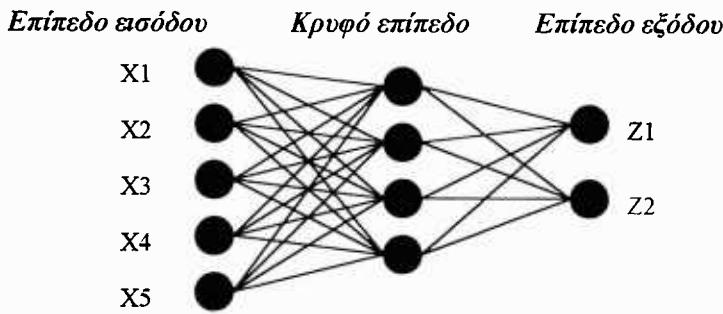
Τα νευρωνικά δίκτυα αποτελούν μία πολύ δυνατή, γενικού σκοπού τεχνική η οποία μπορεί να εφαρμοστεί για πρόβλεψη(*prediction*), *classification* και *clustering*. Η εμφάνιση των νευρωνικών δικτύων έχει σαν στόχο να γεφυρώσει το κενό μεταξύ των υπολογιστών και του ανθρώπινου μναλού. Οι άνθρωποι μπορούν να εξάγουν συμπεράσματα με βάση την εμπειρία τους ενώ οι υπολογιστές βασίζονται σε συγκεκριμένες οδηγίες. Τα νευρωνικά δίκτυα στοχεύουν στο να μειώσουν αυτό το κενό. Όταν χρησιμοποιούνται σε καλά ορισμένο περιβάλλον, η ικανότητα τους να παράγουν και να μαθαίνουν από τα δεδομένα, μιμείται την ικανότητα των ανθρώπων να μαθαίνουν από τις εμπειρίες τους. Αυτή η ικανότητα είναι χρήσιμη για το data mining κάνοντας συγχρόνως τα νευρωνικά δίκτυα μία σημαντική περιοχή για έρευνα, υποσχόμενα νέα και καλύτερα αποτελέσματα στο μέλλον.

Τα νευρωνικά δίκτυα είναι μία προσέγγιση ανάπτυξης και εκτίμησης μαθηματικών δομών με την δυνατότητα να μαθαίνουν. Οι μέθοδοι αυτοί είναι αποτελέσματα ακαδημαϊκών ερευνών με στόχο την μοντελοποίηση συστημάτων μάθησης. Τα νευρωνικά δίκτυα έχουν την ικανότητα να εξάγουν κάποιο συμπέρασμα από πολύπλοκα ή μη ακριβή δεδομένα και μπορούν να χρησιμοποιηθούν για να εξάγουν πρότυπα και να προσδιορίζουν τάσεις οι οποίες είναι πολύ πολύπλοκες για να προσδιοριστούν από ανθρώπους ή από άλλες υπολογιστικές τεχνικές. Ένα εκπαιδευμένο νευρωνικό δίκτυο μπορεί να αντιμετωπίστει ως ένας "ειδικός" για την κατηγορία της πληροφορίας που του δόθηκε να αναλύσει. Έτσι μπορεί να χρησιμοποιηθεί για να κάνει κάποιες προβλέψεις, όταν προκύψουν κάποιες νέες περιπτώσεις.

Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μναλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα δίκτυο το οποίο μπορεί να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων μόλις αυτά παρουσιαστούν μέσα στα δεδομένα,

δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι. Αυτό διακρίνει τα νευρωνικά δίκτυα από τα παραδοσιακά προγράμματα υπολογιστών, τα οποία απλά ακολουθούν οδηγίες σύμφωνα με μία καλά ορισμένη σειρά.

Η δομή των νευρωνικών δικτύων είναι ανάλογη με αυτή του σχήματος 1.4.



Σχήμα 1.4. Δομή ενός νευρωνικού δικτύου.

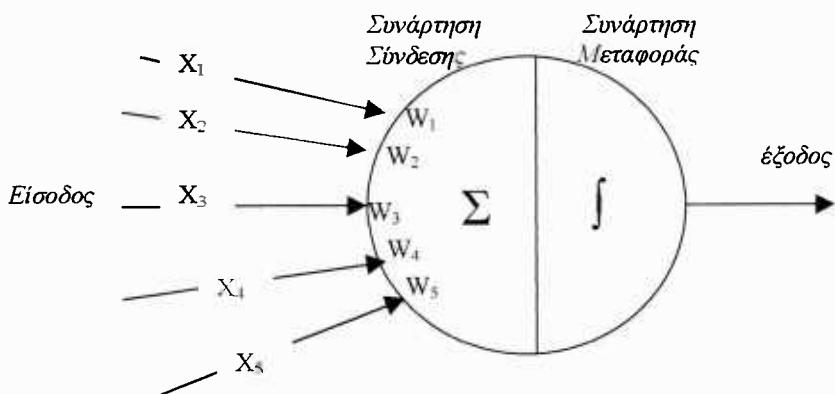
Το αριστερό επίπεδο αναπαριστά το επίπεδο εισόδου, στην περίπτωση του σχήματος έχουμε πέντε εισόδους με ετικέτες X_1, X_2, \dots, X_5 . Το μεσαίο επίπεδο είναι αυτό που καλείται *κρυφό επίπεδο* (*hidden level*), το οποίο έχει μεταβλητό αριθμό κόμβων. Το μεσαίο επίπεδο είναι και αυτό που εκτελεί το μεγαλύτερο μέρος της εργασίας του δικτύου. Το επίπεδο εξόδου (επίπεδο στα δεξιά) έχει δύο κόμβους στο παράδειγμά μας Z_1 και Z_2 , οι οποίες αναπαριστούν τις τιμές εξόδου που προσπαθούμε να προσδιορίσουμε από τις εισόδους. Για παράδειγμα, μπορεί με την βοήθεια ενός κατάλληλα εκπαιδευμένου δικτύου να προβλέψουμε τις πωλήσεις (έξοδος) βασιζόμενοι στις παλιές πωλήσεις, την τιμή και την εποχή (είσοδοι).

1.6.3.1 ΚΟΜΒΟΙ ΝΕΥΡΩΝΙΚΟΥ ΔΙΚΤΥΟΥ

Τα νευρωνικά δίκτυα, όπως προαναφέραμε, αποτελούνται από βασικές μονάδες (κόμβους) που σχεδιάζονται για να μοντελοποιήσουν την συμπεριφορά των βιολογικών νευρώνων (σχήμα 1.5). Κάθε κόμβος στο μεσαίο επίπεδο είναι πλήρως συνδεδεμένος με τις εισόδους, γεγονός που σημαίνει ότι το κρυφό πεδίο βασίζεται σε όλες τις εισόδους τις οποίες και συνδυάζει στις τιμές εξόδου. Ο συνδυασμός αυτός καλείται *συνάρτηση ενεργοποίησης* του κόμβου.

Η συνάρτηση ενεργοποίησης έχει δύο μέρη. Το πρώτο μέρος είναι *η συνάρτηση σύνδεσης* (*combination function*) η οποία συνδυάζει όλες τις εισόδους σε μία απλή τιμή. Κάθε είσοδος έχει το δικό της βάρος. Η πιο κοινή συνάρτηση σύνδεσης είναι το άθροισμα όλων των εισόδων πολλαπλασιασμένων με το αντίστοιχο βάρος τους ($X_1 * W_1 + X_2 * W_2 + \dots + X_N * W_N$). Σε ορισμένες περιπτώσεις είναι χρήσιμες άλλες συναρτήσεις και περιλαμβάνουν το μέγιστο των εισόδων πολλαπλασιασμένων με το βάρος τους, το ελάχιστο, ή το λογικό AND ή OR των τιμών. Ωστόσο, η συνάρτηση που βασίζεται στο άθροισμα των εισόδων πολλαπλασιασμένων με τα βάρη τους δουλεύει καλύτερα στην πράξη.

Το δεύτερο μέρος της συνάρτησης ενεργοποίησης είναι η **συνάρτηση μεταφοράς (transfer function)**, η οποία μεταφέρει την τιμή της συνάρτησης σύνδεσης στην έξοδο. Υπάρχουν τρία είδη συναρτήσεων μεταφοράς: η σιγμοειδής, γραμμική και η συνάρτηση υπερβολικής εφαπτομένης (*hyperbolic tangent*). Η γραμμική συνάρτηση έχει περιορισμένη πρακτική σημασία αντίθετα με τις άλλες δύο (μη γραμμικές συναρτήσεις) οι οποίες παρουσιάζουν μη γραμμική συμπεριφορά.



Σχήμα 1.5. Η μονάδα επεξεργασίας(κόμβος) του νευρωνικού δικτύου.

2^ο ΚΕΦΑΛΑΙΟ

ΜΕΘΟΔΟΙ CLUSTERING

2.1 ΕΙΣΑΓΩΓΗ

Το clustering αποτελεί μία από της βασικές διαδικασίες data mining με κύριο στόχο να κατηγοριοποιήσει ή να ομαδοποιήσει δεδομένα ή αντικείμενα τα οποία παρουσιάζουν κάποια ομοιότητα μεταξύ τους. Το clustering αναφέρεται ως διαδικασία μη εποπτευόμενης μάθησης, που σημαίνει ότι δεν βασίζεται σε προκαθορισμένες κλάσεις (κατηγορίες). Συνεπώς, εάν αποφασίσουμε να εφαρμόσουμε clustering σε ένα σύνολο δεδομένων, δεν υπάρχει κάποιο συγκεκριμένο σύνολο παραδειγμάτων το οποίο θα μπορούσε να μας υποδείξει ποιες είναι οι επιθυμητές σχέσεις που θα πρέπει να ισχύουν μεταξύ των δεδομένων.

Το clustering προσπαθεί να εξάγει από ένα σύνολο δεδομένων, ομάδες αντικειμένων (εγγραφών Β.Δ.) όπου κάθε ομάδα θα αναπαριστά ένα χαρακτηριστικό τμήμα των αντικειμένων τα οποία θα έχουν κάποια κοινά στοιχεία. Για παράδειγμα, εφαρμόζοντας clustering σε ένα σύνολο εγγραφών που περιγράφουν τα άτομα και το ύψος τους, μπορούμε να ομαδοποιήσουμε τα άτομα σε κοντούς, μετρίου αναστήματος και ψηλούς. Με τον τρόπο αυτό, μπορούμε να εξάγουμε πιο εύκολα κανόνες σχετικά με την συμπεριφορά των αντικειμένων μίας συγκεκριμένης ομάδας παρά εξετάζοντας ανεξάρτητες εγγραφές ενός συνόλου δεδομένων. Η ιδέα είναι ότι τα στοιχεία τα οποία ανήκουν στο ίδιο cluster, γεγονός που σημαίνει ότι έχουν παρόμοια χαρακτηριστικά, αναμένεται να συμπεριφέρονται ανάλογα. Οπότε ένας κανόνας ο οποίος είναι έγκυρος για ένα από τα στοιχεία αυτά υπάρχει μεγάλη πιθανότητα να είναι έγκυρος και για τα άλλα στοιχεία που είναι παρόμοια με αυτό.

2.2 ΚΑΤΗΓΟΡΙΕΣ CLUSTERING

Οι διάφοροι μέθοδοι που χρησιμοποιούμε κατά την διαδικασία του clustering μπορούν να κατηγοριοποιηθούν ανάλογα :

- με τον τύπο των μεταβλητών που επιτρέπουν να συμμετέχουν στην βάση δεδομένων που μας ενδιαφέρει,
- τον τρόπο που απεικονίζουν τα clusters,
- τον τρόπο που οργανώνουν τα clusters, δηλαδή iεραρχικά, σε επίπεδες λίστες κτλ.
- τους αλγορίθμους που χρησιμοποιούν.

Έτσι ανάλογα με τον τύπο των μεταβλητών και την θεωρία που αποτελεί την βάση των αλγορίθμων που εφαρμόζονται για το clustering των δεδομένων, μπορούμε να διακρίνουμε τα εξής είδη clustering:

- *Statistical Clustering*
- *Conceptual Clustering*
- *Kohonen Net Clustering*
- *Fuzzy Clustering*

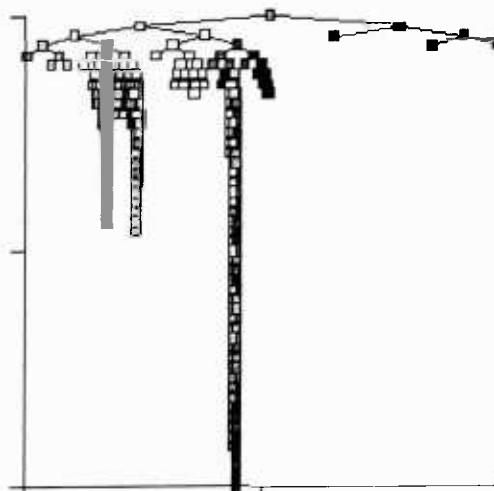
Επίσης ανάλογα με τον τρόπο που οι αλγόριθμοι clustering οργανώνουν τα clusters έχουμε τα εξής είδη clustering[Kaski97]:

- *Iεραρχικό (Hierarchical) clustering*
- *Clustering διαμερισμού (Partitional clustering)*

Για κάθε έναν από τους τύπους αυτούς υπάρχει πλήθος άλλων υποτύπων και αλγορίθμων για την εύρεση των clusters.

2.2.1 ΙΕΡΑΡΧΙΚΟ CLUSTERING

Το *ιεραρχικό clustering* προχωρά διαδοχικά είτε συνδυάζοντας μικρότερα clusters σε μεγαλύτερα ή διασπώντας μεγαλύτερα clusters. Οι μέθοδοι clustering διαφέρουν στο κανόνα με βάση τον οποίο αποφασίζεται ποια από τα μικρότερα clusters θα συγχωνευτούν για την δημιουργία κάποιου μεγαλύτερου, ή ποιο μεγάλο cluster θα διασπαστεί. Το τελικό αποτέλεσμα του αλγορίθμου είναι ένα δέντρο από clusters το οποίο καλείται δενδρογράφημα (σχήμα 2.1) και το οποίο παρουσιάζει τον τρόπο που τα clusters σχετίζονται μεταξύ τους. Εάν κόψουμε το δενδρογράφημα σε κάποιο επίπεδο που επιθυμούμε μπορούμε να έχουμε το clustering των δεδομένων μας σε ομάδες μη σχετιζόμενες.



Σχήμα 2.1. Δενδρογράφημα

2.2.2 PARTITIONAL CLUSTERING

Το *partitional clustering*, βασίζεται στην άμεση αποσύνθεση του συνόλου των δεδομένων σε ένα σύνολο μη σχετιζόμενων clusters. Η συνάρτηση που ο αλγόριθμος clustering προσπαθεί να ελαχιστοποιήσει μπορεί να δίνει έμφαση στην τοπική δομή των δεδομένων, αναθέτοντας clusters στα άκρα της συνάρτησης(ελάχιστο, μέγιστο) ή στην γενική δομή των δεδομένων. Τυπικά, το γενικό κριτήριο είναι η ελαχιστοποίηση κάποιων μέτρων ανομοιότητας μεταξύ των δειγμάτων μέσα σε κάθε ένα από τα clusters, καθώς και η μεγιστοποίηση την ανομοιότητας μεταξύ διαφορετικών clusters.

2.2.3 STATISTICAL CLUSTERING

Αυτή η μορφή clustering έχει τις ρίζες της στο πεδίο της στατιστικής ανάλυσης. Οι αλγόριθμοι παράγουν κλάσεις βασιζόμενοι σε μέτρα αριθμητικής ομοιότητας μεταξύ των αντικειμένων. Περιορίζεται δηλαδή στο ότι μπορεί να εφαρμοστεί σε βάσεις δεδομένων με τύπο γνωρισμάτων αριθμητικές τιμές.

Κάθε αντικείμενο περιγράφεται από ένα σύνολο γνωρισμάτων, των οποίων οι τιμές είναι αριθμητικές. Μία τυπική εγγραφή(αντικείμενο) που αφορά στην περιγραφή κάποιου ατόμου μπορεί να είναι η εξής:

Attribute	Height	Weight	IQ
<i>Value</i>	1.85	180.0	100

Η περιγραφή ενός αντικείμενου μπορεί να αναπαρασταθεί με την βοήθεια ενός διανύσματος ως εξής:

Object1(1.85, 180.0, 100)
Object2(1.75, 195.0, 80)
Object3(1.45, 135.0, 55)

Προκειμένου να μετρήσουμε την ομοιότητα ή την απόσταση μεταξύ των αντικειμένων θα πρέπει να χρησιμοποιήσουμε κάποιο μέτρο απόστασης. Ένα τέτοιο μέτρο μπορεί να είναι η *Ευκλείδεια απόσταση* ή η *City-block Απόσταση*.

$$\text{Ευκλείδεια Απόσταση} = \sqrt{\sum (x_i - y_i)^2}$$

$$\text{City-block Απόσταση} = \sum |x_i - y_i|$$

Όπου x_i και y_i είναι τα στοιχεία των δύο διανυσμάτων των αντικειμένων X και Y.

2.2.4 CONCEPTUAL CLUSTERING

Αντίθετα με το statistical clustering που περιορίζεται σε αντικείμενα με αριθμητικές τιμές, το conceptual clustering μπορεί να εφαρμοστεί σε βάσεις δεδομένων με τύπο

γνωρισμάτων μόνο κείμενο (text). Συνεπώς, το conceptual clustering μπορεί να εφαρμοστεί σε αντικείμενα που έχουν την εξής μορφή:

Attribute Value	Height Tall	Weight Heavy	IQ Average
--------------------	----------------	-----------------	---------------

Οι γεωμετρικές αποστάσεις δεν είναι κατάλληλες στην περίπτωση αυτή προκειμένου να εκτιμηθεί η απόσταση μεταξύ αντικειμένων της παραπάνω μορφής. Μία εναλλακτική διαδικασία που μπορεί να χρησιμοποιηθεί είναι ο αριθμός των γνωρισμάτων που δύο αντικείμενα δεν έχουν κοινά.

Για παράδειγμα έστω τα αντικείμενα:

Object1(Tall, Heavy, Average)
Object2(Tall, Heavy, Low)
Object3(Short, Light, High)

Η απόσταση μεταξύ του αντικειμένου 1 και αντικειμένου 2 είναι 1 καθώς διαφέρουν μόνο στην τιμή του IQ. Η απόσταση μεταξύ του αντικειμένου 2 και 3 είναι 3 καθώς έχουν διαφορετικές τιμές και για τα τρία γνωρίσματα.

2.2.5 FUZZY CLUSTERING

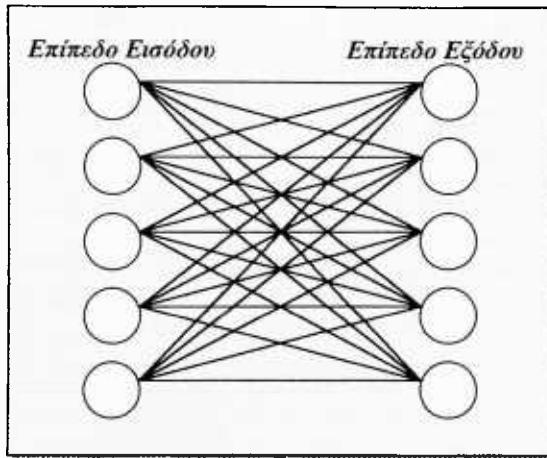
Η μέθοδος fuzzy clustering χρησιμοποιεί fuzzy τεχνικές για την εκτέλεση clustering σε ένα σύνολο δεδομένων. Οι fuzzy αλγόριθμοι συνήθως προσπαθούν να βρουν εκείνα τα σημεία που χαρακτηρίζουν καλύτερα κάθε cluster. Τα σημεία αυτά μπορούν να θεωρηθούν ως το "κέντρο" των clusters και στην συνέχεια προσδιορίζονται οι κατάλληλοι βαθμοί συμμετοχής για κάθε αντικείμενο (εγγραφή) στα clusters. Στο fuzzy clustering δηλαδή θεωρούμε ότι ένα αντικείμενο μπορεί να ανήκει σε περισσότερα από ένα clusters.

2.2.6 KOHONEN NET CLUSTERING

Τα νευρωνικά δίκτυα Kohonen παρέχουν έναν τρόπο κατηγοριοποίησης των δεδομένων μέσω self-organizing δίκτυων τεχνητών νευρώνων. Δύο βασικές έννοιες που κυριαρχούν στα δίκτυα Kohonen και είναι σημαντικό να κατανοήσουμε είναι, η ανταγωνιστική μάθηση (*competitive learning*) και η αυτό-οργάνωση (*self-organization*) [KNN1].

Ανταγωνιστική μάθηση είναι απλά η εύρεση ενός νευρώνα ο οποίος προσεγγίζει περισσότερο το πρότυπο εισόδου. Το δίκτυο στη συνέχεια τροποποιεί αυτό τον νευρώνα και τους γειτονικούς του (ανταγωνιστική μάθηση με αυτό-οργάνωση) έτσι ώστε να μοιάζουν περισσότερο με το πρότυπο.

Ένα δίκτυο Kohonen αποτελείται από ένα επίπεδο κόμβων εισόδου και ένα επίπεδο κόμβων εξόδου [KNN2]. Το επίπεδο εξόδου είναι περισσότερο γνωστό σαν επίπεδο Kohonen (σχήμα2.2).

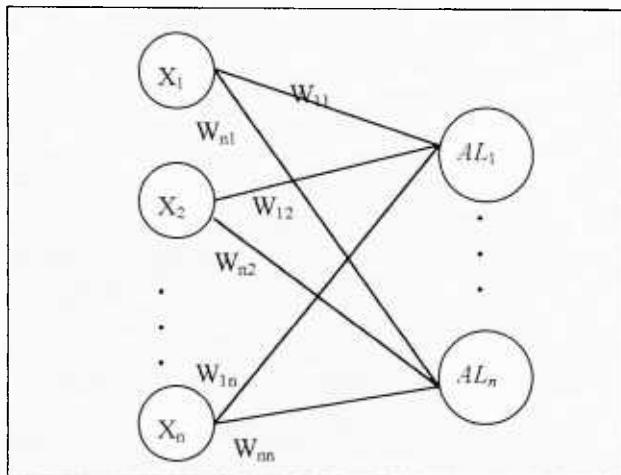


Σχήμα 2.2. Δίκτυο Kohonen

Η βασική ιδέα πίσω από το δίκτυο Kohonen είναι ότι κάθε κόμβος εισόδου θα ανταγωνιστεί με τους άλλους κόμβους εισόδου. Αυτό σημαίνει ότι μόνο μία έξοδος θα είναι ενεργή για μία δεδομένη είσοδο. Σε κάθε νευρώνα εξόδου αντιστοιχεί ένα σύνολο βαρών (διάνυσμα βάρους) κάθε ένα από τα οποία αντιστοιχεί σε ένα από τα δεδομένα εισόδου. Στο δίκτυο για κάθε νευρώνα υπολογίζεται το "επίπεδο ενεργοποίησης"(activation level). Το επίπεδο ενεργοποίησης ενός κόμβου εξόδου j υπολογίζεται ως εξής [KANN](σχήμα 2.3):

$$\text{Activation level}_j = \sqrt{\sum_{i=1}^n (W_{ij} - X_i)^2}$$

Αυτή είναι μία απλή Ευκλείδεια απόσταση μεταξύ των σημείων που αναπαριστώνται από το διάνυσμα του βάρους και το διάνυσμα εισόδου. Ειτι ένας κόμβος του οποίου το διάνυσμα βάρους προσεγγίζει το διάνυσμα εισόδου θα έχει μικρό επίπεδο ενεργοποίησης και ο κόμβος του οποίου το διάνυσμα βάρους διαφέρει από το διάνυσμα εισόδου θα έχει μεγάλο επίπεδο ενεργοποίησης. Ο κόμβος στο δίκτυο με το μικρότερο επίπεδο ενεργοποίησης θεωρείται ο "νικητής" για το τρέχον διάνυσμα εισόδου.



Σχήμα 2.3. Υπολογισμός activation level σε ένα δίκτυο Kohonen

Στην διάρκεια της διαδικασίας εκπαίδευσης το δίκτυο αναπαριστάται κάθε φορά με ένα πρότυπο εισόδου και όλοι οι κόμβοι υπολογίζουν τα επίπεδα ενεργοποίησης τους όπως περιγράψαμε παραπάνω. Ο κόμβος που είναι ο "νικητής" και κάποιοι από τους κόμβους γύρω από αυτόν έχουν την δυνατότητα να διευθετήσουν τα διανύσματα βάρους τους ώστε να προσεγγίσουν το τρέχον διάνυσμα εισόδου καλύτερα. Οι κόμβοι που περιλαμβάνονται στο σύνολο των κόμβων οι οποίοι μπορούν να προσαρμόσουν τα βάρη τους λέμε ότι ανήκουν στην "γειτονιά" του νικητή. Το μέγεθος της γειτονιάς του νικητή διαφοροποιείται κατά την διάρκεια της διαδικασίας εκπαίδευσης. Επίσης το ποσοστό κατά ο οποίο οι κόμβοι που ανήκουν στην γειτονιά του νικητή επιτρέπεται να προσαρμόζουν τα βάρη τους μειώνεται γραμμικά κατά την διάρκεια της περιόδου εκπαίδευσης.

Ο παράγοντας ο οποίος καθορίζει το μέγεθος αλλαγής των βαρών καλείται ρυθμός μάθησης. Οι διευθετήσεις για κάθε στοιχείο του διανύσματος βάρους γίνονται σύμφωνα με την εξίσωση:

$$\Delta W_i = -a(W_i - X_i) \quad (\text{Εξισ. 2.1})$$

όπου a = ρυθμός μάθησης, ΔW = αλλαγή βάρους.

Αλγόριθμος Kohonen

Τα βασικά βήματα του Kohonen αλγορίθμου είναι τα εξής [KNN2]:

Βήμα 1. Για κάθε νευρώνα στο επίπεδο Kohonen λαμβάνεται ένα πλήρες αντίγραφο ενός προτύπου εισόδου.

Βήμα 2. Βρίσκουμε το νευρώνα που είναι ο "νικητής". Ο νικητής είναι αυτός με το μικρότερο επίπεδο ενεργοποίησης:

$$AL_j = \sqrt{\sum_{i=1}^n (W_{ij} - X_i)^2}$$

Βήμα 3. Για κάθε νευρώνα που είναι "νικητής" καθώς και για τους φυσικούς γειτονικούς του κόμβους, χρησιμοποιείται ο ακόλουθος κανόνας εκπαίδευσης για την τροποποίηση των βαρών:

$$\begin{aligned} W_{ij}(t+1) &= W_{ij}(t) + a(t) * gamma(t) * [X_i - W_{ij}(t)] \\ gamma(t) &= exp\{-0.5 * [r_{ij} / sigma(t)]^2\} \end{aligned}$$

όπου a είναι ο ρυθμός μάθησης ο οποίος μειώνεται με το χρόνο (αρχίζει από την τιμή 1 και μειώνεται σταδιακά μέχρι την τιμή 0), r_{ij} είναι η απόσταση μεταξύ του νικητή και του κόμβου που πρόκειται να ενημερωθεί και $sigma$ είναι η ακτίνα γειτονίας η οποία μειώνεται με το χρόνο.

Βήμα 4. Επανάληψη των βημάτων 1-3 για κάθε νέο πρότυπο εισόδου.

Βήμα 5. Επανάληψη βήματος 4 έως ότου όλα τα πρότυπα εισόδου περάσουν(αυτό καθορίζει την τιμή του t).

Βήμα 6. Επανάληψη βήματος 5 για ένα καθορισμένο αριθμό φορών.

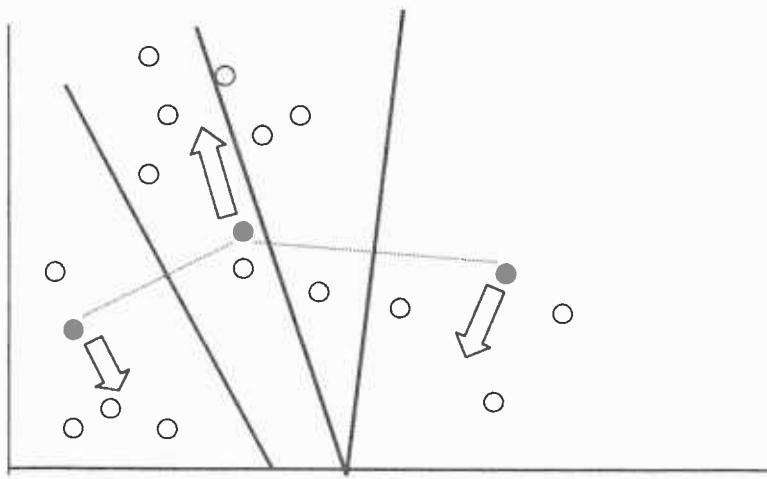
2.3 K-MEANS

Η μέθοδος K-Means αποτελεί μία από τις πιο συχνά χρησιμοποιούμενες μεθόδους clustering [Berry 97]. Ανήκει στην κατηγορία του partitional clustering, βασίζεται δηλαδή στην άμεση αποσύνθεση του συνόλου των δεδομένων σε ένα σύνολο ασυσχέτιστων clusters. Η αντικειμενική συνάρτηση την οποία προσπαθεί να ελαχιστοποιήσει ο αλγόριθμος είναι η μέση τετραγωνική απόσταση των δεδομένων από τα πλησιέστερα κέντρα των clusters[KS 97],

$$E_K = \sum_k \left\| x_k - m_{c(x_k)} \right\|^2 \quad (\text{Εξισ. 2.2})$$

όπου $c(x_k)$ είναι ο δείκτης του κέντρου το οποίο είναι πλησιέστερα στο x_k .

Ο βασικός αλγόριθμος για να ελαχιστοποιήσει την αντικειμενική συνάρτηση αρχίζει θεωρώντας ένα σύνολο από k σημεία-δεδομένα ως τα κέντρα των k clusters (σχήμα 2.4). Αν η σειρά των δεδομένων δεν έχει κάποια ιδιαίτερη σημασία, τότε παίρνουμε τα πρώτα k records. Άλλιώς επιλέγουμε σημεία αντιπροσωπευτικά για τα clusters μας τα οποία απέχουν μεταξύ τους. Καθένα από τα κέντρα αντιπροσωπεύει ένα cluster. Στο δεύτερο βήμα, κάθε σημείο αντιστοιχείται στο cluster του οποίου το κέντρο βρίσκεται πιο κοντά.



Σχήμα 2.4. Αρχικοποίηση K-means



Στη συνέχεια υπολογίζονται τα νέα κέντρα των clusters με χρήση του μέσου όρου των σημείων τους. Για άλλη μια φορά αντιστοιχείται κάθε σημείο στο cluster του οποίου το κέντρο είναι πλησιέστερο. Η διαδικασία επαναλαμβάνεται συνεχώς έως ότου τα όρια των clusters παύουν να μεταβάλλονται, ή η συνάρτηση E δεν μεταβάλλεται σημαντικά.

Ο αλγόριθμος K-Means χρησιμοποιεί σταθερό και δοσμένο εξ' αρχής αριθμό clusters που θα δημιουργηθούν (όσα και τα κέντρα).

2.3.1 ΑΛΓΟΡΙΘΜΟΣ K-MEANS

Στην παράγραφο αυτή περιγράφονται με την μορφή ψευδοκώδικα τα βασικά βήματα του αλγορίθμου K-Means. Ο αλγόριθμος ξεκινά καθορίζοντας με τυχαίο τρόπο c κέντρα που θα αντιπροσωπεύουν τα c clusters. Στην συνέχεια προσδιορίζεται η απόσταση κάθε στοιχείου του συνόλου δεδομένων από το κέντρο κάθε cluster και κάθε στοιχείο τοποθετείται στο cluster από το οποίο απέχει λιγότερο. Τα κέντρα των νέων clusters υπολογίζονται σαν ο μέσος όρος των στοιχείων που ανήκουν μέχρι στιγμής σε κάθε cluster. Η διαδικασία επαναλαμβάνεται μέχρις ότου τα clusters να σταματήσουν να μεταβάλλονται. Αυτό σημαίνει ότι η απόκλιση μεταξύ των κέντρων των clusters που προέκυψαν τελευταία από αυτά της προηγούμενης επανάληψης είναι κοντά στο μηδέν (τα κέντρα ταυτίζονται).

Τα βήματα του αλγορίθμου σε μορφή ψευδοκώδικα είναι τα εξής:

1. Εύρεση των αρχικών κέντρων, v_i $i=1,2,\dots,c$, για τα c clusters.

Για κάθε επανάληψη $r = 1, \dots, r_{max}$

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

$$d_{ki} = (x_k - v_i)^2, k=1,2,\dots,n \quad i=1,2,\dots,c$$

3. Κάθε στοιχείο x_k αντιστοιχίζεται στο cluster για το οποίο ισχύει

$$\min_{k,i} d_{ik}, \forall i, k$$

4. Υπολογισμός των νέων κέντρων των clusters

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

όπου n_i ο αριθμός των στοιχείων που ανήκουν στο i cluster μέχρι στιγμής.

5.

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

stop

else

$r = r + 1$, goto2.

2.3.2 ΠΑΡΑΛΛΑΓΕΣ K-MEANS

Ο K-Means όπως προαναφέρθηκε αποτελεί μία ευρέως αποδεκτή τεχνική clustering, η οποία έχει χρησιμοποιηθεί αποτελεσματικά για clustering σε διάφορα πεδία ορισμού. Ωστόσο, ο αλγόριθμος K-Means δεν είναι μοναδική τεχνική, αλλά έχει διάφορες εκδόσεις και πλήθος παραλλαγών. Οι παραλλαγές αυτές διαφέρουν κυρίως στον τρόπο επιλογής των αρχικών k μέσων (κέντρων) των clusters, στον υπολογισμό της ομοιότητας και στη στρατηγική που χρησιμοποιούν για τον υπολογισμό των μέσων των clusters. Ορισμένες χαρακτηριστικές παραλλαγές του K-Means είναι:

- ο αλγόριθμος *ISODATA* ο οποίος περιλαμβάνει μία διαδικασία για αναζήτηση του καλύτερου αριθμού clusters με βάση κάποιο κόστος εκτέλεσης,
- ο *fuzzy K-Means* ο οποίος επεκτείνει τον κλασικό K-Means αλγόριθμο χρησιμοποιώντας την θεωρία της ασαφής λογικής και
- ο *SAS PROC FASTCLUS*, ο οποίος ελέγχει την διαδικασία clustering υιοθετώντας δύο ακόμα παραμέτρους, την *max_rad* και *min_size*. Η πρώτη παράμετρος ελέγχει τον ελάχιστο αριθμό στοιχείων που μπορεί να έχει κάθε cluster ενώ η δεύτερη καθορίζει ότι η απόσταση κάθε στοιχείου από το κέντρο του cluster δεν πρέπει να είναι μεγαλύτερη του *max_rad*.

Επιπρόσθετα, διάφορα στατιστικά πακέτα όπως το SAS, SPSS και BMPD που χρησιμοποιούν τον K-Means υιοθετούν την δική τους έκδοση το καθένα για τον αλγόριθμο.

2.3.3 ΕΚΛΕΠΤΥΝΣΗ ΑΡΧΙΚΩΝ ΣΗΜΕΙΩΝ ΓΙΑ ΤΟ K-MEANS CLUSTERING

Οι περισσότερες από τις πρακτικές προσεγγίσεις στο clustering χρησιμοποιούν μία επαναληπτική διαδικασία η οποία αποσκοπεί στην σύγκλιση ενός από τα τοπικά ελάχιστα. Η επαναληπτικές αυτές τεχνικές επηρεάζονται σημαντικά από τις αρχικές συνθήκες εκκίνησης του αλγορίθμου clustering. Συνεπώς, θα ήταν σκόπιμο να βρεθεί κάποια διαδικασία για υπολογισμό μίας εκλεπτυσμένης συνθήκης εκκίνησης από μία δεδομένη αρχική η οποία θα βασίζεται σε μία αποδοτική τεχνική για την εκτίμηση των μορφών διασποράς. Μία εκλεπτυσμένη συνθήκη εκκίνησης επιτρέπει σε έναν επαναληπτικό αλγόριθμο να συγκλίνει σε ένα "καλύτερο" τοπικό ελάχιστο.

Μία λύση στο πρόβλημα αρχικοποίησης clustering είναι η παραμετροποίηση κάθε μοντέλου cluster. Αυτή η παραμετροποίηση μπορεί να εκτελεστεί καθορίζοντας τα μέγιστα της συνάρτησης πυκνότητας πιθανότητας των δεδομένων και τοποθετώντας ένα κέντρο cluster σε κάθε μέγιστο. Μία άλλη προσέγγιση είναι να εκτιμήσουμε την πυκνότητα και να προσπαθήσουμε να βρούμε τα μέγιστα της εκτιμούμενης συνάρτησης πυκνότητας. Η εκτίμηση όμως της πυκνότητας σε πολυδιάστατα δεδομένα είναι δύσκολη διαδικασία.

Μία άλλη μέθοδος για την αντιμετώπιση του προβλήματος της αρχικοποίησης ενός αλγορίθμου clustering προτάθηκε από τους Bradley και Fayyad [BF98]. Η μέθοδος βασίζεται στην διαδικασία εκλέπτυνσης των αρχικών σημείων σε σημεία που πιθανά να προσεγγίζουν κάποιο μέγιστο. Μπορεί να εφαρμοστεί σε μία μεγάλη ποικιλία αλγορίθμων clustering, τόσο για διακριτά όσο και για συνεχή δεδομένα, στην συνέχεια όμως η παρουσίαση μας θα επικεντρωθεί στον αλγόριθμο clustering K-means. Ο λόγος που επικεντρωνόμαστε στον K-Means είναι ότι: 1) είναι μία από τις πιο συνήθεις τεχνικές clustering, στην οποία χρησιμοποιείται ένα ευρύ σύνολο εφαρμογών, 2) ανεξάρτητα από το ποιος αλγόριθμος clustering χρησιμοποιείται, ο K-Means χρησιμοποιείται από την μέθοδο εύρεσης εκλεπτυσμένης αρχικοποίησης.

2.3.3.1 ΑΛΓΟΡΙΘΜΟΣ ΕΚΛΕΠΤΥΝΣΗΣ

Ο αλγόριθμος εκλέπτυνσης αρχικά επιλέγει J μικρά τυχαία υποδείγματα από το σύνολο των δεδομένων, S_i , $i=1, 2, \dots, J$. Στα υποδείγματα αυτά εφαρμόζεται clustering μέσω του αλγορίθμου K-Means με την υπόθεση ότι τα κενά clusters στο τέλος θα έχουν τα αρχικά τους κέντρα επανακαθορισμένα και στα υποδείγματα θα έχει γίνει re-clustering. Τα σύνολα CM_i , $i=1, 2, \dots, J$ είναι εκείνες οι λύσεις clustering που προκύπτουν από τα υποδείγματα και οι οποίες δημιουργούν το σύνολο CM . Στο σύνολο CM εφαρμόζεται στην συνέχεια clustering διαμέσου του K-Means λαμβάνοντας ως αρχικοποίηση το CM_i και δημιουργώντας έτσι μία λύση FM_i . Το εκλεπτυσμένο αρχικό σημείο επιλέγεται σαν το σύνολο FM_i το οποίο έχει ελάχιστη απόκλιση από το σύνολο CM .

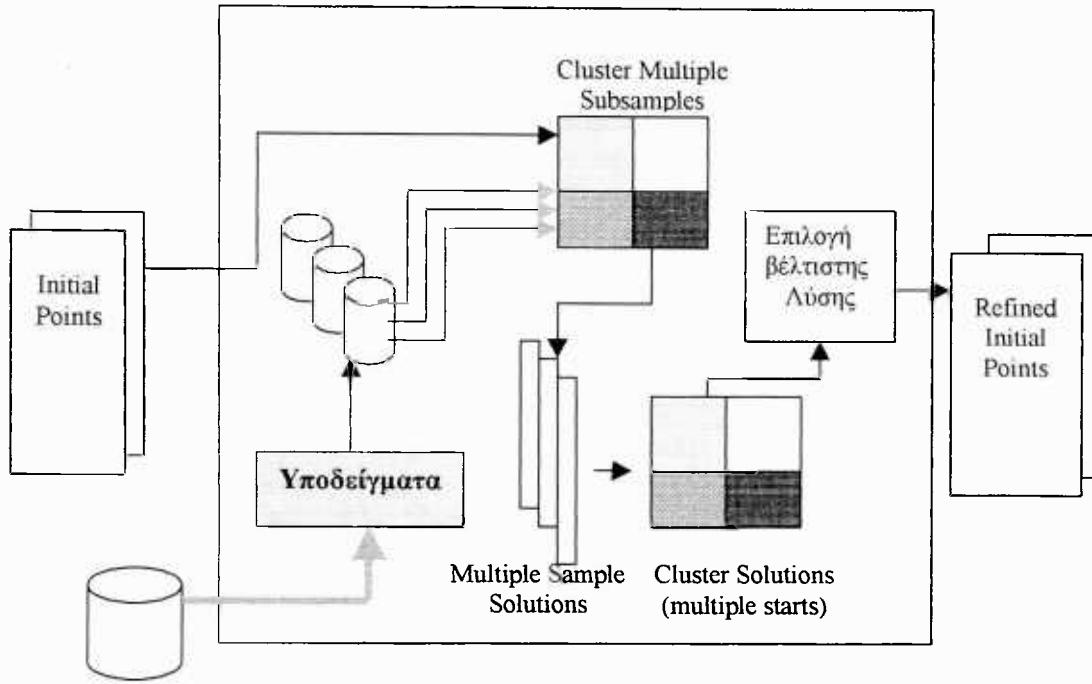
Το clustering του CM είναι μία εξομάλυνση πάνω στο CM_i για να αποφύγουμε τις επιπτώσεις από τους outliers οι οποίοι περιλαμβάνονται στα υποδείγματα S_i . Ο αλγόριθμος εκλέπτυνσης λαμβάνει σαν είσοδο: SP (αρχικά σημεία εκκίνησης), $Data$, K και J (ο αριθμός μικρών υποδειγμάτων που θα ληφθούν από τα δεδομένα).

Algorithm Refine(SP, Data, K, J)

0. CM = \emptyset
1. For i=1,...,J
 - a. Let S_i be a small random subsample of *Data*
 - b. Let $CM_i = \text{KmeansMod}(SP, S_i, K)$
 - c. $CM = CM \cup CM_i$
2. FMS = \emptyset
3. For i = 1, 2, ..., J
 - a. Let $FM_i = \text{Kmeans}(CM_i, CM, K)$
 - b. Let $FMS = FMS \cup FM_i$
4. Let $FM = \text{ArgMin}_{FM_i} \{\text{Distortion}(FM_i, CM)\}$
5. Return (FM)

Οι βασικές συναρτήσεις που χρησιμοποιεί ο αλγόριθμος εκλέπτυνσης είναι: KMeans(), KMeansMod() και Distortion(). Η Kmeans() είναι μία απλή κλήση του κλασικού αλγορίθμου K-Means ο οποίος λαμβάνει ως παραμέτρους εισόδου ένα αρχικό σημείο εκκίνησης, ένα σύνολο δεδομένων και τον αριθμό των clusters, ενώ επιστρέφει το σύνολο των εκτιμούμενων κέντρων για τα K clusters. Ο KMeansMod λαμβάνει τις ίδιες παραμέτρους όπως ο K-Means και εκτελεί την ίδια επαναληπτική διαδικασία όπως ο κλασικός αλγόριθμος K-Means με την εξής όμως τροποποίηση: Εάν κάποιο από τα K clusters δεν έχει στοιχεία από το σύνολο δεδομένων τα οποία να ανήκουν σε αυτό, οι αντίστοιχες αρχικές εκτιμήσεις των κέντρων των κενών clusters ανατίθενται στα στοιχεία τα οποία είναι μακρύτερα από τα κέντρα του cluster που έχουν τοποθετηθεί, ενώ καλείται πάλι ο κλασικός K-Means αλγόριθμος για τα νέα κέντρα που έχουν καθορισθεί.

Η Distortion λαμβάνει τα K εκτιμούμενα κέντρα και τα σύνολα δεδομένων και υπολογίζει το άθροισμα των τετραγώνων των αποστάσεων κάθε σημείου από το κοντινότερο μέσο(κέντρο). Μετρά δηλαδή, το βαθμό προσαρμογής ενός συνόλου clusters σε ένα σύνολο δεδομένων. Ο αλγόριθμος K-Means τερματίζεται σε μία λύση η οποία είναι τοπικά βέλτιστη για αυτή την συνάρτηση απόκλισης(distortion). Η διαδικασία εκλέπτυνσης παρουσιάζεται στο σχήμα 2.5.



Σχήμα 2.5. Διαδικασία Εκλέπτυνσης Αρχικών Σημείων

2.4 PAM (PARTITIONING AROUND MEDOIDS)

Ο PAM αναπτύχθηκε από τους Kaufman και Rousseeuw και αποτελεί μία από τις πιο γνωστές k-medoids μεθόδους clustering[KOP97].

Προκειμένου να βρεθούν k clusters με την προσέγγιση PAM καθορίζεται ένα αντικείμενο αντιπρόσωπος για κάθε cluster. Αυτό το αντικείμενο αντιπρόσωπος καλείται medoid και είναι το αντικείμενο που βρίσκεται πιο κοντά στο κέντρο του cluster. Μόλις επιλεχθούν τα medoids, κάθε μη επιλεγμένο αντικείμενο ομαδοποιείται με το medoid με το οποίο μοιάζει περισσότερο. Ειδικότερα, εάν O_j είναι ένα μη επιλεγμένο αντικείμενο και O_i είναι ένα(επιλεγμένο) medoid, λέμε ότι το O_j ανήκει στο cluster που αντιπροσωπεύεται από το O_i , εάν $d(O_j, O_i) = \min_{O_c} d(O_j, O_c)$, όπου η έκφραση \min_{O_c} δηλώνει το ελάχιστο μεταξύ όλων των medoids O_c και το $d(O_a, O_b)$ δηλώνει την απόσταση μεταξύ των αντικειμένων $d(O_a, O_b)$. Όλες οι τιμές των αποστάσεων μεταξύ των αντικειμένων δίνονται σαν είσοδο στο PAM. Η ποιότητα του clustering μετράται με βάση την μέση διαφοροποίηση ανάμεσα σε ένα αντικείμενο και στο medoid του cluster που ανήκει.

2.4.1 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ

Έστω ότι έχουμε ένα σύνολο ή αντικειμένων το οποίο θέλουμε να διαιρέσουμε σε k clusters με βάση τον αλγόριθμο PAM. Ο PAM ξεκινά με την εύρεση των k medoids, επιλέγοντας αυθαίρετα k αντικείμενα. Στην συνέχεια σε κάθε βήμα, εκτελείται μία ανταλλαγή ανάμεσα σε ένα επιλεγμένο αντικείμενο O_i και σε ένα μη επιλεγμένο O_h

αντικείμενο μέχρις ότου η ανταλλαγή αυτή να οδηγήσει στην βελτίωση της ποιότητας του clustering. Ειδικότερα, για να υπολογίσουμε το αποτέλεσμα μίας τέτοιας ανταλλαγής ανάμεσα στα αντικείμενα O_i και O_h , ο PAM υπολογίζει το κόστος C_{jih} για όλα τα μη επιλεγμένα αντικείμενα O_j . Ο ορισμός του κόστους γίνεται σύμφωνα με τις τέσσερις ακόλουθες εκφράσεις ανάλογα με τις περιπτώσεις των αντικείμενων O_j [RJ94]:

- Το O_j ανήκει στο cluster που αντιπροσωπεύεται από το O_i . Επιπρόσθετα, ας υποθέσουμε ότι το O_j είναι πιο κοντά στο αντικείμενο $O_{j,2}$ από ότι με το αντικείμενο O_h , δηλαδή $d(O_j, O_h) \geq d(O_j, O_{j,2})$, όπου το $O_{j,2}$ είναι το δεύτερο medoid που είναι πιο κοντά στο O_j . Ετσι εάν αντικαταστήσουμε το O_i με το O_h σαν medoid, το O_j θα ανήκει στο cluster που αντιπροσωπεύεται από το $O_{j,2}$. Το κόστος της ανταλλαγής θα δίνεται από την εξίσωση:

$$C_{jih} = d(O_j, O_{j,2}) - d(O_j, O_i) \quad (\text{Εξισ. 2.3})$$

Η ισότητα αυτή δίνει πάντα μη αρνητική τιμή για το κόστος, υποδηλώνοντας ότι το κόστος που προκύπτει για την αντικατάσταση του O_i με το O_h δεν είναι αρνητικό.

- Το O_j ανήκει στο cluster που αντιπροσωπεύεται από το O_i . Άλλα αυτή τη φορά, το O_j είναι λιγότερο κοντά στο αντικείμενο $O_{j,2}$ σε σχέση με το αντικείμενο O_h , δηλαδή $d(O_j, O_h) \leq d(O_j, O_{j,2})$. Ετσι εάν αντικαταστήσουμε το O_i με το O_h σαν medoid, το O_j θα ανήκει στο cluster που αντιπροσωπεύεται από το O_h . Το κόστος της ανταλλαγής θα δίνεται από την εξίσωση:

$$C_{jih} = d(O_j, O_h) - d(O_j, O_i) \quad (\text{Εξισ. 2.4})$$

Η τιμή του κόστους μπορεί να είναι αρνητική ή και θετική ανάλογα με το αν το αντικείμενο O_j προσεγγίζει περισσότερο το O_i ή το O_h .

- Υποθέτουμε ότι το O_j ανήκει σε ένα cluster διαφορετικό από αυτό που αντιπροσωπεύεται από το O_i . Έστω ότι $O_{j,2}$ είναι ο αντιπρόσωπος του cluster. Επίσης, θεωρούμε ότι το O_j προσεγγίζει περισσότερο το O_i από ότι το O_h . Ετσι εάν αντικαταστήσουμε το O_i με το O_h , το O_j θα παραμείνει στο cluster που αντιπροσωπεύεται από το $O_{j,2}$. Το κόστος της ανταλλαγής θα δίνεται από την εξίσωση:

$$C_{jih} = 0 \quad (\text{Εξισ. 2.5})$$

- Το O_j ανήκει στο cluster που αντιπροσωπεύεται από το $O_{j,2}$. Άλλα το O_j είναι λιγότερο κοντά στο αντικείμενο $O_{j,2}$ από ότι με το αντικείμενο O_h . Ετσι εάν αντικαταστήσουμε το O_i με το O_h σαν medoid, το O_j θα μετακινηθεί στο cluster που αντιπροσωπεύεται από το O_h . Το κόστος της ανταλλαγής θα δίνεται από την εξίσωση:

$$C_{jih} = d(O_j, O_h) - d(O_j, O_{j,2}) \quad (\text{Εξισ. 2.6})$$

Το κόστος στην περίπτωση αυτή θα είναι πάντοτε αρνητικό.

Συνδυάζοντας τις τέσσερις παραπάνω περιπτώσεις το συνολικό κόστος αντικατάστασης του αντικειμένου O_i με το O_h δίνεται από την εξίσωση:

$$TC_{ih} = \sum_j C_{jih}. \quad (\text{Εξισ. 2.7})$$

2.4.2 ΑΛΓΟΡΙΘΜΟΣ PAM

Τα βασικά βήματα του PAM σε μορφή ψευδοκώδικα είναι:

1. Επιλογή αυθαίρετα k αντιπροσώπων για τα clusters
2. Υπολογισμός του συνολικού κόστους TC_{ih} για όλα τα ζεύγη των αντικειμένων O_i , O_h όπου το O_i είναι το τρέχον επιλεγμένο αντικείμενο και το O_h είναι ένα μη επιλεγμένο αντικείμενο.
3. Επιλέγουμε το ζεύγος O_i , O_h το οποίο αντιστοιχεί στο $\min_{O_i, O_h} TC_{ih}$. Εάν το συνολικό κόστος είναι αρνητικό αντικαθιστούμε το O_i με το O_h και επιστρέφουμε στο βήμα 2.
4. Διαφορετικά, για κάθε μη επιλεγμένο αντικείμενο, βρίσκουμε το αντικείμενο αντιπρόσωπο που προσεγγίζει περισσότερο. Halt

Ο PAM δουλεύει ικανοποιητικά για μικρά σύνολα δεδομένων (π.χ. 10 αντικείμενα σε 5 clusters). Άλλα αποδεικνύεται μη αποδοτικός για σύνολα δεδομένων μεσαίου και μεγάλου μεγέθους λόγω της μεγάλης πολυπλοκότητας του. Στο βήμα 2 και 3 υπάρχουν συνολικά k(n-k) ζεύγη αντικειμένων O_i , O_h . Για κάθε ζεύγος ο υπολογισμός του συνολικού κόστους απαιτεί την εξέταση (n-k) μη επιλεγμένων αντικειμένων. Συνεπώς τα βήματα 2 και 3 έχουν για μία επανάληψη πολυπλοκότητα $O(k(n-k)^2)$. Είναι λοιπόν φανερό ότι ο PAM έχει μεγάλο κόστος για μεγάλες τιμές του n και k.

2.5 CLARA (CLUSTERING LARGE APPLICATIONS)

Ο αλγόριθμος CLARA σχεδιάστηκε από τους Kaufman και Rousseeuw [KOP97], προκειμένου να διαχειριστούν μεγάλα σύνολα δεδομένων. Η βασική διαφορά ανάμεσα στον CLARA και τον PAM είναι ότι ο πρώτος βασίζεται στην δειγματοποίηση. Ο CLARA αντίθετα με τον PAM δεν βρίσκει αντικείμενα αντιπροσώπους για ολόκληρο το σύνολο δεδομένων, αλλά λαμβάνει με τυχαίο τρόπο ένα δείγμα του συνόλου των δεδομένων, εφαρμόζει στο δείγμα τον PAM και βρίσκει τα medoids του δείγματος. Η ιδέα είναι ότι εάν το δείγμα είναι σχεδιασμένο με εντελώς τυχαίο τρόπο, τότε αναπαριστά ολόκληρο το σύνολο ικανοποιητικά και για το λόγο αυτό τα αντικείμενα αντιπρόσωποι (medoids) του δείγματος θα προσεγγίζουν τα medoids ολόκληρου του συνόλου δεδομένων. Ο αλγόριθμος σχεδιάζει πολλαπλά δείγματα και εξάγει το καλύτερο clustering από τα δείγματα αυτά. Τα πειράματα έχουν αποδείξει ότι 5 δείγματα μεγέθους $40 + 2k$ δίνουν ικανοποιητικά αποτελέσματα.

2.5.1 ΑΛΓΟΡΙΘΜΟΣ CLARA

Στην συνέχεια αναφέρονται υπό μορφή ψευδοκώδικα τα βασικά βήματα του αλγορίθμου [RJ94]:

1. Για $i = 1$ to 5 , επαναλαμβάνουμε τα ακόλουθα βήματα:
2. Σχεδιάζουμε ένα δείγμα $40+2k$ αντικειμένων με τυχαίο τρόπο από το σύνολο των δεδομένων και καλούμε τον αλγόριθμο PAM για να βρούμε τους k αντιπροσώπους για τα clusters.
3. Για κάθε αντικείμενο O_j στο σύνολο δεδομένων, καθορίζουμε πιο από τα k medoids προσεγγίζει περισσότερο το O_j .
4. Υπολογίζουμε την συνολική ανομοιότητα για το clustering που λαμβάνεται από το προηγούμενο βήμα. Εάν αυτή η τιμή είναι μικρότερη από το τρέχον ελάχιστο, χρησιμοποιούμε αυτή την τιμή του ελαχίστου σαν τρέχον ελάχιστο και διατηρούμε τα k medoids που βρήκαμε στο βήμα 2 σαν το καλύτερο σύνολο των medoids που έχουμε μέχρι στιγμής.
5. Επιστρέφουμε στο βήμα 1 και ξεκινάμε με την επόμενη επανάληψη.

Ο CLARA εφαρμόζει τον PAM μόνο σε δείγματα και έτσι σε κάθε επανάληψη η πολυπλοκότητα είναι $O(k(40+k)^2 + k(n-k))$. Συνεπώς ο CLARA είναι πιο αποδοτικός από τον PAM για μεγάλες τιμές του n (μεγάλα σύνολα δεδομένων).

2.6 CLARANS (CLUSTERING LARGE APPLICATIONS BASED ON RANDOMIZED SEARCH)

Ο αλγόριθμος CLARANS [KOP97] προσπαθεί να συνδυάσει τους αλγορίθμους PAM και CLARA εκτελώντας κάθε φορά αναζήτηση μόνο σε ένα υποσύνολο του συνόλου των δεδομένων ενώ δεν περιορίζεται σε κάποιο δείγμα σε μία δεδομένη στιγμή. Ενώ ο CLARA έχει ένα καθορισμένο δείγμα σε κάθε βήμα της αναζήτησης, ο CLARANS σχεδιάζει ένα δείγμα με τυχαίο τρόπο σε κάθε βήμα της αναζήτησης. Η διαδικασία clustering μπορεί να αναπαρασταθεί σαν ένα γράφημα όπου κάθε κόμβος είναι μια πιθανή λύση δηλαδή ένα σύνολο από k medoids. Το clustering που λαμβάνεται μετά την αντικατάσταση ενός medoid καλείται γείτονας (*neighbor*) του τρέχοντος clustering. Ο αριθμός των γειτόνων που μπορούν να δοκιμαστούν τυχαία περιορίζεται από μία παράμετρο που καλείται *maxneighbor*. Εάν βρεθεί ένας καλύτερος γείτονας ο CLARANS μετακινείται στον κόμβο του γείτονα και η διαδικασία ξεκινάει πάλι από τον κόμβο αυτό, ενώ σε διαφορετική περίπτωση το τρέχον clustering παράγει ένα τοπικό βέλτιστο.

Εάν βρεθεί ένα τοπικό βέλτιστο, ο αλγόριθμος CLARANS αρχίζει με ένα νέο τυχαία επιλεγμένο κόμβο για την αναζήτηση ενός νέου τοπικού βέλτιστου. Ο αριθμός των τοπικών βέλτιστων που θα αναζητηθούν καθορίζεται επίσης από μία παράμετρο που καλείται *numlocal*. Ο αλγόριθμος αυτός έχει αποδειχθεί πιο αποδοτικός σε σχέση με τον CLARA και τον PAM και η υπολογιστική πολυπλοκότητα του για κάθε επανάληψη εξαρτάται από τον αριθμό των αντικειμένων, $O(n^2)$. Ωστόσο, λόγω της

τυχαίας προσέγγισης του CLARANS, για μεγάλες τιμές του N , η ποιότητα των αποτελεσμάτων δεν είναι εγγυημένη.

2.6.1 ΑΛΓΟΡΙΘΜΟΣ CLARANS

Τα βασικά βήματα του αλγορίθμου μπορούν να συνοψιστούν στα εξής[RJ94]:

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

Όσο μεγαλύτερος είναι ο αριθμός των γειτόνων(*maxneighbor*) που εξετάζονται τόσο ο αλγόριθμος CLARANS προσεγγίζει τον PAM και η αναζήτηση για την εύρεση του τοπικού ελαχίστου έχει μεγαλύτερη διάρκεια. Άλλα η ποιότητα ενός τέτοιου τοπικού ελαχίστου είναι μεγαλύτερη και έτσι λιγότερα τοπικά ελαχιστά χρειάζεται να εξεταστούν.

2.7 CUBE: ΙΕΡΑΡΧΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ

Ο CUBE είναι ένας ιεραρχικός αλγόριθμος clustering του οποίου τα βασικά χαρακτηριστικά είναι ότι [GRK98]:

- ◆ μπορεί να αναγνωρίζει clusters αυθαίρετων σχημάτων (π.χ. ελλειψοειδή)
- ◆ είναι εύρωστος στην παρουσία των outliers
- ◆ οι απαιτήσεις του σε χώρο αποθήκευσης είναι γραμμική συνάρτηση του αριθμού των στοιχείων εισόδου και η χρονική πολυπλοκότητα του είναι $O(n^2)$ για δεδομένα μικρών διαστάσεων.

Ο αλγόριθμος μπορεί να εφαρμοστεί αποδοτικά και για clustering μεγάλων βάσεων δεδομένων συνδυάζοντας τεχνικές τυχαίας δειγματοποίησης (sampling) και τμηματοποίησης (partitioning). Επομένως, τα δεδομένα που εισάγονται στον αλγόριθμο μπορεί να είναι ένα δείγμα που δημιουργήθηκε τυχαία από τα αυθεντικό σύνολο δεδομένων ή ένα υποσύνολο αυτού του δείγματος εάν εφαρμοστεί η τμηματοποίηση.

2.7.1 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ

Ο αλγόριθμος αρχίζει λαμβάνοντας κάθε σημείο εισόδου σαν ζεχωριστό cluster και σε κάθε βήμα που ακολουθεί συγχωνεύει τα πλησιέστερα ζευγάρια clusters. Προκειμένου να υπολογίσουμε την απόσταση μεταξύ των clusters, αποθηκεύονται για κάθε cluster c αντιπρόσωποι (representatives). Οι αντιπρόσωποι αυτοί καθορίζονται επιλέγοντας αρχικά τα c πιο διάσπαρτα σημεία μέσα σε ένα cluster και στην συνέχεια μετακινούμε τα σημεία προς τον μέσο του cluster κατά ένα ποσοστό α . Η απόσταση μεταξύ των clusters είναι η απόσταση μεταξύ των πιο κοντινών αντιπροσώπων δύο clusters. Έτσι μόνο τα σημεία αντιπρόσωποι ενός cluster χρησιμοποιούνται για να υπολογίσουμε την απόσταση από ένα άλλο cluster.

Οι c αντιπρόσωποι προσπαθούν να προσδιορίσουν το φυσικό σχήμα και την γεωμετρία του cluster. Επιπρόσθετα, μετακινώντας τα διάσπαρτα σημεία προς το μέσο κατά ένα ποσοστό α απομακρύνουμε τις ανωμαλίες και μετριάζουμε τις επιδράσεις των outliers. Ο λόγος που γίνεται αυτό είναι ότι οι outliers βρίσκονται τυπικά μακριά από το κέντρο του cluster και έτσι η συρρίκνωση θα κάνει τους outliers να κινηθούν περισσότερο προς το κέντρο ενώ οι αντιπρόσωποι που θα απομείνουν θα υποστούν ελάχιστη μετακίνηση. Οι μεγάλες μετακινήσεις στους outliers θα μειώσουν την δυνατότητα τους να προκαλέσουν συγχώνευση λάθος clusters. Η παράμετρος α μπορεί επίσης να χρησιμοποιηθεί και για τον έλεγχο του σχήματος των clusters. Μία μικρή τιμή για το α συρρικνώνει τα διάσπαρτα σημεία πολύ λίγο και έτσι ενισχύει την ύπαρξη clusters που δεν είναι σφαιρικά. Αντίθετα, μεγάλες τιμές για το α έχουν σαν αποτέλεσμα την δημιουργία συμπαγών clusters καθώς τα διάσπαρτα σημεία τοποθετούνται πιο κοντά στο μέσο των clusters.

2.7.2 ΑΛΓΟΡΙΘΜΟΣ CUBE

Στην παράγραφο αυτή περιγράφουμε με λεπτομέρεια τον αλγόριθμο clustering CUBE. Οι παράμετροι εισόδου στον αλγόριθμο μας είναι το σύνολο των δεδομένων S το οποίο περιέχει η σημεία d -διαστάσεων και τον επιθυμητό αριθμό των clusters k . Όπως αναφέρθηκε και προηγούμενα, ξεκινώντας με έναν αριθμό στοιχείων τα οποία λαμβάνονται σαν ζεχωριστά clusters, συγχωνεύονται σε κάθε βήμα τα ζεύγη των clusters που είναι πιο κοντά το ένα στο άλλο. Η διαδικασία συνεχίζεται μέχρι να παραμείνουν k clusters.

```

procedure cluster(S, k)
begin
    T := build_kd_tree(S)
    Q := build_heap(S)
    while size(Q) > k do {
        u := extract_min(Q)
        v := u.closest
        delete(Q, v)
        w := merge(u, v)
        delete_rep(T, u); delete_rep(T, v); insert_rep(T, w)
        w.closest := x //x είναι ένα τυχαίο στοιχείο στο cluster Q
        for each x ∈ Q do{
            if dist(w,x) < dist(w, w.closest)
                w.closest:=x
            if x.closest is either u or v {
                if dist(x, x.closest) < dist(x, w)
                    x.closest := closest_cluster(T, x, dist(x, w))
                else
                    x.closest := w
                relocate(Q, x)
            }
            else if dist(x, x.closest) > dist(x, w) {
                x.closest := w
                relocate(Q, x)
            }
        }
        insert(Q, w)
    }
end

```

Διαδικασία συγχώνευσης των clusters

```

procedure cluster(S, k)
begin
    w:=u ∪ v
    w.mean Q = (|u|u.mean + |v|v.mean) / |u|+|v|
    tmpSet := 0
    for i:=1 to c do {
        maxDist := 0
        foreach point p in cluster w do{
            if i=1
                minDist := dist(p, w.mean)
            else
                minDist := min{dist(p,q): q∈tmpSet}
            if (minDist ≥ maxDist){
                maxDist := minDist
                maxPoint := p
            }
        }
    }

```

```

tmpSet := tmpSet ∪ {maxPoint}
}
foreach point p in tmpSet do
    w.rep := w.rep ∪ {p + a*(w.mean - p)}
return w
end

```

Ο αλγόριθμος χρησιμοποιεί για κάθε cluster u τις δομές δεδομένων u.mean και u.rep οι οποίες περιέχουν το μέσο των στοιχείων και το σύνολο των c αντιπροσώπων για το cluster u, αντίστοιχα. Η απόσταση μεταξύ δύο στοιχείων p και q, $\text{dist}(p,q)$, εκφράζει την ομοιότητα μεταξύ των στοιχείων και μπορεί να είναι οποιαδήποτε μετρική ή μη μετρική συνάρτηση. Γενικά η απόσταση μεταξύ των clusters u, v είναι η ελάχιστη απόσταση μεταξύ των αντιπροσώπων της, δηλ.

$$\text{dist}(u,v) = \min \text{ dist}(p, q), \text{ όπου } p \in u.\text{rep} \text{ } q \in v.\text{rep}$$

Επίσης για κάθε cluster u διατηρούμε στην δομή u.closest το πλησιέστερο σε αυτό cluster. Επιπρόσθετα, ο αλγόριθμος διατηρεί μία στοίβα με τα clusters ταξινομημένα κατά την αύξουσα σειρά της απόστασης μεταξύ των u και u.closest. Τέλος χρησιμοποιεί την δομή δεδομένων k-d tree στην οποία αποθηκεύει τα στοιχεία αντιπροσώπους για κάθε cluster. Η δομή k-d tree είναι ένα δυαδικό δέντρο με την διαφορά ότι σε κάθε επίπεδο του δέντρου ελέγχεται μία διαφορετική τιμή κλειδί προκειμένου να καθοριστεί το κλαδί που θα ακολουθηθεί στην συνέχεια. Συνεπώς είναι μία κατάλληλη δομή για αποδοτική αποθήκευση και ανάκτηση πολυδιάστατων στοιχείων δεδομένων. Όταν ένα ζευγάρι clusters συγχωνεύεται, το k-d δέντρο χρησιμοποιείται για να προσδιορίσει το πλησιέστερο cluster για τα clusters που προηγούμενα είχαν σαν πλησιέστερο ένα από τα clusters που συγχωνεύτηκαν.

2.7.3 ΕΠΕΚΤΑΣΕΙΣ ΓΙΑ ΜΕΓΑΛΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ

Γενικά, οι iεραρχικοί αλγόριθμοι δεν είναι άμεσα εφαρμόσιμοι σε μεγάλα σύνολα δεδομένων εξαιτίας της υψηλής πολυπλοκότητας. Προκειμένου ο CUBE να χειριστεί μεγάλες βάσεις δεδομένων χρησιμοποιεί τον συνδυασμό τυχαίας δειγματοποίησης (sampling) και τμηματοποίησης (partitioning). Ένα τυχαίο δείγμα σχεδιάζεται από το σύνολο των δεδομένων με τρόπο ώστε να είναι αντιπροσωπευτικό του συνόλου. Επιλέγεται δηλαδή ο κατάλληλος αριθμός στοιχείων από το αρχικό σύνολο ώστε κατά την εφαρμογή του αλγορίθμου να μην παραληφθούν συγκεκριμένα clusters ή να προσδιοριστούν clusters τα οποία δεν ανταποκρίνονται στα πραγματικά.

Καθώς όμως ο διαχωρισμός μεταξύ των clusters μειώνεται και καθώς τα clusters γίνονται λιγότερο πυκνά, απαιτούνται δείγματα μεγάλου μεγέθους για να διακρίνουμε με επιτυχία τα clusters. Ωστόσο, καθώς το μέγεθος του συνόλου των δεδομένων εισόδου αυξάνεται, η υπολογιστική πολυπλοκότητα για τον αλγόριθμο clustering CURE αυξάνεται σημαντικά. Για το λόγο αυτό προτείνεται ένα απλό σχήμα τμηματοποίησης για να επιταχυνθεί η εκτέλεση του CURE όταν τα μεγέθη των δειγμάτων εισόδου είναι αρκετά μεγάλα. Το δείγμα του συνόλου των δεδομένων μας διαιρείται σε τμήματα στα οποία και εκτελείται ο αλγόριθμος clustering. Στην συνέχεια με βάση τα clusters που έχουν προσδιοριστεί στα τμήματα, εφαρμόζεται ο

αλγόριθμος για την εύρεση των clusters του συνόλου των δεδομένων. Η βασική ιδέα είναι να τμηματοποιήσουμε το δείγμα μας σε p τμήματα, καθένα μεγέθους n/p . Στην συνέχεια εφαρμόζουμε clustering σε κάθε τμήμα μέχρι ο αριθμός των clusters σε κάθε τμήμα να μειωθεί σε n/pq για κάποια σταθερά $q > 1$. Εναλλακτικά, μπορούμε να σταματήσουμε την συγχώνευση των clusters σε ένα τμήμα εάν η απόσταση μεταξύ των πλησιέστερων clusters που πρόκειται να συγχωνευτούν στο επόμενο βήμα ξεπερνά ένα συγκεκριμένο όριο. Έχοντας παράγει n/pq clusters για κάθε τμήμα, μπορούμε να εκτελέσουμε ένα δεύτερο clustering στα n/q clusters των τμημάτων.

2.8 DBSCAN

Ο DBSCAN είναι ένας αλγόριθμος clustering ο οποίος βασίζεται στην πυκνότητα. Η βασική ιδέα είναι ότι η περιοχή που εκτείνεται σε συγκεκριμένη ακτίνα(Eps) γύρω από κάθε αντικείμενο ενός cluster (γειτονιά αντικειμένου) θα πρέπει να περιέχει έναν ελάχιστο αριθμό από αντικείμενα.

Βασικές έννοιες αλγορίθμου

Στην συνέχεια θα αναφερθούμε εν συντομίᾳ στις κυριότερες έννοιες που αποτελούν την βάση του αλγορίθμου DBSCAN [EKSWX98]. Θεωρώντας λοιπόν ότι έχουμε ένα σύνολο αντικειμένων D στο οποίο η γειτονιά κάθε αντικειμένου εκτείνεται σε ακτίνα Eps γύρω από αυτό και ο ελάχιστος αριθμός στοιχείων που μπορεί να περιέχει είναι $MinPts$, μπορούμε να ορίσουμε τις εξής έννοιες:

- Ένα αντικείμενο p είναι **άμεσα πυκνά-προσεγγίσιμο** από ένα αντικείμενο q εάν
 1. το αντικείμενο ανήκει στο υποσύνολο των αντικειμένων που βρίσκονται στην γειτονιά του q
 2. ο αριθμός των αντικειμένων που περιέχονται στην γειτονιά του q είναι μεγαλύτερο από ένα όριο $MinPts$.
- Ένα αντικείμενο p είναι **πυκνά-προσεγγίσιμο** από ένα αντικείμενο q , $p >_D q$, εάν υπάρχει μια σειρά από αντικείμενα p_1, \dots, p_n , $p_1=q$, $p_n=p$ τέτοια ώστε το p_{i+1} να είναι άμεσα πυκνά-προσεγγίσιμο από το p_i .
- Ένα αντικείμενο p είναι **πυκνά-συνδεδεμένο** με ένα αντικείμενο q εάν υπάρχει ένα αντικείμενο o τέτοιο ώστε τόσο το p όσο και το q να είναι πυκνά-προσεγγίσιμα από το o .
- Ένα *cluster* C στο σύνολο των δεδομένων D είναι ένα μη-κενό υποσύνολο του D το οποίο ικανοποιεί τις ακόλουθες συνθήκες:
 1. Για κάθε $p, q \in C$: εάν $p \in C$ και $q >_D p$, τότε $q \in C$
 2. Για κάθε $p, q \in C$: το p είναι πυκνά-συνδεδεμένο με το q .
- Έστω ότι C_1, C_2, \dots, C_n είναι clusters του συνόλου δεδομένων D . Ορίζουμε ως **θόρυβο(noise)** το σύνολο των αντικειμένων στην βάση δεδομένων D τα οποία δεν ανήκουν σε κανένα cluster C_i .

Επίσης στο clustering τα αντικείμενα διακρίνονται στα αντικείμενα πυρήνα(*core objects*) τα οποία είναι αντικείμενα που ικανοποιούν την υπόθεση 2 του πρώτου ορισμού και στα αντικείμενα όχι-πυρήνα(*non-core objects*) τα οποία είναι όλα τα αντικείμενα που δεν ανήκουν στην κατηγορία των αντικειμένων πυρήνα.

2.8.1 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ

Ο αλγόριθμος αυτός απαιτεί από τον χρήστη να καθορίσει δύο παραμέτρους οι οποίες χρησιμοποιούνται για να ορίσουν την ελάχιστη πυκνότητα για το clustering. Οι παράμετροι αυτοί είναι: η ακτίνα *Eps* στην οποία θα εκτείνεται η γειτονιά κάθε στοιχείου του συνόλου των δεδομένων και ο ελάχιστος αριθμός των σημείων *MinPts* που μπορεί να υπάρχουν στην γειτονιά.

Ο αλγόριθμος αρχίζει με ένα τυχαίο στοιχείο p του συνόλου και ανακτά όλα τα στοιχεία τα οποία είναι πυκνά προσεγγίσιμα από το p . Εάν το στοιχείο p είναι ένα αντικείμενο πυρήνα, ο αλγόριθμος ορίζει ένα cluster. Εάν το στοιχείο p είναι ένα ακραίο στοιχείο, κανένα αντικείμενο δεν είναι πυκνά-προσεγγίσιμο από το p και το p συμπεριλαμβάνεται στο θόρυβο. Τότε, ο DBSCAN λαμβάνει το επόμενο στοιχείο της βάσης δεδομένων.

Ενώ ο αλγόριθμος μπορεί να βρει clusters με αυθαίρετα σχήματα, έχει αρκετά προβλήματα. Τα κυριότερα από αυτά είναι:

- Επηρεάζεται από τις τιμές των παραμέτρων *Eps* και *MinPts*, οι οποίες είναι δύσκολο να προσδιοριστούν
- Όπως όλοι οι ιεραρχικοί αλγόριθμοι πάσχει από το πρόβλημα της ευρωστίας καθώς στην περίπτωση που υπάρχει μία πυκνή σειρά σημείων που συνδέει δύο clusters ο DBSCAN μπορεί να τελειώσει συγχωνεύοντας τα δύο clusters.
- Δεν εφαρμόζει κάποια μορφή preclustering αλλά εφαρμόζεται απευθείας στο σύνολο των δεδομένων με αποτέλεσμα να καθίσταται ασύμφορος για μεγάλες βάσεις δεδομένων λόγω του κόστους I/O.
- Η χρήση δείγματος για να περιοριστεί το μέγεθος της εισόδου κατά την εφαρμογή των αλγορίθμων που βασίζονται στην πυκνότητα δεν είναι εφικτή. Ο λόγος είναι ότι ακόμα και αν το δείγμα είναι μεγάλο, μπορεί να υπάρχουν μεγάλες διακυμάνσεις στην πυκνότητα των σημείων μέσα σε κάθε cluster στο τυχαίο δείγμα.

2.8.2 ΑΛΓΟΡΙΘΜΟΣ DBSCAN

Ο αλγόριθμος clustering DBSCAN μπορεί να περιγραφεί με την μορφή ψευδοκώδικα ως εξής:

Algorithm DBSCAN (D, Eps, MinPts)

//Προϋπόθεση: Όλα τα αντικείμενα στο σύνολο δεδομένων D δεν έχουν τοποθετηθεί σε clusters.

FORALL objects o in D DO:

IF o δεν έχει ταξινομηθεί

Κάλεσε την συνάρτηση *expand_cluster* προκειμένου να κατασκευαστεί ένα cluster.

με ακτίνα Eps και ελάχιστο αριθμό στοιχείων MinPts το οποίο θα περιέχει το o.

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

```

Ανάκτηση της Eps-γειτονιάς  $N_{EPS}(o)$  του o;
if |  $N_{EPS}(o)$  | < MinPts //δηλ. ο δεν είναι ένα αντικείμενο πυρήνα
    Σημείωσε το o σαν θόρυβο, RETURN;
else //το o είναι αντικείμενο πυρήνα
    Επέλεξε ένα νέο cluster_id και σημείωσε όλα τα αντικείμενα στο  $N_{EPS}(o)$ 
    με το τρέχον cluster_id;
    ώθησε όλα τα αντικείμενα από το  $N_{EPS}(o)$ -{o} στην στοίβα seeds;
    while not seeds.empty() do
        currentObject:=seed.top();
        ανάκτηση της Eps-γειτονιάς του τρέχοντος αντικειμένου;
        if |  $N_{EPS}(currentObject)$  | ≥ MinPts
            επέλεξε όλα τα αντικείμενα  $N_{EPS}(currentObject)$  που δεν έχουν
            ταξινομηθεί ακόμα ή έχουν σημειωθεί ως θόρυβος,
            τοποθέτησε τα μη ταξινομημένα αντικείμενα στην στοίβα seeds και
            σημείωσε όλα τα αυτά αντικείμενα με το τρέχον cluster_id;
        seeds.pop();
    RETURN.

```

2.8.3 INCREMENTAL DBSCAN

To datawarehousing παρέχει πολλές ευκαιρίες για την εκτέλεση εργασιών data mining όπως clustering και classification. Τυπικά, οι ενημερώσεις συγκεντρώνονται και εφαρμόζονται περιοδικά στα δεδομένα που υπάρχουν στην αποθήκη δεδομένων(datawarehouse). Συνεπώς, όλα τα πρότυπα που εξάγονται από το warehouse με κάποιον αλγόριθμο data mining θα πρέπει επίσης να ενημερωθούν.

Ο DBSCAN απαιτεί μόνο μία συνάρτηση απόστασης και έτσι μπορεί να εφαρμοστεί σε οποιαδήποτε βάση δεδομένων που περιέχει δεδομένα από τον μετρικό χώρο. Επίσης λόγω του ότι ο αλγόριθμος βασίζεται στην πυκνότητα για την εκτέλεση του clustering έχει αποδειχθεί ότι μπορεί να υποστηρίξει αποτελεσματικά το data mining σε περιβάλλον datawarehouse. Η εισαγωγή ή διαγραφή ενός αντικειμένου επηρεάζει το τρέχον clustering μόνο στην γειτονιά αυτού του αντικειμένου.

Γενικά, έχει αποδειχθεί ότι κατά την εισαγωγή ή διαγραφή ενός αντικειμένου p, το σύνολο των αντικειμένων που επηρεάζονται(δηλ. αντικείμενα τα οποία μπορεί να μεταβάλλουν την συμμετοχή τους στα clusters), είναι τα αντικείμενα που ανήκουν στην γειτονιά του αντικειμένου p καθώς και όλα τα αντικείμενα που είναι πυκνά προσεγγίσιμα από ένα από αυτά τα αντικείμενα στο $D \cup \{p\}$. Αντίθετα, η συμμετοχή των άλλων αντικειμένων, που δεν ανήκουν στο σύνολο των επηρεαζόμενων αντικειμένων, στα clusters δεν μεταβάλλεται. Συνεπώς, με βάση τον αλγόριθμο DBSCAN μπορούν να σχεδιαστούν αποδοτικοί αλγόριθμοι ώστε να υποστηρίζουν τις εισαγωγές και διαγραφές στο clustering.

2.9 SCALING KAI WEIGHTING

Το *scaling* έχει να κάνει με διαφορετικές μεταβλητές που μετρώνται σε διαφορετικές μονάδες μέτρησης. Σκοπός μας είναι να φέρουμε όλες τις μεταβλητές σε συγκρίσιμα διαστήματα, ώστε μεταβολές μίας μεταβλητής να μην εμφανιστούν ως περισσότερο σημαντικές από ότι μεταβολές κάποιας άλλης μεταβλητής. Τρεις κοινοί τρόποι για scaling είναι [Bergy 97]:

1. Διαιρούμε κάθε μεταβλητή με τον μέσο όρο όλων των τιμών που λαμβάνει.
2. Διαιρούμε κάθε μεταβλητή με το εύρος του πεδίου τιμών της (διαφορά μεταξύ της μικρότερης και μεγαλύτερης τιμής που λαμβάνει η μεταβλητή) αφού αφαιρέσουμε την κατώτερη τιμή.
3. Αφαιρούμε τον μέσο όρο από κάθε μεταβλητή και μετά διαιρούμε με την τυπική απόκλιση. Η διαδικασία αυτή scaling "καλείται μετατροπή σε Z-τιμή".

Συνήθως η κλιμάκωση γίνεται μετατρέποντας όλες τις μεταβλητές και τις τιμές τους στο κοινό διάστημα 0 έως 1 ή -1 έως 1. Με τον τρόπο αυτό, τουλάχιστον οι αναλογίες των μεταβολών που παρατηρούνται στις μεταβλητές με διαφορετικές μονάδες μέτρησης είναι συμβατές. Επίσης θα ήταν χρήσιμο να αναφέρουμε, ότι η τρίτη μέθοδος scaling που αναφέραμε έχει αποδειχθεί σε πολλές περιπτώσεις να είναι ευαίσθητη στους outliers. Για να αντιμετωπιστεί το πρόβλημα αυτό προτείνεται η εξαίρεση του 1-5% των δεδομένων από τον υπολογισμό της μέσης τιμής και τυπικής απόκλισης[Raat93].

Το *weighting* υλοποιεί το διαφορετικό ενδιαφέρον που μπορεί να έχουμε για κάποιες μεταβλητές σε σχέση με τις άλλες. Δίνοντας διαφορετικά βάρη στις μεταβλητές, δίνουμε μεγαλύτερη σημασία στα μεγέθη της μεταβλητής με μεγαλύτερο βάρος. Για παράδειγμα, εάν μας ενδιαφέρουν περισσότερο οι άνθρωποι που έχουν παιδιά παρά ο αριθμός των πιστωτικών καρτών που διαθέτουν, τότε θα ήταν σκόπιμο στο αποτέλεσμα του clustering να μεροληπτήσουμε υπέρ του αριθμού των παιδιών, πολλαπλασιάζοντας το αντίστοιχο πεδίο με κάποιο βάρος υψηλότερο από ότι το πεδίο που αφορά των αριθμό των πιστωτικών καρτών. Η διαδικασία επιλογής βαρών (weights) είναι ένα από τα προβλήματα βελτιστοποίησης (optimization problems) και μπορεί να επιτευχθεί με χρήση γενετικών αλγορίθμων (genetic algorithms).

2.10 ΑΛΓΟΡΙΘΜΟΙ CLUSTERING ΓΙΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΜΕ ΛΕΚΤΙΚΕΣ ΤΙΜΕΣ

Οι περισσότεροι από τους κλασικούς αλγορίθμους clustering, μερικούς από τους οποίους περιγράψαμε στις προηγούμενες παραγράφους, περιορίζουν την εφαρμογή τους σε σύνολα δεδομένων με αριθμητικές τιμές. Σε πολλές όμως περιπτώσεις οι data mining εφαρμογές περιλαμβάνουν και μη αριθμητικά δεδομένα (categorical data). Η παραδοσιακή προσέγγιση μετατροπής των μη λεκτικών δεδομένων σε αριθμητικές τιμές δεν παράγει πάντοτε αποτελέσματα που να έχουν κάποιο νόημα, ιδιαίτερα όταν στο πεδίο ορισμού των δεδομένων δεν υπάρχει κάποια διάταξη. Έτσι θα ήταν χρήσιμο να βρεθούν κάποιες άλλες διαδικασίες για την εφαρμογή του clustering και σε λεκτικά δεδομένα.

Στην συνέχεια θα περιγράψουμε κάποιους αλγορίθμους τόσο ιεραρχικούς όσο και partitional, οι οποίοι μπορούν να εφαρμοστούν αποτελεσματικά για clustering μη αριθμητικών γνωρίσματων. Στην πρώτη κατηγορία των ιεραρχικών αλγορίθμων ανήκει ο ROCK ενώ στην δεύτερη ανήκουν οι αλγόριθμοι k-modes και k-prototypes, οι οποίοι είναι βασισμένη στο K-Means αλγόριθμο.

2.10.1 ROCK (ROBUST CLUSTERING ALGORITHM FOR CATEGORICAL ATTRIBUTES)

Ο αλγόριθμος ROCK ανήκει στην κατηγορία των ιεραρχικών αλγορίθμων clustering και αντίθετα με άλλους παραδοσιακούς αλγορίθμους μπορεί να χειριστεί αποτελεσματικά boolean και categorical γνωρίσματα. Τα μέτρα απόστασης που χρησιμοποιούν οι παραδοσιακοί αλγόριθμοι clustering δεν είναι κατάλληλα για δεδομένα μη αριθμητικά. Για το λόγο αυτό ο ROCK εισάγει δύο νέες έννοιες στις οποίες βασίζεται για να εκτιμήσει την ομοιότητα/ εγγύτητα μεταξύ των στοιχείων ενός συνόλου δεδομένων. Οι έννοιες αυτές είναι η έννοια του γείτονα(*neighbor*) και των δεσμών(*links*) οι οποίες ορίζονται ως εξής[GRK99]:

- *Γείτονες (Neighbors)*. Οι γείτονες ενός σημείου είναι εκείνα τα σημεία τα οποία παρουσιάζουν σημαντική ομοιότητα με αυτό. Θεωρούμε την $\text{sim}(p_i, p_j)$ ως την συνάρτηση ομοιότητας με βάση την οποία εκτιμούμε την εγγύτητα μεταξύ δύο σημείων και η οποία κυμαίνεται μεταξύ του 0 και 1. Η συνάρτηση μπορεί να είναι ένα οποιαδήποτε καλά ορισμένο μέτρο απόστασης ή ακόμα και μία μη μετρική συνάρτηση(π.χ. μία συνάρτηση ομοιότητας που παρέχεται από ειδικούς στο πεδίο που ανήκουν τα στοιχεία που συγκρίνουμε). Δεδομένου λοιπόν μίας συνάρτησης ομοιότητας και ενός ορίου θ ($\theta \in [0, 1]$), ένα ζεύγος σημείων p_i, p_j είναι γείτονες εάν ισχύει η ακόλουθη ανισότητα:

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

- *Δεσμοί (Links)*. Ο δεσμός $\text{link}(p_i, p_j)$ ορίζεται ως ο αριθμός των κοινών γειτόνων μεταξύ των στοιχείων p_i, p_j .

Το clustering ενός συνόλου δεδομένων που βασίζεται μόνο στην ομοιότητα ή εγγύτητα μεταξύ των στοιχείων του συνόλου δεν έχει αρκετά καλά αποτελέσματα στην διάκριση δύο "όχι τόσο καλά διαχωρίσμων" clusters διότι είναι δυνατόν σημεία τα οποία ανήκουν σε διαφορετικά cluster να είναι γείτονες. Το να είναι δύο στοιχεία που ανήκουν σε διαφορετικά clusters γείτονες, είναι τελείως διαφορετικό με το να έχουν έναν μεγάλο αριθμό κοινών γειτόνων δηλαδή σημείων που να είναι γείτονες και για τα δύο στοιχεία. Η διαπίστωση αυτή καθιστά αναγκαία την χρήση της έννοιας των δεσμών για να καθοριστεί πότε δύο στοιχεία μπορούν να ανήκουν στο ίδιο cluster. Εάν λοιπόν ο αριθμός $\text{link}(p_i, p_j)$ είναι μεγάλος τότε είναι πολύ πιθανό τα στοιχεία p_i, p_j να ανήκουν στο ίδιο cluster. Σε αυτό το μέτρο βασίζεται και ο ROCK για να καθορίσει τα στοιχεία τα οποία μπορούν να συγχωνευτούν σε ένα cluster.

2.10.1.1 ΣΥΝΑΡΤΗΣΗ ΚΡΙΤΗΡΙΟ (CRITERION FUNCTION)

Ένα από τα βασικά θέματα που πρέπει να προσδιοριστούν από μία μέθοδο clustering είναι η εύρεση των καλύτερων clusters. Θα πρέπει επομένως να καθορισθεί με ποιον τρόπο θα μπορούσαμε να προσδιορίσουμε τα βέλτιστα clusters. Εάν κάποιος μπορούσε να χαρακτηρίσει διαμέσου μαθηματικών τα "βέλτιστα clusters", θα μπορούσαμε να δημιουργήσουμε αλγορίθμους οι οποίοι θα βοηθούσαν στην εύρεση αυτών των clusters.

Ένας συνήθης τρόπος για την εύρεση των βέλτιστων clusters είναι ο καθορισμός συναρτήσεων κριτηρίων. Τα clusters που μεγιστοποιούν την συνάρτηση είναι και τα βέλτιστα clusters. Καθώς ενδιαφέρομαστε σε κάθε cluster να έχουμε ένα υψηλό βαθμό συνεκτικότητας, θα θέλαμε να μεγιστοποιήσουμε το άθροισμα των δεσμών $link(p_q, p_r)$ για κάθε ζευγάρι σημείων p_q, p_r που ανήκουν σε ένα cluster και που την ίδια στιγμή ελαχιστοποιούν το άθροισμα των δεσμών $link(p_q, p_s)$ για τα σημεία p_q, p_s σε διαφορετικά clusters. Αυτό οδηγεί στην ακόλουθη συνάρτηση κριτήριο (Εξισ. 2.8) η οποία θα θέλαμε να μεγιστοποιείται για k clusters.

$$E_l = \sum_{i=1}^k n_i \sum_{p_q, p_r \in C_i} \frac{link(p_q, p_r)}{n_i^{1+2f(\theta)}} \quad (\text{Εξισ. 2.8})$$

όπου C_i δηλώνει το cluster i μεγέθους n_i .

Η λογική για την συνάρτηση κριτήριο είναι η ακόλουθη. Καθώς ένας από τους στόχους μας είναι να μεγιστοποιήσουμε την τιμή $link(p_q, p_r)$ για όλα τα ζεύγη τιμών p_q, p_r , μία απλή συνάρτηση κριτήριο όπως η $\sum_{i=1}^k \sum_{p_q, p_r \in C_i} link(p_q, p_r)$ η οποία αναπαριστά απλά το άθροισμα των δεσμών μεταξύ των σημείων στο ίδιο cluster, θα ήταν κατάλληλη. Ωστόσο, αν και αυτή η συνάρτηση διαβεβαιώνει ότι σημεία με ένα μεγάλο αριθμό δεσμών μεταξύ τους ανατίθεται στο ίδιο cluster, δεν μπορεί να αποτρέψει το clustering στο οποίο όλα τα σημεία ανατίθενται σε ένα μόνο cluster. Έτσι δεν δίνει την δυνατότητα σε σημεία με μικρό αριθμό δεσμών μεταξύ τους να διαμοιράζονται σε διαφορετικά clusters.

Προκειμένου να αντιμετωπιστεί το πρόβλημα αυτό, διαιρούμε το συνολικό αριθμό των δεσμών μεταξύ των σημείων σε clusters C_i με τον αναμενόμενο αριθμό των δεσμών στο C_i . Ο συνολικός αριθμός δεσμών σε ένα cluster C_i υπολογίζεται ίσος με $n_i^{1+2f(\theta)}$, όπου $f(\theta)$ είναι μία συνάρτηση η οποία εξαρτάται από το σύνολο των δεδομένων καθώς και το είδος των clusters που μας ενδιαφέρουν. Η συνάρτηση αυτή έχει την εξής σημαντική ιδιότητα: κάθε σημείο που ανήκει σε ένα cluster C_i έχει περίπου $n_i^{f(\theta)}$ γείτονες στο C_i . Εάν μία τέτοια συνάρτηση υπάρχει τότε μπορούμε να υποθέσουμε ότι τα σημεία εκτός του cluster C_i έχουν μικρό αριθμό δεσμών με τα σημεία που ανήκουν στο cluster και κάθε σημείο στο C_i έχει $n_i^{f(\theta)}$ δεσμούς - ένα για κάθε ζευγάρι σημείων. Έτσι θα έχουμε $n_i^{1+2f(\theta)}$ ως αναμενόμενο αριθμό δεσμών ανάμεσα στα σημεία του cluster C_i . Διαιρώντας με τον αναμενόμενο αριθμό δεσμών στη συνάρτηση κριτήριο E_l αποτρέπουμε σημεία με πολύ λίγους δεσμούς ανάμεσα τους να τεθούν στο ίδιο cluster καθώς η ανάθεση τους στο ίδιο cluster μπορεί να προκαλέσει την αύξηση του αναμενόμενου αριθμού των δεσμών για το cluster περισσότερο από τον πραγματικό αριθμό των δεσμών και το αποτέλεσμα θα είναι η μικρότερη τιμή για την συνάρτηση κριτήριο.

2.10.1.2 ΜΕΤΡΟ ΠΟΙΟΤΗΤΑΣ

Η συνάρτηση κριτήριο στην οποία αναφερθήκαμε παραπάνω μπορεί να χρησιμοποιηθεί για την εκτίμηση της ποιότητας των clusters. Τα καλύτερα σημεία του clustering όπως είδαμε έχουν σαν αποτέλεσμα υψηλότερες τιμές για την συνάρτηση κριτήριο. Καθώς ο στόχος είναι να βρούμε ένα clustering του συνόλου των δεδομένων μας που θα μεγιστοποιεί την συνάρτηση κριτήριο, θα χρησιμοποιήσουμε ένα μέτρο παρόμοιο με την συνάρτηση προκειμένου να καθορίσουμε τα καλύτερα ζευγάρια σημείων για να συνδυαστούν σε κάθε βήμα του αλγορίθμου iεραρχικού clustering ROCK. Για κάθε ζεύγος clusters C_i, C_j , έστω ότι $link[C_i, C_j]$ αποθηκεύει τον αριθμό των δεσμών μεταξύ αυτών clusters ο οποίος ισούται με $\sum_{i=1}^k \sum_{p_q, p_r \in C_i} link(p_q, p_r)$. Τότε, μπορούμε να ορίσουμε το μέτρο ποιότητας $g(C_i, C_j)$ για την συσχέτιση των clusters C_i, C_j ως εξής:

$$g(C_i, C_j) = \frac{link[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \quad (\text{Εξισ. 2.9})$$

Τα ζεύγη των clusters για τα οποία το παραπάνω καλύτερο μέτρο ποιότητας είναι μέγιστο είναι το καλύτερο ζεύγος των clusters που πρόκειται να συσχετιστούν σε κάθε βήμα του αλγορίθμου clustering. Ζεύγη clusters λοιπόν με μεγάλο αριθμό δεσμών, είναι γενικά, καλοί υποψήφιοι για συσχέτιση. Ωστόσο, η χρησιμοποίηση του αριθμού των δεσμών ανάμεσα σε ζεύγη clusters σαν έναν δείκτη της ποιότητας συσχέτισης αυτών μπορεί να μην είναι κατάλληλο. Η προσέγγιση αυτή μπορεί να είναι κατάλληλη για καλά διαχωρίσιμα clusters, αλλά στην περίπτωση των outliers ή στην περίπτωση clusters με στοιχεία που είναι γείτονες, ένα μεγάλο cluster μπορεί να σκιάσει άλλα clusters και έτσι σημεία από διαφορετικά clusters μπορεί να συσχετιστούν σε ένα μόνο cluster. Αυτό συμβαίνει γιατί ένα μεγάλο cluster θα έχει τυπικά μεγαλύτερο αριθμό δεσμών με άλλα clusters.

Προκειμένου όμως να περιορίσουμε το πρόβλημα, διαιρούμε τον αριθμό των δεσμών μεταξύ των clusters με τον αναμενόμενο αριθμό δεσμών μεταξύ αυτών. Ετσι, εάν κάθε σημείο σε ένα cluster C_i έχει $n_i^{f(\theta)}$ γείτονες, τότε ο αναμενόμενος αριθμός δεσμών μεταξύ των σημείων του cluster είναι περίπου $n_i^{1+2f(\theta)}$. Καθώς για μεγάλα clusters, μπορούμε να υποθέσουμε ότι σημεία εκτός του cluster συντελούν ελάχιστα στον αριθμό των δεσμών μεταξύ σημείων ενός cluster, ο αναμενόμενος αριθμός δεσμών μεταξύ σημείων του cluster είναι περίπου $n_i^{1+2f(\theta)}$. Επομένως, εάν δύο αρκετά μεγάλα clusters με μέγεθος n_i, n_j συγχωνευτούν, ο αριθμός των δεσμών ανάμεσα στα ζεύγη των σημείων στο clusters που προέκυψε από την συγχώνευση είναι $(n_i + n_j)^{1+2f(\theta)}$, ενώ ο αριθμός των δεσμών σε κάθε cluster(πριν την συγχώνευση) είναι $n_i^{1+2f(\theta)}$. και $n_j^{1+2f(\theta)}$, αντίστοιχα. Ετσι, ο αναμενόμενος αριθμός των δεσμών μεταξύ ζευγών σημείων όπου κάθε ένα θα προέρχεται από διαφορετικό cluster θα είναι $(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}$. Χρησιμοποιούμε λοιπόν τον παράγοντα αυτό ως παράγοντα κανονικοποίησης στο μέτρο ποιότητας ώστε να μας οδηγήσει σε clusters με υψηλές τιμές για την συνάρτηση κριτήριο.

2.10.1.3 ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΟΥ ROCK

Ο αλγόριθμος εφαρμόζεται σε ένα δείγμα του συνόλου των δεδομένων το οποίο επιλέγεται με τυχαίο τρόπο. Ο αλγόριθμος λαμβάνει σαν είσοδο τον αριθμό του

συνόλου των π σημείων του δείγματος στα οποία πρόκειται να γίνει clustering καθώς και ο αριθμός k των clusters. Η διαδικασία ξεκινά με τον υπολογισμό του αριθμού των δεσμών ανάμεσα στα ζεύγη των σημείων. Για κάθε cluster i, διατηρούμε σε μία στοίβα κάθε cluster j για το οποίο ο αριθμός των δεσμών link[i, j] δεν είναι 0. Τα clusters j στην στοίβα q[i] ταξινομούνται με φθίνουσα σειρά του μέτρου ποιότητας για σύνδεση των clusters.

Επιπρόσθετα με την στοίβα Q ο αλγόριθμος διατηρεί και μία στοίβα με όλα τα clusters. Επιπρόσθετα τα clusters στο Q ταξινομούνται με φθίνουσα σειρά των μέτρων ποιότητας των clusters. Έτσι, g(j, max(q[j])) χρησιμοποιείται για να ταξινομηθούν τα διάφορα clusters j στο Q, όπου το max(q[j]) είναι το καλύτερο cluster το οποίο μπορεί να συγχωνευτεί με cluster j. Σε κάθε βήμα, το μέγιστο cluster j στο Q και το μέγιστο cluster στο q[j] είναι το καλύτερο ζεύγος των clusters που πρόκειται να συγχωνευτούν.

```
procedure cluster(S, k)
begin
    link := compute_links(S)
    for each s ∈ S do
        q[s]:=build_local_heap(link, s)
    Q := build_global_heap(S,q)
    while size(Q) > k do {
        u:= extract_max(Q)
        v:= max(q[u])
        delete(Q, v)
        w:=merge(u,v)
        for each x ∈ q[v] ∪ q[u] do{
            link[x,w]:=link[x,u]+link[x,v]
            delete(q[x],w,g(x,w); insert(q[w],x,g(x,w))
            update(Q,x,q[x])}
        insert(Q,w,q[w])
        deallocate(q[u]); deallocate(q[v])
    }
end
```

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

Με βάση την παραπάνω περιγραφή του αλγορίθμου μπορούμε να δούμε ότι η πολυπλοκότητα για τον υπολογισμό των δεσμών ανάμεσα στα σημεία είναι $O(n^2 m_a)$ για τον μέσο αριθμό των γειτόνων m_a . Ο χρόνος για να δημιουργήσουμε μία τοπική

στοίβα, αρχικά είναι $O(n)$ (μία στοίβα για ένα σύνολο n clusters εισόδου μπορεί να δημιουργηθεί σε χρόνο ο οποίος είναι γραμμικά ανάλογος με τον αριθμό των clusters). Η γενική στοίβα έχει επίσης το πολύ n clusters αρχικά και μπορεί να δημιουργηθεί σε χρόνο $O(n)$. Η πολυπλοκότητα για τα βήματα στο while-loop εκτελούνται σε χρόνο $O(n)$. Το εσωτερικό for-loop κυριαρχεί την πολυπλοκότητα του while-loop. Καθώς το μέγεθος για κάθε τοπική ουρά μπορεί να είναι στην χειρότερη περίπτωση n και το ένα w cluster που προκύπτει από την συγχώνευση άλλων clusters μπορεί να χρειάζεται να εισαχθεί σε $O(n)$ τοπικές ουρές, η χρονική πολυπλοκότητα για το for-loop είναι $O(n \log n)$ και επομένως για το while-loop $O(n^2 \log n)$ στη χειρότερη περίπτωση. Με βάση λοιπόν την μέχρι τώρα ανάλυση η πολυπλοκότητα του αλγορίθμου ROCK σε ότι αφορά τον υπολογισμό της λίστας των γειτονικών κόμβων και δεσμών, είναι $O(n^2 + nm_m m_a + n^2 \log n)$.

2.10.2 ΑΛΓΟΡΙΘΜΟΙ CLUSTERING ΒΑΣΙΣΜΕΝΟΙ ΣΤΟΝ K-MEANS ΓΙΑ ΛΕΚΤΙΚΑ ΔΕΔΟΜΕΝΑ

Σε σχέση με άλλους αλγορίθμους ο K-Means και οι παραλλαγές του προσαρμόζεται καλά στην διαδικασία data mining λόγω της αποδοτικότητας του στην επεξεργασία μεγάλων συνόλων δεδομένων. Ωστόσο, η χρήση τους περιορίζεται συχνά σε αριθμητικά δεδομένα λόγω του ότι αυτοί οι αλγόριθμοι ελαχιστοποιούν την συνάρτηση κόστους υπολογίζοντας τους μέσους των clusters. Για το σκοπό αυτό τα τελευταία χρόνια έχουν γίνει κάποιες προσπάθειες για ανάπτυξη αλγορίθμων clustering που θα επεκτείνουν την βασική λογική του K-Means και για clustering σε γνωρίσματα με λεκτικές τιμές (categorical attributes) [Huang 97]. Οι κυριότεροι αλγόριθμοι clustering οι οποίοι βασίζονται στον K-Means αλγόριθμο είναι ο k-prototypes και ο k-mode. Οι αλγόριθμοι αυτοί σχεδιάστηκαν από τον Huang και έχουν αποδειχθεί αποτελεσματικοί για μεγάλα σύνολα λεκτικών δεδομένων σε σχέση με άλλους αλγορίθμους κυρίως ierarchical ή k-means καθίστανται μη αποδοτικοί για μεγάλα σύνολα δεδομένων.

2.10.2.1 ΑΛΓΟΡΙΘΜΟΣ K-PROTOTYPES

Ο k-prototypes σχεδιάστηκε για clustering μεγάλων συνόλων βάσεων δεδομένων με τιμές αριθμητικές και λεκτικές. Στον αλγόριθμο ορίζεται ένα μέτρο ανομοιότητας το οποίο λαμβάνει υπόψη του γνωρίσματα τόσο με αριθμητικές όσο και με λεκτικές τιμές. Επίσης θεωρεί ότι s_n είναι το μέτρο ανομοιότητας σε αριθμητικά γνωρίσματα το οποίο ορίζεται από την Eukleidεια απόσταση και s_c είναι το μέτρο ανομοιότητας για λεκτικά γνωρίσματα το οποίο ορίζεται σαν ο αριθμός των αταίριαστων κατηγοριών μεταξύ δύο αντικειμένων. Το μέτρο ανομοιότητας ανάμεσα στα δύο αντικείμενα ορίζεται ως $s_n + \gamma s_c$, όπου γ είναι ένα βάρος για την εξισορρόπηση των δύο μερών και την αποφυγή της εύνοιας κάποιου από τους τύπους των γνωρισμάτων.

Η διαδικασία clustering του k-prototypes είναι ανάλογη με τον αλγόριθμο K-Means εκτός από την νέα μέθοδο που χρησιμοποιείται για την ενημέρωση των λεκτικών τιμών των προτύπων(κέντρων) των clusters. Ένα πρόβλημα που προκύπτει από την χρήση αυτού του αλγορίθμου είναι η επιλογή του κατάλληλου βάρους. Μία πρόταση είναι να χρησιμοποιηθεί σαν βάση για την επιλογή του βάρους η τυπική απόκλιση των αριθμητικών γνωρισμάτων.

2.10.2.2 ΑΛΓΟΡΙΘΜΟΣ K-MODES

Ο αλγόριθμος k-modes προτάθηκε από τον Huang [Huang 1997] όπως και ο k-prototypes και αποτέλεσε μία απλούστευση του δεύτερου καθώς λαμβάνει υπόψη του μόνο γνωρίσματα με λεκτικές τιμές. Συνεπώς, το βάρος γ δεν είναι πλέον απαραίτητο στον αλγόριθμο καθώς δεν λαμβάνεται υπόψη ο παράγοντας s_n . Η προσέγγιση αυτή θεωρεί ότι εάν στο σύνολο δεδομένων περιλαμβάνονται και αριθμητικά γνωρίσματα τότε γίνεται κατηγοριοποίηση αυτών χρησιμοποιώντας τον αλγόριθμο K-Means ή κάποια από τις παραλλαγές του.

Γενικά, ο k-modes αλγόριθμος βασίζεται στον K-Means στον οποίο όμως έχουν γίνει οι εξής τρεις τροποποιήσεις:

- χρησιμοποιούνται διαφορετικά μέτρα ανομοιότητας έτσι ώστε να μπορούν να εφαρμοστούν σε λεκτικές τιμές,
- αντικαταστάθηκαν τα K-Means με τα k-modes, και
- χρησιμοποιούνται μέθοδοι βασισμένη στην συχνότητα εμφάνισης των τιμών προκειμένου να ενημερωθούν τα κέντρα των clusters, δηλαδή τα modes.

Στην συνέχεια παρουσιάζουμε με μεγαλύτερη λεπτομέρεια τις τροποποιήσεις αυτές.

2.10.2.2.1 ΜΕΤΡΑ ΑΝΟΜΟΙΟΤΗΤΑΣ

Έστω ότι X, Y είναι δύο αντικείμενα με τη λεκτικά γνωρίσματα. Το μέτρο ανομοιότητας μεταξύ του X και Y μπορεί να οριστεί με βάση την συνολική ανομοιότητα μεταξύ των λεκτικών γνωρισμάτων των δύο αντικείμενων. Όσο πιο μικρός είναι ο αριθμός των αταίριαστων τιμών των αντίστοιχων γνωρισμάτων των δύο αντικείμενων, τόσο περισσότερο όμοια μπορούν να θεωρηθούν τα δύο αντικείμενα. Τυπικά,

$$d(X, Y) = \sum_{j=1}^m \delta(x_j, y_j) \quad (\text{Εξισ. 2.10})$$

όπου

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad (\text{Εξισ. 2.11})$$

Το $d(X, Y)$ δίνει ίση σημαντικότητα σε κάθε κατηγορία ενός γνωρίσματος. Εάν λάβουμε υπόψη μας τις συχνότητες των κατηγοριών σε ένα σύνολο δεδομένων, μπορούμε να ορίσουμε το μέτρο ανομοιότητας ως εξής

$$d_{x^2}(X, Y) = \sum_{j=1}^m \frac{(n_{x_j} + n_{y_j})}{n_{x_j} n_{y_j}} \delta(x_j, y_j) \quad (\text{Εξισ. 2.12})$$

όπου n_{x_j}, n_{y_j} είναι αριθμοί των αντικειμένων σε ένα σύνολο δεδομένων τα οποία έχουν κατηγορίες x_j, y_j για το γνώρισμα j . Η απόσταση που ορίζεται σύμφωνα με τον παραπάνω τύπο είναι όμοια με την chi-square απόσταση και καλείται chi-square απόσταση.

2.10.2.2.2 ΕΥΡΕΣΗ MODE ΓΙΑ ΕΝΑ ΣΥΝΟΛΟ

Εστω ότι X είναι ένα σύνολο αντικειμένων που περιγράφονται από τα λεκτικά γνωρίσματα A_1, A_2, \dots, A_m .

Mode ενός αντικειμένου X είναι ένα διάνυσμα $Q = [q_1, q_2, \dots, q_m]$ το οποίο ελαχιστοποιεί την εξίσωση

$$D(X, Q) = \sum_{i=1}^n d(X_i, Q) \quad (\text{Εξισ. 2.13})$$

Όπου $X = \{X_1, X_2, \dots, X_n\}$ και d μπορεί να οριστεί με βάση την εξίσωση (Εξισ. 2.10) ή με βάση την εξίσωση (Εξισ. 2.12). Εδώ, το Q δεν είναι απαραίτητο να είναι στοιχείο του X .

Εάν $n_{ck,j}$ είναι ο αριθμός των αντικειμένων που έχουν κατηγορία $c_{k,j}$ στο γνώρισμα A_j και $f_r(A_j = c_{k,j} | X) = n_{ck,j}/n$ η σχετική συχνότητα της κατηγορίας $c_{k,j}$ στο X .

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

$$f_r(A_j = q_j | X) \geq f_r(A_j = c_{k,j} | X) \quad (\text{Εξισ. 2.14})$$

για $q_j \neq c_{k,j}$ για όλα τα $j = 1 \dots m$.

Το παραπάνω θεώρημα ορίζει έναν τρόπο για να βρούμε ένα Q για ένα δεδομένο X και για το λόγο αυτό είναι σημαντικό γιατί επιτρέπει να χρησιμοποιούμε τον K-Means για clustering λεκτικών τιμών χωρίς να περιορίζεται η αποδοτικότητα. Το mode βέβαια ενός συνόλου δεδομένων X δεν είναι μοναδικό. Για παράδειγμα το mode ενός συνόλου $\{[a, b], [a, c], [c, b], [b, c]\}$ μπορεί να είναι είτε $[a, b]$ ή $[a, c]$.

2.10.2.2.3 Ο ΑΛΓΟΡΙΘΜΟΣ K-MODES

Εστω $\{S_1, S_2, \dots, S_k\}$ είναι μία τμηματοποίηση του X , όπου $S_i \neq \emptyset$ για $1 \leq i \leq k$ και $\{Q_1, Q_2, \dots, Q_k\}$ τα modes του $\{S_1, S_2, \dots, S_k\}$. Το συνολικό κόστος της τμηματοποίησης ορίζεται ως εξής:

$$E = \sum_{i=1}^k \sum_{l=1}^n y_{i,l} d(X_i, Q_l) \quad (\text{Εξισ. 2.15})$$

Όπου $y_{i,l}$ είναι ένα στοιχείο του πίνακα συμμετοχής $Y_{n \times 1}$ και d η απόσταση όπως ορίζεται στη (Εξισ. 2.11) ή (Εξισ. 2.12).

Παρόμοια με τον K-Means αλγόριθμο, ο αντικειμενικός σκοπός του clustering είναι να βρούμε ένα σύνολο $\{Q_1, Q_2, \dots, Q_k\}$ το οποίο μπορεί να ελαχιστοποιήσει το E . Τα βασικά βήματα του αλγορίθμου k-modes είναι τα ακόλουθα :

1. Επιλογή k αρχικών modes, ένα για κάθε cluster.
2. Ανάθεση ενός αντικειμένου στο cluster του οποίου το mode είναι πιο κοντά στο αντικείμενο σύμφωνα με το d . Ενημέρωση του mode του cluster μετά από κάθε ανάθεση σύμφωνα με το Θεώρημα.

3. Αφού όλα τα αντικείμενα έχουν τοποθετηθεί σε clusters, γίνεται επανέλεγχος της ανομοιότητας των αντικειμένων ως προς τα τρέχοντα modes. Εάν για ένα αντικείμενο βρεθεί ότι βρίσκεται πιο κοντά στο mode ενός άλλου cluster από ότι στο mode του τρέχοντος cluster, επανατοποθετείται στο αντικείμενο στο άλλο cluster και ενημερώνονται ανάλογα τα modes των clusters.
4. Επαναλαμβάνεται το βήμα 3 μέχρις ότου κανένα αντικείμενο να μην αλλάξει clusters μετά από τον πλήρη έλεγχο όλου του συνόλου δεδομένων.

Όπως και ο K-Means έτσι και ο αλγόριθμος k-modes παράγει βέλτιστες λύσεις οι οποίες εξαρτώνται από τα αρχικά modes και την διάταξη των αντικειμένων στο σύνολο των δεδομένων. Για το λόγο αυτό χρησιμοποιούνται διάφορες τεχνικές για τον ορισμό των αρχικών k modes. Μία μέθοδος είναι να επιλεχθούν οι k πρώτες εγγραφές ως τα k αρχικά modes του αλγορίθμου. Μία δεύτερη μέθοδος που μπορεί να εφαρμοστεί αποτελείται από τα εξής βήματα:

1. Υπολογίζουμε τις συχνότητες εμφάνισης όλων των κατηγοριών για όλα τα γνωρίσματα και τα αποθηκεύουμε σε έναν πίνακα με φθίνουσα σειρά των συχνοτήτων, ως εξής:

$c_{1,1}$	$c_{1,2}$	$c_{1,3}$	$c_{1,4}$
$c_{2,1}$	$c_{2,2}$	$c_{2,3}$	$c_{2,4}$
$c_{3,1}$		$c_{3,3}$	$c_{3,4}$
$c_{4,1}$		$c_{4,3}$	
		$c_{5,3}$	

Στον παραπάνω πίνακα $c_{i,j}$ 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. Κατανέμουμε τις πιο συχνές κατηγορίες ομοιόμορφα στα k αρχικά modes. Για παράδειγμα αναθέτουμε, $Q_1 = [q_{1,1} = c_{1,1}, q_{1,2} = c_{2,2}, q_{1,3} = c_{3,3}, q_{1,4} = c_{1,4}]$, $Q_2 = [q_{2,1} = c_{2,1}, q_{2,2} = c_{1,2}, q_{2,3} = c_{4,3}, q_{2,4} = c_{2,4}]$ και $Q_3 = [q_{3,1} = c_{3,1}, q_{3,2} = c_{2,2}, q_{3,3} = c_{1,3}, q_{3,4} = c_{3,4}]$.
3. Αρχίζουμε με Q_1 . Επιλέγουμε την εγγραφή που είναι περισσότερο όμοια με το Q_1 και αντικαθιστούμε το Q_1 με την εγγραφή αυτή σαν το πρώτο αρχικό mode. Η ίδια διαδικασία συνεχίζεται και με τις υπόλοιπες εγγραφές. Με τον τρόπο αυτό προσπαθούμε να αποφύγουμε την εμφάνιση κενών clusters.

Γενικά, ο σκοπός της μεθόδου επιλογής των αρχικών modes είναι να κάνουμε διαφορετικά τα αρχικά modes και έτσι να καταλήξουμε σε καλύτερα αποτελέσματα clustering.

3^ο ΚΕΦΑΛΑΙΟ

FUZZY CLUSTER ΑΝΑΛΥΣΗ

3.1 ΕΙΣΑΓΩΓΗ

Η cluster ανάλυση βασίζεται στην τμηματοποίηση μίας συλλογής δεδομένων σε έναν αριθμό υποομάδων, όπου τα αντικείμενα σε ένα cluster (υποομάδα) θα παρουσιάζουν ένα συγκεκριμένο βαθμό ομοιότητας ή εγγύτητας μεταξύ τους. Το *crisp clustering* αναθέτει κάθε στοιχείο (διάνυσμα γνωρισμάτων) σε ένα και μόνο ένα από τα clusters, με βαθμό συμμετοχής ίσο με την μονάδα, θεωρώντας ότι μεταξύ των clusters υπάρχουν καλά ορισμένα όρια. Ωστόσο, αυτό το μοντέλο συχνά δεν ανταποκρίνεται στα πραγματικά δεδομένα, όπου τα όρια μεταξύ των υποομάδων μπορεί να είναι ασαφή και όπου απαιτείται μία πιο λεπτομερή περιγραφή της συμμετοχής ενός αντικειμένου σε ένα cluster καθώς υπάρχουν περιπτώσεις που δεν μπορούμε να τοποθετήσουμε κάποιο στοιχείο σε κάποιο συγκεκριμένο cluster.

Η αντιμετώπιση πολλών προβλημάτων κυρίως στο χώρο των επιστημών ζωής φαίνεται να αντιμετωπίζονται αποτελεσματικότερα με λήψη αποφάσεων σε ένα ασαφές περιβάλλον. Για το λόγο αυτό, από πολύ νωρίς οι προσπάθειες των επιστημόνων στράφηκαν στην ανάπτυξη μίας οικογένειας fuzzy clustering αλγορίθμων, οι οποίοι αποτελούν στην πραγματικότητα επεκτάσεις των κλασικών αλγορίθμων σε ασαφές περιβάλλον. Χαρακτηριστικές εργασίες στον τομέα αυτό είναι του Bezdek ο οποίος το 1973 ανέπτυξε μία οικογένεια fuzzy clustering αλγορίθμων, οι οποίοι βασίζονται στην ασαφή επέκταση του κριτηρίου των ελαχίστων τετραγώνων και απέδειξε την σύγκλιση των αλγορίθμων σε ένα τοπικό ελάχιστο. Επίσης σχετικοί αλγόριθμοι, λαμβάνοντας υπόψη διαφορετικά cluster σχήματα έχουν προταθεί από τους Bezdek, Dunn και Gustafson, Kessel.

Γενικά, η fuzzy cluster ανάλυση τμηματοποιεί τα στοιχεία ενός συνόλου δεδομένων σε clusters καθορίζοντας κάποιο βαθμό συμμετοχής μεταξύ 0 και 1 για κάθε δείγμα δεδομένων που τοποθετείται σε κάποιο cluster. Έτσι ένα στοιχείο μπορεί να ανήκει σε περισσότερα από ένα clusters με διαφορετικό όμως βαθμό συμμετοχής στο καθένα ανάλογα με την εγγύτητα ή ομοιότητα που παρουσιάζει σε σχέση με τα άλλα στοιχεία των clusters.

3.2 FUZZY C-MEANS ΑΛΓΟΡΙΘΜΟΣ ΚΑΙ ΟΙ ΠΑΡΑΛΛΑΓΕΣ ΤΟΥ

Ένας από τους βασικούς αλγορίθμους για fuzzy clustering είναι ο Fuzzy C-Means. Μέχρι τώρα έχουν εμφανιστεί διάφορες παραλλαγές του Fuzzy C-Means Clustering, οι οποίες σχετίζονται με τα διάφορα σχήματα των clusters (hyperelipsoidal, spherical, linear, κ.λ.π.). Τα σχήματα των clusters καθορίζονται από την συγκεκριμένη μορφή μίας εκ των προτέρων θεωρούμενης αντικειμενικής συνάρτησης. Το κοινό στοιχείο σε όλες τις εναλλακτικές προσεγγίσεις clustering είναι ο τρόπος βελτιστοποίησης της αντικειμενικής συνάρτησης η οποία αξιολογεί μία δεδομένη ασαφή ανάθεση δεδομένων σε clusters [HBD96].

Ο στόχος του clustering είναι να βρεθούν ομάδες όμοιων στοιχείων (clusters) σε ένα σύνολο n αντικειμένων $O = \{o_1, o_2, \dots, o_n\}$. Για να περιγράψουμε τα αποτελέσματα του clustering, χρησιμοποιούμε πίνακες $U = [U_{ik}] \in R^{c \times n}$, όπου το c είναι ο αριθμός των θεωρούμενων clusters και n είναι ο αριθμός των αντικειμένων που ταξινομούνται σε clusters. Ο αριθμός U_{ik} αναπαριστά τον βαθμό συσχέτισης του αντικειμένου o_k με το cluster i .

Η ασαφής διαίρεση (fuzzy partitioning) των δεδομένων επιτρέπει τη συνολική ($=1$) συμμετοχή κάθε αντικειμένου να κατανέμεται ανάμεσα στα c clusters έτσι ώστε το U να μπορεί να είναι ένα οποιοδήποτε στοιχείο του συνόλου των ασαφών c-τμημάτων (Ruspini, 1969):

$$\begin{aligned} M_{fcn} = \{U \in R^{c \times n} | & U \in [0,1] \text{ για } 1 \leq i \leq c \text{ και } 1 \leq k \leq n, \\ & \sum_{i=1}^c U_{ik} = 1 \text{ για } 1 \leq k \leq n, \\ & \sum_{k=1}^n U_{ik} > 0 \text{ για } 1 \leq i \leq c\} \end{aligned}$$

Εστω λοιπόν ότι έχουμε ένα σύνολο από δεδομένα $X = \{x_1, \dots, x_n\} \subset R^s$ τα οποία θέλουμε να διαιρέσουμε σε c clusters. Οι αλγόριθμοι c-means υποθέτουν ότι τα clusters έχουν την μορφή "σφαίρας" και προσπαθούν να επιτύχουν την ταυτόχρονη τμηματοποίηση των δεδομένων και να υπολογίσουν τα κέντρα των clusters u_1, u_2, \dots, u_c .

Η μέθοδος fuzzy c-means βασίζεται στην επαναληπτική βελτιστοποίηση της αντικειμενικής συνάρτησης:

$$J_m(U, V) = \sum_{i=1}^c \sum_{k=1}^n U_{ik}^m d^2(x_k, v_i) \quad (\text{Εξισ. 3.1})$$

όπου $U \in M_{fcn}$, $V = [v_1, \dots, v_c] \in R^{s \times c}$ ο πίνακας προτύπων (κέντρα των clusters), c ο αριθμός των θεωρούμενων clusters, n ο αριθμός των δεδομένων και $m \geq 1$ παράμετρος ασάφειας. Η παράμετρος m είναι ένας δείκτης της ασάφειας που λαμβάνουμε στα clusters. Ειδικότερα, ισχύει ότι όταν $m \rightarrow 1$ τότε τα clusters τείνει να είναι

σαφή(crisp), δηλαδή $U_{ik} \rightarrow 1$ ή $U_{ik} \rightarrow 0$, ενώ για $m \rightarrow \infty$ έχουμε $U_{ik} \rightarrow 1/c$. Συνήθως επιλέγουμε $m=2$.

Η αντικειμενική συνάρτηση (Εξισ. 3.1) διασφαλίζει ότι κανένα cluster δεν θα είναι τελείως κενό και διαβεβαιώνει ότι για κάθε σύνολο δεδομένων η κατανομή των στοιχείων μπορεί να γίνει σε διαφορετικά clusters αλλά το σύνολο των βαθμών συμμετοχής κάθε στοιχείου στο σύνολο των clusters θα είναι ίση με την μονάδα. Η βελτιστοποίηση της αντικειμενικής συνάρτησης προέρχεται λαμβάνοντας το ελάχιστο της συνάρτησης, σύμφωνα με τον εξής τύπο

$$\min_{U, u_1, u_2, \dots, u_N} J_m(U, V)$$

Σύμφωνα με τον παραπάνω τύπο λαμβάνοντας το τοπικό ελάχιστο ή το οριακό σημείο του $J_m(U, V)$ μπορούμε να υπολογίσουμε τις νέες τιμές κάθε φορά της μήτρας συμμετοχής

$$U_{ik} = 1 / \sum_{j=1}^c \left(\frac{d(x_k - v_i)}{d(x_k - v_j)} \right)^{2/(m-1)} \quad (\text{Εξισ. 3.2})$$

Τα πρότυπα (κέντρα) των clusters λαμβάνονται ως η μέση τιμή των προτύπων

$$v_i = \sum_{k=1}^N U_{ik}^m x_k / \sum_{k=1}^N U_{ik}^m \quad (\text{Εξισ. 3.3})$$

Στην απλή περίπτωση του FCM χρησιμοποιείται ως μέτρο απόστασης d η Ευκλείδεια απόσταση. Στην περίπτωση αυτή τα clusters που προκύπτουν έχουν σχήμα σφαιρικό και περίπου το ίδιο μέγεθος (αριθμό στοιχείων).

Προκειμένου να γίνει καλύτερα κατανοητή η διαδικασία του fuzzy clustering παρουσιάζουμε στην συνέχεια τα βασικά βήματα του αλγορίθμου fuzzy c-means με την μορφή ψευδοκώδικα.

3.2.1 FUZZY C-MEANS ΑΛΓΟΡΙΘΜΟΣ ΓΙΑ OBJECT-DATA

Τα βήματα του αλγορίθμου Fuzzy C-Means είναι τα εξής:

Βήμα 1. Αρχικοποίηση.

Θεωρούμε μία αρχική τμηματοποίηση $U^{(0)}$ και ορίζουμε ένα κριτήριο σταματήματος ϵ .

Στη συνέχεια για κάθε επανάληψη r , $r=1, \dots, r_{\max}$ έχουμε:

Βήμα 2. Ενημέρωση αποστάσεων(Update distances). Υπολογίσουμε τους νέους μέσους για τα clusters και τις αποστάσεις σύμφωνα με τους τύπους

$$v_i^{(r)} = \sum_{k=1}^n \left(U_{ik}^{(r)} \right)^m x_k \Bigg/ \sum_{k=1}^n \left(U_{ik}^{(r)} \right)^m \quad \text{για } 1 \leq i \leq c, \quad (\text{Εξισ. 3.4})$$

$$\left(d_{ik}^{(r)} \right)^2 = \|x_k - v_i^{(r)}\|^2 \quad \text{για } 1 \leq i \leq c, 1 \leq k \leq n.$$

Βήμα 3. Ενημέρωση τμημάτων (Update partition). Χρησιμοποιούμε τις νέες αποστάσεις για να υπολογίσουμε το νέο τμήμα $U^{(r+1)}$ μέσω της fuzzy φόρμουλας ενημέρωσης όπως αυτή ορίζεται από τον (Bezdek, 1981). Για $k=1, \dots, n$

Εάν $d_{ik}^{(r)} > 0$, για όλα τα $i=1, 2, \dots, c$ τότε

$$U_{ik}^{(r+1)} = \left[\sum \left[\left(d_{ik}^{(r)} \right)^2 / \left(d_{jk}^{(r)} \right)^2 \right] \right]^{-1/(m-1)} \quad (\text{Εξισ. 3.5})$$

Διαφορετικά εάν ένα τουλάχιστον $d_{ik}^{(r)} = 0$,

$$U_{ik}^{(r+1)} = 0 \text{ εαν } d_{ik}^{(r)} > 0 \text{ και} \quad (\text{Εξισ. 3.6})$$

$$U_{ik}^{(r+1)} > 0 \text{ εαν } d_{ik}^{(r)} = 0, \quad U_{ik}^{(r+1)} \in [0,1] \text{ και} \quad \sum_{i=1}^c U_{ik}^{(r+1)} = 1$$

Βήμα 4. Ελεγχος απόκλισης (Convergence check). Η διαδικασία σταματάει όταν $\|U^{(r+1)} - U^{(r)}\| < \epsilon$, αλλιώς επιστρέφουμε στο βήμα 2 με $r = r+1$.

3.2.2 ΠΑΡΑΛΛΑΓΕΣ ΤΟΥ FUZZY C-MEANS (FCM)

Υιοθετώντας διαφορετικούς ορισμούς για τα μέτρα απόστασης οι Gustafson, Kessel και οι Gath, Geva σχεδίασαν μεθόδους fuzzy clustering οι οποίοι αναζητούν υπερ-ελλειψοειδή clusters (hyper-ellipsoidal clusters) των οποίων το μέγεθος μπορεί να ποικίλει (δεν έχουν όλα τα clusters τον ίδιο περίπου αριθμό στοιχείων). Και στις δύο παραπάνω περιπτώσεις εκτός από τα πρότυπα v_i και τους βαθμούς συμμετοχής U_{ik} , για κάθε cluster i ορίζεται μία μήτρα συνδιακύμανσης (covariance matrix) C_i . Ο αλγόριθμος που πρότειναν ο Gustafson και Kessel (GK) αντικαθιστά την Ευκλείδεια απόσταση με την απόσταση που ορίζεται από την εξίσωση

$$d^2(x_k, v_i) = (\rho_i \det C_i)^{1/p} (x_k - c_i)^T C_i^{-1} (x_k - c_i) \quad (\text{Εξισ. 3.7})$$

Ενώ η μέθοδος που προτάθηκε από τους Gath και Geva (GG) βασίζεται στην κανονική κατανομή και χρησιμοποιεί την απόσταση

$$d^2(x_k, v_i) = \frac{(\det C_i)^{1/2}}{p_i} \exp\left(\frac{(x_k - c_i)^T C^{-1}(x_k - c_i)}{2}\right) \quad (\text{Εξισ. 3.8})$$

όπου το p_i δίνεται από την εξίσωση (Εξισ. 3.9)

$$p_i = \frac{\sum_{k=1}^n (U_{ik})^m}{\sum_{j=1}^c \sum_{k=1}^n (U_{jk})^m} \quad (\text{Εξισ. 3.9})$$

έτσι ώστε και στις δύο περιπτώσεις για κάθε cluster ο αντίστροφος και η ορίζουσα του πίνακα θα πρέπει να ορίζονται σε κάθε επανάληψη. Για τον GK το μέγεθος κάθε cluster θα πρέπει να καθορίζεται εκ των προτέρων με την τιμή του ρ_i , ενώ στον GG το μέγεθος των clusters δεν χρειάζεται να ορίζεται εκ των προτέρων.

Οι αλγόριθμοι GG και GK όπως τους περιγράψαμε παραπάνω είναι αρκετά πολύπλοκοι λόγω των υπολογισμών που απαιτούν σε σχέση με την μήτρα συνδιακύμανσης. Επίσης η εξαγωγή κανόνων παρουσιάζει αρκετά προβλήματα καθώς οι αλγόριθμοι οδηγούν στον καθορισμό υπερ-ελλειψοειδή clusters. Συνεπώς θα ήταν χρήσιμο να περιορίσουμε τους πίνακες συνδιακύμανσης που χρησιμοποιούν οι αλγόριθμοι GG και GK σε διαγώνιους πίνακες. Κατά την εφαρμογή των αλγορίθμων η ενημέρωση της μήτρας C_i θα πρέπει να γίνεται με τρόπο ώστε να διατηρείται διαγώνια και να συμβάλλει στην τροποποίηση της Ευκλείδειας απόστασης μέσω των (Εξισ. 3.7) και (Εξισ. 3.8). Για το λόγο αυτό, απαιτείται μία νέα φόρμουλα για την ενημέρωση της μήτρας συνδιακύμανσης, η οποία μπορεί να γίνει με βάση τους εξής τύπους:

$$C_u^{(i)} = \frac{\left(\rho_i \prod_{a=1}^p \sum_{k=1}^n (U_{ik})^m (x_{ka} - v_{ia})^2\right)^{1/p}}{\sum_{k=1}^n (U_{ik})^m (x_{ku} - v_{iu})^2} \quad \text{and} \quad C_u^{(i)} = \frac{\sum_{k=1}^n (U_{ik})^m}{\sum_{k=1}^n (U_{ik})^m (x_{ku} - v_{iu})^2}$$

οι οποίοι αποτελούν τα βασικά σχήματα ενημέρωσης για τους αλγορίθμους GG και GK, αντίστοιχα. Επίσης με $c_u^{(i)}$ δηλώνει το υ διαγώνιο στοιχείο του πίνακα C_i και x_{ka} v_{ia} είναι η α συντεταγμένη του διανύσματος x_k και v_i . Με βάση τους παραπάνω τύπους μπορούμε να διαπιστώσουμε ότι για την μήτρα συνδιακύμανσης δεν απαιτείται ο υπολογισμός ούτε του αντίστροφου πίνακα ούτε της ορίζουσας και έτσι οι αντίστοιχοι αλγόριθμοι είναι απλούστεροι και πιο γρήγοροι σε σχέση με τους αυθεντικούς αλγορίθμους GG και GK.

3.3 ΕΦΑΡΜΟΓΗ ΤΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ C-MEANS CLUSTERING ΣΕ ΣΧΕΣΙΑΚΑ ΔΕΔΟΜΕΝΑ

Ο στόχος της cluster ανάλυσης όπως έχουμε αναφέρει και σε προηγούμενες παραγράφους είναι να βρεθούν clusters (ομάδες όμοιων αντικειμένων) σε ένα σύνολο O αντικειμένων $O = \{o_1, o_2, \dots, o_n\}$. Το σύνολο των αντικειμένων που προαναφέρθηκε συχνά περιγράφεται χρησιμοποιώντας αριθμητικά δεδομένα αντικειμένων (object data) ή αριθμητικά σχεσιακά δεδομένα (relational data). Ένα σύνολο από αριθμητικά δεδομένα αντικειμένων είναι της μορφής $X = \{x_1, \dots, x_n\} \subset R^s$, όπου για κάθε k , x_k δίνει μετρήσεις από τις διαφορετικά χαρακτηριστικά (όπως ύψος, βάρος, κλπ) για κάθε αντικείμενο o_k . Τα αριθμητικά σχεσιακά δεδομένα περιγράφουν το σύνολο O έμμεσα δίνοντας μετρήσεις διαφοράς (ή ομοιότητας) ανάμεσα σε κάθε ζεύγος αντικειμένων στο O . Συνεπώς τα σχεσιακά δεδομένα μπορούν να αναπαρασταθούν χρησιμοποιώντας ένα πίνακα R , όπου R_{jk} ($1 \leq j, k \leq n$) είναι ο βαθμός ανομοιότητας μεταξύ των αντικειμένων o_j και o_k . Οι βασικές ιδιότητες που θεωρούμε ότι έχει ο πίνακας ανομοιότητας R είναι [HBD96]:

$$\begin{aligned} R_{jk} &\geq 0, \quad 1 \leq j, k \leq n, \\ R_{jk} &= R_{kj}, \quad 1 \leq j, k \leq n, \\ R_{jj} &= 0, \quad 1 \leq j \leq n. \end{aligned}$$

To clustering σε ένα σύνολο αντικειμένων χρησιμοποιώντας σχεσιακά δεδομένα μπορεί να γίνει με διάφορους τρόπους περιλαμβάνοντας γραφο-θεωρητικές μεθόδους καθώς και μεθόδους αντικειμενικών συναρτήσεων. Επίσης έχουν αναπτυχθεί και άλλες μέθοδοι που βασίζονται στην βελτιστοποίηση για αριθμητικά σχεσιακά δεδομένα από τους Ruspinī(1969), Roubens(1978) και Windham(1985). Οι Hathaway, Bezdek, Davenport (1996) έθεσαν τις αρχές για την ανάπτυξη μίας σχεσιακής έκδοσης του c-means αλγορίθμου ώστε να αντιμετωπίστούν προβλήματα clustering που περιλαμβάνουν σχεσιακά δεδομένα. Στην συνέχεια γίνεται μία σύντομη αναφορά στον τρόπο που μπορεί να μετατραπεί ο αλγόριθμος fuzzy-c means ώστε να υποστηρίζει clustering σχεσιακών δεδομένων.

3.3.1 ΑΛΓΟΡΙΘΜΟΣ C-MEANS ΓΙΑ ΣΧΕΣΙΑΚΑ ΔΕΔΟΜΕΝΑ.

Η σχεσιακή έκδοση του αλγορίθμου c-means είναι εφικτή εάν μπορούμε να υπολογίσουμε τις αποστάσεις, που αναφέρθηκαν στο βήμα 2 του αλγορίθμου fuzzy c-means για object data, άμεσα από τα σχεσιακά δεδομένα. Σύμφωνα με τον Hathaway et al [HBD96][HB94] αυτό είναι εφικτό. Συγκεκριμένα, αναφέρεται ότι εάν $X = \{x_1, x_2, \dots, x_n\} \subset R^s$ είναι ένα δεδομένο σύνολο αντικειμένων, μπορούμε να ορίσουμε τον αντίστοιχο πίνακα ανομοιότητας R , με βάση τον τύπο $R_{jk} = \|x_j - x_k\|^2$, για $1 \leq j, k \leq n$. Τότε οι τιμές απόστασης υπολογίζονται στο βήμα 2 του αλγορίθμου διαμέσου των τύπων

$$v_i^{(r)} = \left(\left(U_{i1}^{(r)} \right)^m, \dots, \left(U_{in}^{(r)} \right)^m \right)^T \Big/ \sum_{k=1}^n \left(U_{ik}^{(r)} \right)^m \text{ για } 1 \leq i \leq c, \quad (\text{Εξισ. 3.10})$$

$$(d_{ik}^{(r)})^2 = \left(R v_i^{(r)} \right)_k - \frac{1}{2} \left(v_i^{(r)} \right)^T R \left(v_i^{(r)} \right) \text{ για } 1 \leq i \leq c, 1 \leq k \leq n.$$

Επιπρόσθετα, η ακολουθία των τμημάτων $\{U^{(r)}\}$ που προκύπτουν από την εκτέλεση του αλγορίθμου fuzzy c-means για σχεσιακά δεδομένα είναι η ίδια με την ακολουθία που προκύπτει από την εκτέλεση του βήματος 2 του fuzzy c-means αλγορίθμου που περιγράψαμε παραπάνω, εάν αντικαταστήσουμε τις εξισώσεις (Εξισ. 3.4) με τις (Εξισ. 3.10) και χρησιμοποιήσουμε σχεσιακά δεδομένα.

Σύμφωνα με τα παραπάνω προκύπτει ότι ο πίνακας R υπολογίζεται με βάση την Euclidean απόσταση. Ωστόσο, δεν υπάρχει κανένας περιορισμός ότι ο πίνακας ανομοιότητας R θα βασίζεται πάντα στην Euclidean απόσταση.

Το πρόβλημα του να μην βασίζεται στην Euclidean απόσταση ο πίνακας R , μπορεί να αντιμετωπιστεί εφαρμόζοντας έναν κατάλληλο μετασχηματισμό. Η εκτενής αναφορά για την μετατροπή αυτή γίνεται στην εργασία των Hathaway and Bazdek, 1994 [HB94]. Σύμφωνα λοιπόν με την εργασία αυτή η μετατροπή δεν είναι τίποτα άλλο παρά η προσθήκη ενός θετικού αριθμού β σε όλα τα στοιχεία του R που δεν βρίσκονται στην διαγώνιο. Μπορούμε να δηλώσουμε τον πίνακα που προκύπτει από την μετατροπή ως R_β και οποίος ορίζεται τυπικά ως εξής

$$\left(R_\beta \right)_{jk} = \begin{cases} R_{jk} + \beta, & \text{εαν } j \neq k, \\ 0, & \text{εαν } j = k. \end{cases} \quad (\text{Εξισ. 3.11})$$

Η επιλογή του κατάλληλου β βασίζεται στο παρακάτω θεώρημα το οποίο αναφέρεται από την θεωρία που σχετίζεται με την μετατροπή αυτή.

Θεώρημα. Εάν ο πίνακας R δεν ικανοποιεί την Euclidean απόσταση, τότε υπάρχει ένας θετικός αριθμός β ώστε R_β να είναι Euclidean για όλα τα $\beta > \beta_0$ και να μην είναι Euclidean για όλα τα $\beta < \beta_0$.

Οι προτεινόμενες μέθοδοι για τον υπολογισμό του β χρησιμοποιούν ένα οικονομικό σχήμα το οποίο δυναμικά υπολογίζει μία καλή τιμή για το β στην διάρκεια των επαναλήψεων. Το σχήμα αυτό μπορεί να ερμηνευτεί στην γραμμική άλγεβρα σαν προσέγγιση μίας συγκεκριμένης τιμής του R .

Με βάση τα παραπάνω ο αλγόριθμος fuzzy c-means μπορεί να χρησιμοποιηθεί με τις ακόλουθες αλλαγές:

1. Προσθέτουμε στο βήμα 1 μία αρχικοποίηση για το β , $\beta=0$, και
2. αντικαθιστούμε το βήμα 2 με τα εξής:

Ενημέρωση αποστάσεων. Υπολογίζουμε τους νέους μέσους και τις αποστάσεις ως εξής

$$v_i^{(r)} = \left(\left(U_{i1}^{(r)} \right)^m, \dots, \left(U_{in}^{(r)} \right)^m \right)^T \Bigg/ \sum_{k=1}^n \left(U_{ik}^{(r)} \right)^m \text{ για } 1 \leq i \leq c, \quad (\text{Εξισ. 3.12})$$

$$\left(d_{ik}^{(r)} \right)^2 = \left(R v_i^{(r)} \right)_k - \frac{1}{2} \left(v_i^{(r)} \right)^T R \left(v_i^{(r)} \right) \text{ για } 1 \leq i \leq c, 1 \leq k \leq n.$$

Εάν $\left(d_{ik}^{(r)} \right) < 0$ για i και k τότε υπολογίζουμε

$$\begin{aligned} \Delta\beta &= \max \left\{ -2 \left(d_{ik}^{(r)} \right)^2 / \left(\| v_i^{(r)} - e_k \|^2 \right) \right\}, \text{ μεταβάλλοντας} \\ \left(d_{ik}^{(r)} \right)^2 &\leftarrow \left(d_{ik}^{(r)} \right)^2 + \left(\Delta\beta / 2 \right) \| v_i^{(r)} - e_k \|^2 \text{ για } 1 \leq i \leq c, 1 \leq k \leq n, \\ \beta &\leftarrow \beta + \Delta\beta. \end{aligned} \quad (\text{Εξισ. 3.13})$$

Όπου το e_k δηλώνει την k -οστή στήλη του μοναδιαίου πίνακα $I \in \mathbb{R}^n$

3.3.2 ΠΑΡΑΤΗΡΗΣΕΙΣ ΓΙΑ ΤΟΝ FCM ΓΙΑ ΣΧΕΣΙΑΚΑ ΔΕΔΟΜΕΝΑ

Ένα πολύ ενδιαφέρον πρόβλημα στο οποίο το σχεσιακό c-means clustering είναι πιθανότατα σημαντικό αφορά την τμηματοποίηση ετερογενών δεδομένων, τα οποία μπορούν να περιλαμβάνουν αριθμητικά δεδομένα, κατηγορήματα, διαστήματα κ.α. Χρησιμοποιώντας μία προσέγγιση σχεσιακών δεδομένων, εκείνο που απαιτείται για clustering των ετερογενών δεδομένων είναι ο προσδιορισμός ενός κατάλληλου μέτρου ανομοιότητας μεταξύ των δεδομένων. Μετά τον υπολογισμό της ομοιότητας/ανομοιότητας (similarity/dissimilarity), το σχεσιακό c-means clustering μπορεί να εκτελεστεί χωρίς καμία άλλη περιτλοκή. Αυτό που θα πρέπει να σημειώσουμε στο σημείο αυτό είναι ότι η προσέγγιση του Relational Fuzzy C-Means(RFCM) παράγει τμηματοποίηση των δεδομένων, χωρίς όμως να προσδιορίζει πρότυπα (κέντρα) για τα clusters, τα οποία μπορεί να είναι χρήσιμα για μερικές εφαρμογές.

Επίσης ένας γενικός κανόνας είναι ότι η εκτέλεση του Fuzzy C-Means για αντικείμενα είναι πιο φθηνή υπολογιστικά από την εκτέλεση του αντίστοιχου αλγορίθμου για σχεσιακά δεδομένα. Μία εξαίρεση αποτελεί η εφαρμογή ενός τύπου c-means clustering, χρησιμοποιώντας κάποια άλλη νόρμα εκτός της νόρμας εσωτερικού γινομένου (inner product norm). Αυτό συμβαίνει γιατί στην περίπτωση των δεδομένων αντικειμένων, ο υπολογισμός του εσωτερικού γινομένου αυξάνει σημαντικά την πολυπλοκότητα κατά την διαδικασία της βελτιστοποίησης. Από την άλλη πλευρά, ο υπολογισμός των σχεσιακών δεδομένων από τα αντικείμενα, χρησιμοποιώντας οποιαδήποτε νόρμα είναι μία σχετικά απλή διαδικασία. Βλέπουμε γενικά ότι κάθε μία από τις δύο προσεγγίσεις για clustering σχεσιακών δεδομένων έχει κάποια πλεονεκτήματα και μειονεκτήματα κατά περίπτωση.

3.4 NOISE FUZZY CLUSTERING ΑΛΓΟΡΙΘΜΟΣ

Στις προηγούμενες παραγράφους αναφερθήκαμε στην διαδικασία μετατροπής των fuzzy c-means αλγορίθμων για σχεσιακά δεδομένα. Ένα άλλο βασικό θέμα το οποίο συχνά εμφανίζεται κατά την διαδικασία του clustering και επηρεάζει τα αποτελέσματα του clustering είναι η αντιμετώπιση των outliers (σημεία θορύβου).

Ο fuzzy c-means αλγόριθμος που προτάθηκε από τον Bezdek (1981) δουλεύει καλά για μία ποικιλία εφαρμογών. Ωστόσο, οι περιορισμοί που θέτει FCM (το άθροισμα των βαθμών συμμετοχής ενός στοιχείο στο σύνολο των clusters να είναι ίσο με 1), αναγκάζει ακόμα περισσότερο τα σημεία θορύβου(outlying points) να ομαδοποιούνται σε clusters με την ίδια συνολική συμμετοχή (=1) όπως και τα άλλα σημεία. Αυτό μπορεί να επηρεάζει την ακρίβεια τόσο των κέντρων των clusters όσο και την τελική τμηματοποίηση. Για να αντιμετωπιστεί το πρόβλημα αυτό, ο Dave (1991) πρότεινε να συμπεριληφθεί ο όρος του "cluster-θορύβου(cluster-noise)", ο οποίος θα μπορούσε να χρησιμοποιηθεί για το διαχωρισμό των outliers από τα κεντρικά clusters έτσι ώστε να μην υποβιβάζεται η ποιότητα της cluster ανάλυσης. Βέβαια αυτή η προσέγγιση απαιτεί μία τεχνική κατάλληλη για εύρεση των outliers.

Με βάση τον αλγόριθμο noise clustering του Dave(1991), μπορούμε να θεωρήσουμε ότι τα δεδομένα του πίνακα R έχουν c "πραγματικά" clusters συν ένα cluster θορύβου. Η αντικειμενική συνάρτηση της fuzzy clustering μεθόδου θα έχει την μορφή

$$J_m(U, V) = \sum_{i=1}^c \sum_{k=1}^n U_{ik}^m \|x_k - v_i\|^2 + \sum_{k=1}^n U_{c+1}^m \delta^2 \quad (\text{Εξισ. 3.14})$$

όπου δ^2 αναπαριστά την τετραγωνική "απόσταση" κάθε στοιχείου από το cluster θορύβου. Με βάση λοιπόν τα παραπάνω μπορούμε να διατυπώσουμε μία ισοδύναμη έκδοση του fuzzy clustering αλγορίθμου όπως παρουσιάζεται στην συνέχεια.

Αλγόριθμος noise clustering

Ο αλγόριθμος noise clustering όπως προτάθηκε από τον Dave (1991) αποτελείται από τα ακόλουθα βήματα:

Βήμα 1. Καθορισμός του αριθμού των (πραγματικών) clusters c , του συντελεστή ασάφειας $m > 1$, του κριτηρίου σταματήματος ϵ και της απόστασης του noise cluster δ . Επίσης ορίζουμε μία αρχική τμηματοποίηση $U^{(0)} \in M_{fc+ln}$.

Στην συνέχεια για κάθε επανάληψη r , $r=1, \dots, r_{max}$.

Βήμα 2. Υπολογισμός των νέων αποστάσεων για $i = 1, \dots, c$ σύμφωνα με το Βήμα 2 της παραγράφου 3.2.1. Θέτουμε $d''_{c+1k} = \delta$, για $k = 1, \dots, n$.

Βήμα 3. Υπολογισμός των νέων βαθμών συμμετοχής μέσω των τύπων του αναφέρθηκαν στην παράγραφο 3.2.1(Βήμα 3), αντικαθιστώντας το c με $c+1$.

Βήμα 4.

```

If |U(r+1) - U(r)| < ε then
    Stop
Else
{
    Go to Βήμα 2,
    r = r + 1
}

```

Η έκδοση του παραπάνω αλγορίθμου για σχεσιακά δεδομένα μπορεί να προκύψει εύκολα υιοθετώντας τις εξής τροποποιήσεις: 1) προσθέτουμε στο βήμα 1 την αρχικοποίηση για την παράμετρο $\beta=0$, και 2) χρησιμοποιούμε για τον υπολογισμό της απόστασης την εξίσωση (Εξισ. 3.10) της παραγράφου 3.3.1.

3.5 CONDITIONAL FUZZY C-MEANS CLUSTERING

Σύμφωνα με τις μεθόδους clustering τα πρότυπα ανάλογα με την εγγύτητα που παρουσιάζουν μεταξύ τους στο χώρο των χαρακτηριστικών ταξινομούνται σε κάποια clusters. Εάν σε σχέση με κάθε πρότυπο (pattern) θεωρήσουμε και μία βοηθητική μεταβλητή τότε τα πρότυπα μπορούν να δομηθούν στις διάφορες κατηγορίες όχι μόνο με βάση την εγγύτητα τους στο χώρο των χαρακτηριστικών αλλά και με βάση τις τιμές της μεταβλητής που χρησιμοποιείται στα δεδομένα. Το clustering που περιλαμβάνει τόσο τα χαρακτηριστικά όσο και την βοηθητική μεταβλητή καλείται *conditional clustering*[Pedrycz95]. Το *conditional clustering* διακρίνεται ανάμεσα στις άλλες τεχνικές του data mining και μπορεί να θεωρηθεί σαν μία γλωσσολογικά προσανατολισμένη(linguistically oriented) αναζήτηση για εξαρτήσεις σε σύνολα δεδομένων πολλαπλών μεταβλητών.

Υποθέτουμε ότι έχουμε ένα σύνολο από δεδομένα $X=\{x_1, x_2, \dots, x_n\}$ τα οποία θέλουμε να διαιρέσουμε σε c clusters και $U \in M_{fcn}$ είναι η μήτρα συμμετοχής. Επίσης θεωρούμε ότι για κάθε δείγμα x_k ($1 \leq k \leq N$) ορίζονται κάποιες μεταβλητές συνθήκης f_1, f_2, \dots, f_N αντίστοιχα. Η τιμή f_k περιγράφει το επίπεδο ανάμιξης του x_k στα clusters που κατασκευάζονται. Η συσχέτιση του f_k με τις υπολογιζόμενες τιμές συμμετοχής του x_k , $U_{1k}, U_{2k}, \dots, U_{ck}$ μπορεί να γίνει με τους εξής τρόπους:

- Το f_k μπορεί να υπολογιστεί ως το άθροισμα των εισόδων της k -οστής στήλης της συνάρτησης συμμετοχής, δηλαδή

$$\sum_{i=1}^c U_{ik} = f_k, \quad k = 1, 2, \dots, N \quad (\text{Εξισ. 3.15})$$

- Το f_k μπορεί να θεωρηθεί ως η μέγιστη τιμή των συναρτήσεων συμμετοχής της αντίστοιχης στήλης

$$\max_{i=1 \dots c} U_{ik} = f_k \quad (\text{Εξισ. 3.16})$$

Έτσι η οικογένεια των πινάκων συμμετοχής είναι

$$\begin{aligned} \mathbf{U} = \{U_{ik} \in [0,1] \mid & \sum_{i=1}^c U_{ik} = f_k \quad \forall k| \\ & 0 < \sum_{k=1}^N U_{ik} < N \quad \forall i\} \end{aligned} \quad (\text{Εξισ. 3.17})$$

Λαμβάνοντας το τοπικό ελάχιστο της αντικειμενικής συνάρτησης μπορούμε όπως και προηγούμενα να υπολογίσουμε τον πίνακα συμμετοχής σύμφωνα με την εξίσωση

$$U_{ik} = f_k \left/ \sum_{j=1}^c \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{2/(p-1)} \right. \quad (\text{Εξισ. 3.18})$$

Θα πρέπει να επισημάνουμε ότι στην περίπτωση του conditional c-means clustering οι τιμές των συναρτήσεων συμμετοχής δεν έχουν άθροισμα 1.

Η μέθοδος μπορεί να επεκταθεί σε διάφορες μεταβλητές συνθήκης, όπως για παράδειγμα κάποιους λεκτικούς όρους. Ορίζοντας κάποιους λεκτικούς όρους για κάθε μεταβλητή αναπτύσσουμε μία σύνθετη λογική φόρμουλα που χρησιμοποιείται για να οδηγήσουμε τον αλγόριθμο clustering. Για παράδειγμα εάν θέλουμε να ορίσουμε ομάδες για το σύνολο των δεδομένων X λαμβάνοντας υπόψη την λογική έκφραση

$$(f \text{ and } g) \text{ or } h \text{ (π.χ. } (y_1 \text{ is } small \text{ and } y_2 \text{ is } large) \text{ or } y_3 \text{ is } medium))$$

κατά την βελτιστοποίηση της αντικειμενικής συνάρτησης που αναφέραμε κατά την περιγραφή του βασικού Fuzzy C-Means αλγορίθμου, θα πρέπει να λάβουμε υπόψη μας και τους περιορισμούς f, g, h . Οπότε έχουμε

$$\min_{U, u_1, u_2, \dots, u_N} J_m(U, V)$$

υπό την προϋπόθεση ότι $U \in U(f, g, h)$.

Η οικογένεια των πινάκων συμμετοχής στην οποία ανήκει το U , ορίζεται ως εξής

$$\begin{aligned} \mathbf{U}(f, g, h) = \{U_{ik} \in [0,1] \mid & \sum_{i=1}^c U_{ik} = L(f_k, g_k, h_k) \quad \forall k| \\ & 0 < \sum_{k=1}^N U_{ik} < N \quad \forall i\} \end{aligned} \quad (\text{Εξισ. 3.19})$$

όπου L είναι η λογική έκφραση των f, g, h fuzzy sets.

Θα πρέπει τέλος να σημειωθεί ότι η προσέγγιση για το conditional clustering που προτάθηκε από τον Pedrycz και την οποία περιγράψαμε παραπάνω, είναι μία γενική προσέγγιση. Αυτό σημαίνει ότι και άλλες προσεγγίσεις όπως ο fuzzy c-lines, ο relational Fuzzy C-Means κλπ μπορούν να συνδυαστούν με την προσέγγιση του conditional clustering, λαμβάνοντας υπόψη κάποιες υποσυνθήκη μεταβλητές στην διαδικασία του clustering.

4^ο ΚΕΦΑΛΑΙΟ

ΑΞΙΟΛΟΓΗΣΗ ΠΟΙΟΤΗΤΑΣ CLUSTERING

4.1 ΕΙΣΑΓΩΓΗ

Ο αντικειμενικός σκοπός των περισσοτέρων μεθόδων clustering είναι η παροχή χρήσιμης πληροφορίας ομαδοποιώντας τα στοιχεία ενός συνόλου δεδομένων σε clusters. Τα μέλη κάθε cluster θα πρέπει να παρουσιάζουν κάποια ομοιότητα μεταξύ τους ενώ τα μέλη διαφορετικών clusters θα πρέπει να διαφέρουν όσο το δυνατό περισσότερο. Η εκτίμηση της ομοιότητας ή διαφοροποίησης μεταξύ των στοιχείων μέσα στο ίδιο cluster ή μεταξύ διαφορετικών clusters πραγματοποιείται με την βοήθεια κάποιων μέτρων ομοιότητας/απόστασης.

Στο κεφάλαιο αυτό θα αναφερθούμε σε μερικά από τα βασικά μέτρα αξιολόγησης των clusters καθώς και στις βασικές τεχνικές χρήσης των μέτρων αυτών για την εκτίμηση της ποιότητας των clusters που προκύπτουν από μία διαδικασία clustering. Ο βασικός στόχος χρήσης των μέτρων ποιότητας είναι η επιλογή εκείνου του σχήματος clustering που θα μας δώσει την καλύτερη τμηματοποίηση των δεδομένων για την εφαρμογή μας.

4.2 ΜΕΤΡΑ ΠΟΙΟΤΗΤΑΣ CLUSTERING

Ο κύριος στόχος μας κατά την εφαρμογή μίας προσέγγισης clustering είναι να προσδιορίσουμε ποια είναι τα καλύτερα clusters για ένα σύνολο δεδομένων, δηλαδή τα clusters των οποίων τα μέλη να έχουν μεγάλο βαθμό ομοιότητας. Εάν αναπαραστήσουμε τα δεδομένα μας στο χώρο μπορούμε εύκολα να κατανοήσουμε ότι παρόμοια θα είναι τα δεδομένα τα οποία βρίσκονται κοντά το ένα στο άλλο. Ενώ καλά διαχωρισμένα είναι τα clusters των οποίων τα στοιχεία απέχουν περισσότερο. Συνεπώς αυτό που χρειαζόμαστε είναι μια τμηματοποίηση των δεδομένων μας με τρόπο ώστε [Begety 97]:

- τα μέλη κάθε cluster να είναι όσο το δυνατόν πιο κοντά το ένα στο άλλο και
- τα clusters να απέχουν μεταξύ τους.

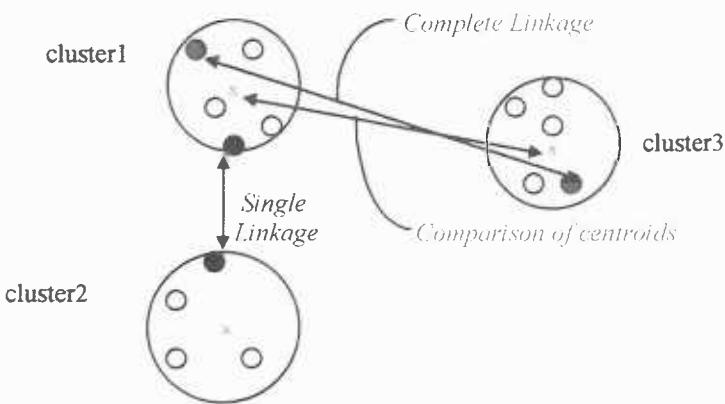
Ένα κοινό μέτρο για το πρώτο κριτήριο (ομοιότητα μέσα σε ένα cluster) είναι η διακύμανση (variance), δηλαδή το άθροισμα των τετραγωνικών διαφορών κάθε αντικειμένου από το κέντρο του cluster στο οποίο ανήκει (συνήθως λαμβάνουμε ως

κέντρο το μέσο όρο των στοιχείων που ανήκουν στο cluster). Ο στόχος μας είναι η ελαχιστοποίηση της διακύμανσης.

Για το δεύτερο κριτήριο (απόσταση μεταξύ διαφορετικών clusters) μετράμε την απόσταση μεταξύ των διαφορετικών clusters που έχουν προκύψει, την οποία θέλουμε να μεγιστοποιήσουμε. Υπάρχουν τρεις προσεγγίσεις που μπορούν να χρησιμοποιηθούν για την εκτίμηση της απόστασης μεταξύ δύο clusters:

- **Single linkage** : Μετρά την απόσταση μεταξύ των πιο κοντινών στοιχείων που ανήκουν στα clusters. Αυτή η μέθοδος παράγει clusters με την ιδιότητα ότι κάθε μέλος ενός cluster είναι περισσότερο στενά συσχετισμένο με τα μέλη του cluster που ανήκει παρά με οποιοδήποτε στοιχείο εκτός του cluster.
- **Complete linkage** : Μετρά την απόσταση μεταξύ των πιο απομακρυσμένων στοιχείων των clusters. Η μέθοδος αυτή παράγει clusters με την ιδιότητα ότι όλα τα μέλη των clusters απέχουν μεταξύ τους κάποια γνωστή μέγιστη απόσταση.
- **Comparison of centroids** : Μετρά την απόσταση μεταξύ των κέντρων των clusters. Το κέντρο ενός cluster είναι συνήθως ο μέσος όρος των στοιχείων του cluster.

Το σχήμα 4.1 αναπαριστά διαγραμματικά τις τρεις προσεγγίσεις εκτίμησης της απόστασης μεταξύ των clusters.



Σχήμα 4.1. Προσεγγίσεις μέτρησης της απόστασης μεταξύ των clusters.

Η επιλογή των μέτρων που θα χρησιμοποιηθούν για την εκτίμησης της ποιότητας συνήθως εξαρτάται και από την προσέγγιση clustering που χρησιμοποιούμε. Θα πρέπει λοιπόν τα μέτρα ποιότητας που χρησιμοποιούμε να λαμβάνουν υπόψη τους τον τρόπο με τον οποίο προσδιορίζονται τα clusters κατά την διαδικασία clustering. Γενικά, ένα μέτρο εκτίμησης ποιότητας clusters θα πρέπει να λαμβάνει το μέτρο ομοιότητας ή το μέτρο απόστασης που χρησιμοποιείται κατά την διαδικασία δημιουργίας των clusters και να το χρησιμοποιεί για να αντιπαραβάλλει την μέση απόσταση μέσα στα clusters (εκτίμησης πυκνότητας clusters) με την μέση απόσταση μεταξύ των clusters.

4.3 ΜΕΤΡΟ ΠΟΙΟΤΗΤΑΣ ΓΙΑ CRISP CLUSTERING

Ένα κριτήριο αξιολόγησης του clustering θα πρέπει να λαμβάνει υπόψη με βάση όσα αναφέραμε παραπάνω την απόσταση μεταξύ των στοιχείων μέσα σε κάθε cluster καθώς και την απόσταση μεταξύ των clusters. Θα πρέπει δηλαδή να μετρά την πυκνότητα και διαφοροποίηση των clusters που προκύπτουν από την εφαρμογή μίας διαδικασίας clustering. Στην συνέχεια θα περιγράψουμε κάποιες τέτοιες συναρτήσεις ποιότητας clustering.

4.3.1 ΔΕΙΚΤΗΣ ΔΙΑΧΩΡΙΣΜΟΥ (SEPARATION INDEX)

Ένα καλά ορισμένο κριτήριο αξιολόγησης για crisp clustering είναι ο δείκτης διαχωρισμού (separation index) που προτάθηκε από τον Dunn και ο οποίος προσδιορίζει "πυκνά, καλά διαχωρισμένα" ("compact, separate" - CS) clusters και ορίζεται ως εξής [XB91]:

$$D_1 = \min_{1 \leq i \leq c} \left\{ \min_{i+1 \leq j \leq c-1} \left\{ \frac{\text{dis}(v_i, v_j)}{\max_{1 \leq k \leq c} \{ \text{dia}(v_k) \}} \right\} \right\} \quad (\text{Εξισ. 4.1})$$

Όπου

$$\text{dia}(v_k) = \max_{X_i, X_j \in v_k} d(X_i, X_j) \quad (\text{Εξισ. 4.2})$$

$$\text{dis}(v_i, v_j) = \min_{X_i \in v_i, X_j \in v_j} d(X_i, X_j)$$

όπου d είναι μέτρο απόστασης στο \mathbb{R}^p . Το CS clustering του X ορίζεται ως εξής

$$\max_{2 \leq c \leq n} \left\{ \max_{\Omega_c} D_1 \right\} \quad (\text{Εξισ. 4.3})$$

όπου Ω_c δηλώνει τους βέλτιστους υποψηφίους για συγκεκριμένο c . Το κύριο πρόβλημα με την άμεση εφαρμογή αυτού του μέτρου αξιολόγησης είναι η υπολογιστική του πολυπλοκότητα καθώς ο υπολογισμός τους D_1 γίνεται απαγορευτικός με την αύξηση του αριθμού των clusters c και του αριθμού των στοιχείων n .

Ένα άλλο κριτήριο αξιολόγησης το οποίο επίσης μετρά την πυκνότητα και τον διαχωρισμό των clusters εισάγεται από τους Davies και Bouldin. Η βασική διαφορά του μέτρου αυτού από το D_1 είναι ότι λαμβάνει υπόψη την τη μέση περίπτωση χρησιμοποιώντας το μέσο σφάλμα για κάθε κλάση.

4.4 ΑΞΙΟΛΟΓΗΣΗ FUZZY CLUSTERING

Η εφαρμογή κάποιας clustering προσέγγισης ορίζει για όλα τα δεδομένα ενός συνόλου σε ποιο cluster ανήκουν. Στην περίπτωση που τα όρια μεταξύ των τμημάτων είναι συγκεκριμένα έχουμε το λεγόμενο *crisp clustering*. Αντίθετα, όταν εφαρμόζουμε το *fuzzy clustering* τα όρια μεταξύ των τμημάτων που παράγονται είναι ασαφή. Αυτό σημαίνει ότι κάθε δείγμα από τα δεδομένα ενός ασαφούς τμήματος ανήκει σε διαφορετικές κλάσεις με διαφορετικούς βαθμούς συμμετοχής.

Καθώς όμως οι αλγόριθμοι fuzzy clustering χαρακτηρίζονται ως unsupervised, τα τελικά τμήματα των δεδομένων που προκύπτουν θα πρέπει να αξιολογηθούν ως προς την εγκυρότητα τους. Γενικά, η εκτίμηση της εγκυρότητας των clusters μπορεί να απαντήσει μεταξύ άλλων και σε ερωτήσεις που αφορούν το πόσο καλά είναι τα τμήματα στα οποία διαιρέθηκαν τα δεδομένα, εάν υπάρχει κάποια καλύτερη τμηματοποίηση που μπορεί να εφαρμοστεί, κλπ. Επίσης εάν ο αριθμός των τμημάτων δεν είναι εκ των προτέρων γνωστός ένας δείκτης εγκυρότητας (validation index) μπορεί να βοηθήσει στο να επιλεχθεί ο καλύτερος δυνατός αριθμός κλάσεων.

Γενικά, το δίλημμα του να αποφασίσουμε τον αριθμό των clusters καθώς και να εκτιμήσουμε την δομή κάθε τμήματος που προκύπτει, υπεισέρχεται σε αυτό που καλούμε *εγκυρότητα cluster (cluster validity)*. Τα βασικά βήματα για να επιτευχθεί αυτό είναι τα εξής:

1. Όλες οι παράμετροι για την μέθοδο clustering είναι καθορισμένες εκτός από τον αριθμό των clusters.
2. Διαφοροποιώντας τον αριθμό των clusters μεταξύ του 2 και μίας μεγαλύτερης τιμής c_{max} και εφαρμόζοντας έναν αλγόριθμο clustering, για κάθε αριθμό clusters $c_i \in \{2, 3, \dots, c_{max}\}$ βρίσκουμε μία διαφορετική τμηματοποίηση για τα δεδομένα.
3. Εφαρμόζουμε ένα δείκτη εγκυρότητας για κάθε τμήμα, που λαμβάνεται από το βήμα 2 προκειμένου να ορίσουμε μία τιμή για την εκτίμηση της εγκυρότητας. Ο πραγματικός αριθμός των κλάσεων των δεδομένων μπορεί να καθοριστεί με βάση την ακραία τιμή των δεικτών εγκυρότητας για όλα τα clusters c_i .

Η κλασική προσέγγιση στην εκτίμηση της ποιότητας των clusters για fuzzy clustering βασίζεται στην άμεση αξιολόγηση των fuzzy c-τμημάτων.

4.4.1 ΚΛΑΣΙΚΕΣ ΤΕΧΝΙΚΕΣ ΕΚΤΙΜΗΣΗΣ ΠΟΙΟΤΗΤΑΣ CLUSTERING ΓΙΑ TON FUZZY C-MEANS

Ο αλγόριθμος FCM (Fuzzy C-Means) μπορεί να τμηματοποιήσει ένα σύνολο δεδομένων για ένα καθορισμένο αριθμό clusters. Ένας από τους αντικειμενικούς σκοπούς όμως μίας διαδικασίας εκτίμησης της ποιότητας των clusters είναι να καθοριστεί αυτόματα ο βέλτιστος αριθμός clusters. Αυτό είναι επιθυμητό, για παράδειγμα, όταν ο FCM χρησιμοποιείται για clustering εικόνων καθώς ο αριθμός των clusters στην εικόνα δεν είναι γνωστός. Η εκτίμηση της εγκυρότητας ενός παραγόμενου από τον FCM ασαφούς τμήματος μπορεί να επιτευχθεί με την βοήθεια ενός δείκτη εγκυρότητας cluster (cluster validity index). Εάν ορίσουμε τον ελάχιστο και μέγιστο αριθμό clusters ως c_{min} και c_{max} αντίστοιχα, τότε για κάθε αριθμό clusters

$c \in [c_{\min}, c_{\max}]$ μπορεί να προκύψει μία διαφορετική τμηματοποίηση των δεδομένων με την εφαρμογή του αλγορίθμου FCM. Στην συνέχεια μπορούμε να υπολογίσουμε για κάθε τιμή του c τον δείκτη εγκυρότητας των clusters. Συγκρίνοντας όλες τις τιμές των δεικτών εγκυρότητας για όλους τους δυνατούς αριθμούς clusters, μπορούμε να προσδιορίσουμε ποιος είναι ο βέλτιστος αριθμός clusters.

Διάφορες τεχνικές εγκυρότητας έχουν διατυπωθεί για το fuzzy c-means αλγόριθμο μέχρι σήμερα (Bezdek(1974), Windham (1981), Backer και Jain(1981), Xie και Beni (1991) κλπ.). Στην συνέχεια θα γίνει μία σύντομη αναφορά στα κυριότερα από τα μέτρα που έχουν διατυπωθεί για εκτίμηση της ποιότητας των clusters[Dave96][RR98].

4.4.1.1 PARTITION COEFFICIENT (PC)

Ο Bezdek (1974) όρισε τον συντελεστή συμμετοχής (partition coefficient -PC) ως ένα μέτρο εκτίμησης της ποιότητας των clusters. Εάν $U \in M_{fc}$ είναι μία fuzzy c-τμηματοποίηση n δεδομένων, ο συντελεστής συμμετοχής δίνεται από την εξίσωση (Εξισ. 4.4)

$$V_{PC}(U; c) = \sum_{k=1}^n \sum_{i=1}^c (U_{ik})^2 / n \quad (\text{Εξισ. 4.4})$$

όπου το $V_{PC}(U ; c) \in [1/c, 1]$, γεγονός που δείχνει την εξάρτηση του μέτρου από τον αριθμό των clusters. Ωστόσο, ένας απλός γραμμικός μετασχηματισμός μπορεί να εφαρμοστεί σε αυτό το μέτρο ώστε να καταστήσει τα αποτελέσματα ανεξάρτητα από τον αριθμό των clusters, c . Τα μειονεκτήματα του συντελεστή συμμετοχής, όπως προσδιορίστηκαν από τον Bezdek (1974), είναι η μονότονη εξάρτηση του μέτρου από τον αριθμό των clusters και η έλλειψη άμεσης σύνδεσης με κάποιες ιδιότητες των ίδιων των δεδομένων και της γεωμετρίας των clusters.

4.4.1.2 ENTRÓPIΑ ΤΜΗΜΑΤΟΣ(PARTITION ENTROPY - PE)

Η έννοια της συσχέτισης με βάση την εντροπία έχει συζητηθεί στην στατιστική και έχει εφαρμοστεί στα ασαφή σύνολα. Ένα σχήμα που χρησιμοποιεί ένα μέτρο εντροπίας βασιζόμενο στα ασαφή σύνολα θα πρέπει να αποκτήσει την ελάχιστη τιμή του μέτρου για μία καλή τμηματοποίηση των δεδομένων. Ο Bezdek (1981) όρισε την *partition entropy* ενός fuzzy c - τμήματος U ως εξής

$$V_{PE}(U; c) = - \sum_{k=1}^n \sum_{i=1}^c (U_{ik}) \log_a (U_{ik}) / n \quad (\text{Εξισ. 4.5})$$

όπου η λογαριθμική βάση $a \in (1, \infty)$.

Σύμφωνα με τον Bezdek(1981) ο περιορισμός που θέτει η partition entropy μπορεί να αποδοθεί στο γεγονός ότι βασίζεται μόνο στους βαθμούς συμμετοχής ενώ δεν λαμβάνει υπόψη καθόλου την γεωμετρία των κέντρων των clusters.

4.4.1.3 ΣΥΝΑΡΤΗΣΗ ΑΞΙΟΛΟΓΗΣΗΣ FUZZY CLUSTERING COMPACTNESS AND SEPARATION

Οι Xie και Beni [XB91] εισάγουν έναν νέο δείκτη εκτίμησης εγκυρότητας ο οποίος λαμβάνει υπόψη του τόσο την πυκνότητα μέσα στα clusters όσο και την απόσταση μεταξύ των διαφορετικών clusters. Ειδικότερα, θεωρώντας ένα fuzzy clustering ενός συνόλου δεδομένων $X = \{x_j: j=1, \dots, n\}$ σε c clusters, με $v_i (i=1, 2, 3, \dots, n)$ τα κέντρα των clusters και $U_{ij} (i=1, 2, \dots, n, j=1, 2, \dots, n)$ οι βαθμοί συμμετοχής του στοιχείου j στο cluster i , οι Xie και Beni όρισαν ένα δείκτη αξιολόγησης clustering με βάση τους παρακάτω ορισμούς:

- Ορίζουμε σαν ασαφή απόκλιση του x_j από το cluster i την απόσταση

$$d_{ij} = U_{ij} \|x_j - v_i\| \quad (\text{Εξισ. 4.6})$$

όπου $\|\dots\|$ δηλώνουμε την Ευκλείδεια απόσταση.

- Ο ασαφής αριθμός των διανυσμάτων μέσα σε ένα ασαφές cluster ορίζεται ως το άθροισμα των βαθμών συμμετοχής των στοιχείων στο συγκεκριμένο cluster.

$$n_i = \sum_j U_{ij} \quad (\text{Εξισ. 4.7})$$

- Για κάθε cluster i , ορίζουμε την διακύμανση του cluster i ως το άθροισμα των τετραγώνων της ασαφούς απόκλισης κάθε στοιχείου. Δηλαδή

$$\sigma_i = \sum_{i,j} (d_{ij})^2 \quad (\text{Εξισ. 4.8})$$

Ενώ με $\sigma = \sum_i \sigma_i$ δηλώνουμε τη συνολική διακύμανση του συνόλου των δεδομένων μας.

Το σ_i και σ εξαρτώνται από το σύνολο των δεδομένων, αλλά κυρίως από την ασαφή τμηματοποίηση δηλ. από τους βαθμούς συμμετοχής U_{ij} και v_i .

Έτσι εάν εφαρμόσουμε το fuzzy c-means με $m=2$ τότε η τιμή της διακύμανσης θα ισούται με την αντικειμενική συνάρτηση J .

- Ο λόγος της συνολικής διακύμανσης ως προς το μέγεθος του συνόλου δεδομένων, ορίζει την συνολική πυκνότητα π της ασαφούς τμηματοποίησης του συνόλου δεδομένων.

$$\pi = \sigma/n \quad (\text{Εξισ. 4.9})$$

Η τιμή του π δηλώνει πόσο συμπαγές είναι κάθε cluster. Ειδικότερα, όσο μικρότερη είναι η τιμή του π τόσο πιο πυκνά είναι τα clusters. Το π είναι συνάρτηση της κατανομής των χαρακτηριστικών του ίδιου του συνόλου και επιτρόσθετα η συνάρτηση του πως κατανέμουμε τα δεδομένα σε clusters. Ωστόσο, είναι ανεξάρτητο από τον αριθμό των σημείων. Για ένα δεδομένο σύνολο, μία μικρή τιμή για το π υποδηλώνει ότι έχουμε πετύχει μία καλή τμηματοποίηση δηλαδή μία τμηματοποίηση με αρκετά πυκνά clusters.

- Η ποσότητα $\pi_i = \sigma_i / n_i$ καλείται πυκνότητα του cluster i .

Καθώς n_i είναι ο αριθμός των διανυσμάτων σε ένα cluster i, s_i/n_i θα είναι η μέση διακύμανση στο cluster i. Εχοντας ορίσει το πι, έχουμε την δυνατότητα να ορίσουμε την πυκνότητα μίας ασαφούς τμηματοποίησης ως $\pi = \sum_i (\pi_i)/c$ δηλ. η μέση πυκνότητα ανά cluster ή ως $\pi = \max_i \pi_i$ δηλ. η χειρότερη περίπτωση. Αποδεικνύεται ότι και οι δύο αυτοί εναλλακτικοί ορισμοί έχουν τα ίδια αποτελέσματα με τον προηγούμενο ορισμό του $\pi = \sigma/n$.

- Διαφοροποίηση μεταξύ των fuzzy clusters ονομάζουμε την ελάχιστη απόσταση μεταξύ των κέντρων των clusters, δηλαδή

$$s = (d_{\min})^2 \quad (\text{Εξισ. 4.10})$$

$$d_{\min} = \min_{i,j} \|v_i - v_j\|$$

Μία μεγάλη τιμή του s δηλώνει την ύπαρξη καλά διαχωρισμένων clusters.

Με βάση τους παραπάνω ορισμούς [(Εξισ. 4.9), (Εξισ. 4.10)] η συνάρτηση ποιότητας *compactness and separation* μπορεί να οριστεί ως εξής

$$S = \pi / s = s / n(d_{\min})^2 \quad (\text{Εξισ. 4.11})$$

Μία μικρή τιμή για το S υποδηλώνει μία τμηματοποίηση στην οποία όλα τα clusters είναι πυκνά και καλά διαχωρισμένα μεταξύ τους.

Για τους αλγορίθμους FCM (fuzzy c-means), το μέτρο αυτό είναι απλά η αντικειμενική συνάρτηση διατρούμενη από τον αριθμό των δεδομένων και την ελάχιστη απόσταση μεταξύ των κέντρων των cluster.

$$V_{XB}(U; V; X) = \frac{\sum_{i=1}^c \sum_{k=1}^n (U_{ik})^m \|x_k - v_i\|^2}{n(\min\{v_i - v_j\})} \quad (\text{Εξισ. 4.12})$$

Το μέτρο αυτό συνδυάζει την ιδέα της πυκνότητας(compactness) και διαχωρισμού(separation) των clusters και εξ ορισμού στοχεύει στο να εργάζεται μόνο με συμπαγή και καλά διαχωρισμένα clusters. Επιπρόσθετα, η ιδέα πίσω από αυτό το μέτρο είναι να δούμε πόσο καλά διαχωρισμένα είναι τα clusters.

4.4.1.4 ΆΛΛΑ ΚΛΑΣΙΚΑ ΜΕΤΡΑ

Οι Backer και Jain(1981) αντιμετωπίζουν το πρόβλημα εκτίμησης της εγκυρότητας διαμέσου ενός μέτρου που βασίζεται στην αποσύνθεση των ασαφών συνόλων. Η μέθοδος αυτή χρησιμοποιεί την θεωρία ασαφών συνόλων για να εκτιμήσει την απόσταση μεταξύ των fuzzy clusters. Ειδικότερα, σχετίζει την δυνατότητα διαχωρισμού (separability) των δεδομένων και σαν συνέπεια το πόσο "καλό" είναι ένα τμήμα, με το μέγεθος της ασάφειας στα κενά διαστήματα ανάμεσα στα ασαφή σύνολα. Το μέτρο αυτό όπως και το PC μέτρο, χρησιμοποιεί το ασαφές αλγεβρικό άθροισμα, αλλά η διατύπωση του έχει σαν αποτέλεσμα την μείωση της εξάρτησης

του από τον αριθμό των clusters, *c*. Αυτό έχει σαν αποτέλεσμα την διακύμανση του μέτρου αυτού μεταξύ του 0 και 1, γεγονός που θα πρέπει να δίνει την δυνατότητα για καλύτερη διάκριση οποιασδήποτε λεπτής διαφοράς μεταξύ των τμημάτων. Έτσι έχει πλεονέκτημα έναντι του συντελεστή συμμετοχής(ΠC).

Οι Fukuyama και Sugeno(1989) στην προσπάθεια τους να ορίσουν έναν δείκτη εκτίμησης εγκυρότητας των clusters ο οποίος θα βασιζόταν στα ίδια τα δεδομένα και δεν θα είχε τα μειονεκτήματα των μέτρων V_{PC} και V_{PE} δημιουργησαν έναν δείκτη ο οποίος χρησιμοποιεί κατά τον υπολογισμό του τόσο το σύνολο των δεδομένων όσο και τα πρότυπα των clusters(κέντρα των clusters). Ο δείκτης αυτός δίνεται από την εξίσωση:

$$V_{FS,m}(U;V;X) = \sum_{i=1}^c \sum_{k=1}^n (U_{ik})^m \left(\|x_k - v_i\|^2 - \|v_i - \bar{v}\|^2 \right) \quad (\text{Εξισ. 4.13})$$

όπου \bar{v} είναι ο συνολικός μέσος όλων των δεδομένων x_k .

Στον Πίνακα 4.1 συνοψίζονται τα βασικότερα μέτρα ποιότητας *fuzzy clustering*.

Validity Index	Functional Index	Βέλτιστος αριθμός cluster
Partition Coefficient	$V_{PC}(U;c) = \sum_{k=1}^n \sum_{i=1}^c (U_{ik})^2 / n$	Max{ $V_{PC}(U;c;m)$ }
Partition Entropy	$V_{PE}(U;c) = - \sum_{k=1}^n \sum_{i=1}^c (U_{ik}) \log_a (U_{ik}) / n$	Min{ $V_{PE}(U;c,m)$ }
Fukuyama	$V_{FS,m}(U;V;X) = \sum_{i=1}^c \sum_{k=1}^n (U_{ik})^m \left(\ x_k - v_i\ ^2 - \ v_i - \bar{v}\ ^2 \right)$	Min($V_{FS,m}(U;c,m)$)
and Sugeno		
Xie and Beni	$V_{XB}(U;V;X) = \frac{\sum_{i=1}^c \sum_{k=1}^n (U_{ik})^m \ x_k - v_i\ ^2}{n \left(\min \{ v_i - v_j \} \right)}$	Min{ $V_{XB}(U;c,m)$ }

Πίνακας 4.1. Κλασικά μέτρα Ποιότητας *Fuzzy Clustering*

4.4.2 ΜΕΤΡΟ ΕΚΤΙΜΗΣΗΣ ΤΗΣ ΠΥΚΝΟΤΗΤΑΣ (COMPACTNESS) ΚΑΙ ΤΟΥ ΔΙΑΧΩΡΙΣΜΟΥ (SEPARATION) ΤΩΝ FUZZY C-ΤΜΗΜΑΤΩΝ

Η πραγματική εκτίμηση της εγκυρότητας για το FCM πρέπει να λαμβάνει υπόψη της τόσο την πυκνότητα μέσα στα ίδια τα τμήματα που προκύπτουν όσο και το διαχωρισμό μεταξύ των διαφορετικών fuzzy c-tμημάτων. Μία καλή τμηματοποίηση των δεδομένων απαιτεί μέγιστη πυκνότητα για κάθε cluster(τμήμα) που προκύπτει ενώ η απόσταση μεταξύ των clusters να είναι όσο το δυνατόν μεγαλύτερη. Εάν μία συνάρτηση εγκυρότητας λάβει υπόψη της μόνο την απαίτηση για την πυκνότητα, τότε η καλύτερη τμηματοποίηση λαμβάνεται όταν κάθε στοιχείο του συνόλου των δεδομένων λαμβάνεται σαν ξεχωριστό cluster. Αντίθετα, εάν το κριτήριο στο οποίο βασίζεται η συνάρτηση εγκυρότητας είναι η απόσταση μεταξύ των clusters, τότε το καλύτερο τμήμα θα είναι το ίδιο το σύνολο των δεδομένων, καθώς η απόσταση μεταξύ του μοναδικού cluster(του συνόλου δεδομένων) και του εαυτού του είναι μηδέν. Συνεπώς μία αξιόπιστη συνάρτηση εγκυρότητας θα πρέπει να λαμβάνει υπόψη της και τα δύο κριτήρια και θα είναι βέλτιστη για την τμηματοποίηση που θα συνδυάζει βέλτιστη τιμή και για τα δύο κριτήρια.

Οι Rezaee, Lelieveldt, Reiber [RR98] στην προσπάθεια τους να σχεδιάσουν ένα δείκτη εγκυρότητας ο οποίος θα συνδυάζει και τα δύο κριτήρια(compactness, separation) όρισαν τον δείκτη V_{CWB} (Compose Within and Between scattering). Στην συνέχεια ακολουθεί μία σύντομη περιγραφή για τον ορισμό αυτού του δείκτη.

4.4.2.1 Ο ΔΕΙΚΤΗΣ COMPOSE WITHIN AND BETWEEN SCATTERING

Θεωρούμε ότι έχουμε διαιρέσει το σύνολο δεδομένων $X=\{x_1, \dots, x_n | x_i \in R^p\}$ σε c clusters και v_i είναι τα κέντρα των c clusters έτσι ώστε $V=\{v_1, \dots, v_n\}$ και έχουμε τον πίνακα συμμετοχής $U=[U_{ik} (i=1,2,\dots,c; k=1,2,\dots,n)]$ του οποίου τα στοιχεία δείχνουν το βαθμό συμμετοχής του στοιχείου x_k στο cluster i. Με βάση τις υποθέσεις αυτές μπορούμε να ορίσουμε τις εξής τιμές:

- Η διακύμανση των στοιχείων του συνόλου των δεδομένων X καλείται $\sigma(X) \in R^p$. Η τιμή της p διάστασης του $\sigma(X)$ ορίζεται από την εξίσωση (Εξισ. 4.14):

$$\sigma_x^p = \frac{1}{n} \sum_{k=1}^n \left(x_k^p - \bar{x}^p \right)^2 \quad (\text{Εξισ. 4.14})$$

$$\text{με } \bar{x}^p = \frac{1}{n} \sum_{k=1}^n x_k^p, \forall x_k \in X$$

- Η ασαφής απόκλιση του cluster i καλείται $\sigma(v_i) \in R^p$. Η τιμή της p διάστασης του $\sigma(v_i)$ ορίζεται από την (Εξισ. 4.15):

$$\sigma_{v_i}^p = \frac{1}{n} \sum_{k=1}^n U_{ik} \left(x_k^p - v_i^p \right)^2 \quad (\text{Εξισ. 4.15})$$

- Η μέση διασπορά για c clusters ορίζεται από την (Εξισ. 4.16):

$$Scat(c) = \frac{\frac{1}{c} \sum_{i=1}^c \|\sigma(v_i)\|}{\|\sigma(X)\|} \quad (\text{Εξισ. 4.16})$$

οπου $\|x\| = (x^T x)^{1/2}$

- Η συνάρτηση για την εκτίμηση της συνολικής διαφοροποίησης της απόστασης μεταξύ των clusters, Disc(c), ορίζεται με βάση την (Εξισ. 4.17):

$$Disc(c) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^c \left(\sum_{i=1}^c \|v_k - v_i\| \right)^{-1} \quad (\text{Εξισ. 4.17})$$

όπου $D_{\max} = \text{maximum}(\|v_i - v_j\| \forall i, j \in \{2, 3, \dots, c\})$ είναι η μέγιστη απόσταση ανάμεσα στα κέντρα των clusters. Η $D_{\min} = \text{minimum}(\|v_i - v_j\| \forall i, j \in \{2, 3, \dots, c\})$ είναι η ελάχιστη απόσταση ανάμεσα στα κέντρα των clusters.

Με βάση τα παραπάνω ο δείκτης εγκυρότητας μπορεί να οριστεί σε συνδυασμό με τις δύο τελευταίες εξισώσεις [(Εξισ. 4.16), (Εξισ. 4.17)] ως εξής:

$$V_{CWB}(U, V) = \alpha \text{Scatt}(c) + \text{Dis}(c). \quad (\text{Εξισ. 4.18})$$

όπου α είναι ένας συντελεστής βάρους ίσος με $\text{Dis}(c_{\max})$.

Ο πρώτος όρος του V_{CWB} δηλαδή $\text{Scatt}(c)$ δηλώνει τον μέσο όρο της απόκλισης μέσα στα clusters για κάθε αριθμό c των clusters. Μία μικρή τιμή για τον όρο αυτό υποδηλώνει ένα συμπαγές τμήμα. Όσο η απόκλιση μεταξύ των στοιχείων μέσα στα clusters μεγαλώνει, αυτά γίνονται λιγότερο συμπαγή και για το λόγο αυτό το $\text{Scatt}(c)$ είναι μία καλή ένδειξη για την μέση πυκνότητα στα clusters. Ο δεύτερος όρος του δείκτη εγκυρότητας, ο $\text{Dis}(c)$, εκφράζει την συνολική διαφοροποίηση μεταξύ των clusters. Γενικά, αυτός ο όρος θα αυξάνεται με την αύξηση του αριθμού των clusters και επηρεάζεται από την γεωμετρία των κέντρων των clusters. Επειδή οι τιμές των δύο όρων του V_{CWB} είναι διαφορετικής κλίμακας, απαιτείται ένας παράγοντας α προκειμένου να εξισορροπηθούν οι δύο όροι.

Ο αριθμός των clusters, ο οποίος ελαχιστοποιεί το δείκτη εγκυρότητας V_{CWB} μπορεί να ληφθεί σαν βέλτιστη τιμή για τον αριθμό των κλάσεων(clusters) των δεδομένων.

4.4.2.2 ΑΞΙΟΛΟΓΗΣΗ CLUSTERS ΠΟΥ ΠΡΟΚΥΠΤΟΥΝ ΑΠΟ ΤΟΝ ΑΛΓΟΡΙΘΜΟ FCM

Εάν ορίσουμε τη μέγιστη και ελάχιστη τιμή για τον αριθμό των clusters καθώς και τις τιμές μεταξύ των οπίων θα κυμαίνεται η παράμετρος ασάφειας $m \in (1, \dots, \infty)$ μπορούμε εφαρμόζοντας τον παρακάτω αλγόριθμο να βρούμε για κάθε τιμή της παραμέτρου ασάφειας το βέλτιστο αριθμό clusters (c_{opt})[RR98]. Σαν μέτρο αξιολόγησης της ποιότητας των clusters που προκύπτουν από την εφαρμογή του

αλγορίθμου FCM χρησιμοποιούμε το μέτρο V_{CWB} που περιγράψαμε λεπτομερέστερα παραπάνω.

Στην συνέχεια παρουσιάζονται τα βασικά βήματα του αλγορίθμου με μορφή ψευδοκώδικα:

1. Ορίζουμε την μέγιστη τιμή c_{max} και ελάχιστη τιμή c_{min} του αριθμού των clusters. Επίσης ορίζουμε τις τιμές m_{min} και $m_{max} \in (1, \dots, \infty)$ ως την μέγιστη και ελάχιστη τιμή αντίστοιχα για την παράμετρο ασάφειας, m .
2. Αρχικοποίηση: $m = m_{min}$
 $c = c_{max}, c_{opt,m} = c;$
3. If ($m < m_{max}$)
Εφαρμογή FCM στο σύνολο των δεδομένων, με σκοπό να καθοριστούν τα clusters και οι βαθμοί συμμετοχής μ_{ik} (συμμετοχή του στοιχείου k στο clusters i).
else stop.
4. If ($c=c_{max}$)
{ $a=Dis(c_{max}); indexValue=V_{CWB}(m,c)$ }
else if ($V_{CWB}(m,c) < indexValue$)
{ $c_{opt}=c; indexValue = V_{CWB}(m,c);$ }
5. $c=c-1$,
if ($c=c_{min}-1$)
{ $m=m+0.05$ },
goto 3.

Μετά το βήμα 3 έχει προκύψει μία fuzzy τμηματοποίηση του συνόλου των δεδομένων για συγκεκριμένη τιμή του m . Αυτή η τμηματοποίηση αξιολογείται στο βήμα 4 με την βοήθεια του μέτρου V_{CWB} . Στην παράμετρο V_{CWB} (c, m) "indexValue" αποθηκεύεται η ελάχιστη τιμή των $V_{CWB}(c, m)$ μέχρι στιγμής. Μετά το βήμα 5 ο βέλτιστος αριθμός clusters $c_{opt} \in [c_{min}, c_{max}]$ που βρίσκουμε θα αντιστοιχεί στην ελάχιστη τιμή για το $V_{CWB}(c, m)$. Τα βήματα 3 -5 επαναλαμβάνονται για διαφορετικές τιμές της παραμέτρου ασάφειας(m) και έτσι μπορούμε να έχουμε για κάθε τιμή της παραμέτρου m το βέλτιστο αριθμό των clusters $c_{opt,m}$. Συνεπώς με βάση τον παραπάνω αλγόριθμο έχουμε την fuzzy τμηματοποίηση του συνόλου των δεδομένων στο βέλτιστο αριθμό clusters για κάθε τιμή του m . Θα πρέπει όμως να επιλέξουμε εκείνη την τμηματοποίηση που ανταποκρίνεται καλύτερα στις απαιτήσεις μας και η οποία περιέχει το μεγαλύτερο πληροφοριακό περιεχόμενο για την εφαρμογή μας.

Επίσης θα πρέπει να σημειώσουμε ότι ο αλγόριθμος για την αξιολόγηση των clusters και την επιλογή του καλύτερου αριθμού clusters μπορεί να εφαρμοστεί υιοθετώντας κάποιο άλλο μέτρο αξιολόγησης. Εάν λοιπόν εκτιμήσουμε ότι κάποιο άλλο μέτρο αξιολόγησης clustering ανταποκρίνεται καλύτερα στις απαιτήσεις και ανάγκες της εφαρμογής μπορούμε να προσαρμόσουμε κατάλληλα τον αλγόριθμο ώστε να επιτύχουμε τα καλύτερα δυνατά αποτελέσματα.

5^ο ΚΕΦΑΛΑΙΟ

ΥΠΟΣΤΗΡΙΞΗ ΑΒΕΒΑΙΟΤΗΤΑΣ ΣΤΟ DATA MINING

5.1 ΕΙΣΑΓΩΓΗ

Η διαδικασία εξόρυξης γνώσης από μεγάλα σύνολα δεδομένων έχει σαν κύριο στόχο την αναζήτηση έγκυρων, χρήσιμων προτύπων μέσα από τα σύνολα δεδομένων τα οποία παρουσιάζουν ενδιαφέρον. Προκειμένου όμως η διαδικασία αυτή να έχει θετικά αποτελέσματα και να συμβάλλει στην αποτελεσματική εκμετάλλευση της γνώσης που εξάγεται, θα πρέπει οι διάφορες μορφές γνώσεις που προκύπτουν από την διαδικασία *data mining* (π.χ. κανόνες, κατηγοριοποιήσεις κλπ) να είναι πλήρως κατανοητές σε μη ειδικούς. Αυτή αποτελεί μία βασική απαίτηση τόσο του επιστημονικού όσο και του επιχειρηματικού κόσμου ώστε να μπορεί να υποστηριχθεί μία μεγάλη μερίδα στελεχών κατά την διαδικασία λήψης αποφάσεων. Μία άλλη αναγνωρισμένη απαίτηση είναι η διαχείριση της αβεβαιότητας κατά την διαδικασία KDD, δηλαδή πως μπορούμε να αναπαραστήσουμε την αβεβαιότητα(uncertainty) στα στάδια της διαδικασίας *data mining*.

Η βασική προσέγγιση για την αντιμετώπιση προβλημάτων ασάφειας και την ικανοποίηση των παραπάνω απαιτήσεων είναι να συνδυάσουμε τεχνικές *Data Mining* με την θεωρία της *Ασαφούς Λογικής (Fuzzy Logic)* [Vaz98]. Η προσέγγιση αυτή αποτελεί και το βασικό αντικείμενο του κεφαλαίου αυτού. Συγκεκριμένα, θα αναφερθούμε στην δημιουργία ενός σχήματος *data mining* το οποίο συνδυάζοντας στοιχεία από την θεωρία *Fuzzy Sets* και *Fuzzy Logic* μπορεί να συμβάλλει στην εξαγωγή χρήσιμων προτύπων από μεγάλα σύνολα δεδομένων.

5.2 FUZZY LOGIC

Σύμφωνα με την *Boolean* λογική κάθε αντικείμενο ανήκει σε ένα και μόνο σύνολο. Στις περισσότερες όμως περιπτώσεις που αντιμετωπίζουμε στον πραγματικό κόσμο δεν μπορούμε να εκφράσουμε με βεβαιότητα ότι ένα αντικείμενο ανήκει σε κάποια συγκεκριμένη κατηγορία ή γενικότερα ότι κάτι ισχύει με βεβαιότητα. Το πρόβλημα αυτό της ασάφειας έρχεται να αντιμετωπίσει η ασαφής λογική εισάγοντας την έννοια του βαθμού βεβαιότητας (*degree of belief*).

Η ασαφής λογική αποτελεί επέκταση της κλασικής λογικής και εισήχθη από τον Zadeh στην προσπάθεια να αναπαραστήσει και να διαχειριστεί δεδομένα τα οποία δεν ήταν σαφή. Κλειδί της θεωρίας αυτής αποτελεί ο βαθμός βεβαιότητας, ο οποίος εκφράζει την πίστη με την οποία ισχύει κάποια πρόταση.

Στην κλασική λογική η έκφραση

"X is in A"

θα μπορούσε να είναι αληθής ή ψευδής δηλαδή ο βαθμός συμμετοχής του X στο A μπορούσε να πάρει τιμές στο σύνολο $\{0, 1\}$. Σύμφωνα όμως με την ασαφή λογική η παραπάνω φράση μπορεί να ισχύει με κάποιο βαθμό βεβαιότητας παίρνοντας τιμές στο διάστημα $[0,1]$.

Αμεση συσχετισμένη με την ασαφή λογική είναι η θεωρία ασαφών συνόλων η οποία αποτελεί επέκταση της κλασικής θεωρίας συνόλων. Στην θεωρία αυτή η βασική έννοια είναι το **ασαφές σύνολο (fuzzy set)**. Ένα ασαφές σύνολο έχει ως χαρακτηριστικό ότι αποτελείται από τα στοιχεία ενός συνόλου S τα οποία ανήκουν σε αυτό με κάποιο βαθμό βεβαιότητας. Έτσι ένα σύνολο $S = \{s_i\}$ ορίζει το ασαφές σύνολο F_S το οποίο αποτελείται από ζεύγη της μορφής (s_i, μ_i) , όπου s_i είναι στοιχείο του συνόλου S και $\mu_i \in [0,1]$ είναι ο βαθμός συμμετοχής του s_i στο σύνολο. Η μαθηματική πρόταση $x \in A$, όπου A είναι ασαφές σύνολο, αποτελεί πρόταση της ασαφούς λογικής.

5.3 FUZZY LOGIC AND DATA MINING

Η ασαφής λογική έχει πολλαπλές εφαρμογές σε διάφορες επιστημονικές περιοχές. Σημαντικό όμως είναι το ενδιαφέρον που παρουσιάζει σε συνδυασμό με τεχνικές data mining. Ο συνδυασμός αυτός μπορεί να βοηθήσει στην αντιμετώπιση προβλημάτων που παρουσιάζονται στην διαδικασία εξόρυξης χρήσης γνώσης από μεγάλα σύνολα δεδομένων. Τα δεδομένα που χρησιμοποιούμε στην διαδικασία data mining είναι δεδομένα που προέρχονται από τον πραγματικό κόσμο με αποτέλεσμα η γνώση που περιέχουν να είναι ασαφής. Ο στόχος μας όμως είναι να επεξεργαστούμε όσο υπόψη την ασάφεια που εμπεριέχουν να οδηγηθούμε σε συμπεράσματα χρήσιμα και κατανοητά.

Για παράδειγμα, εφαρμόζοντας τις κατάλληλες τεχνικές μπορεί από ένα υποσύνολο του σχήματος μίας βάσης δεδομένων για πωλήσεις να προκύψει ο εξής κανόνας:

Μισθός_πελάτη[180000, 300000] and Ηλικία_πελάτη[25-40] \Leftrightarrow τιμή[13000, 20000]

Στον παραπάνω κανόνα το διάστημα $[180000, 300000]$ για το μισθό του πελάτη δεν δίνει σαφή εικόνα για το τι αντιπροσωπεύει στην πλήρη κλίμακα των μισθών καθώς και σε τι ποσοστό πληθυσμού αντιστοιχεί. Συνεπώς, ένας αναλυτής δεν μπορεί να έχει άμεση και πλήρη αντίληψη των υποθέσεων και του τελικού συμπεράσματος που προκύπτει από τον κανόνα. Η χρήση λεκτικών χαρακτηρισμών για τα γνωρίσματα που χρησιμοποιούνται στον κανόνα θα μπορούσε να βοηθήσει στην καλύτερη κατανόηση αυτού. Έτσι προκύπτει η απαίτηση για κατανοητούς κανόνες

σαν αποτέλεσμα των επεξεργασιών data mining. Η απαίτηση αυτή θα μπορούσε να ικανοποιηθεί με την κατηγοριοποίηση των δεδομένων σε κατηγορίες που θα αναπαρίστανται από λεκτικές τιμές. Με τον τρόπο αυτό θα μπορέσουμε να διατυπώνουμε κανόνες οι οποίοι θα είναι πιο κοντά στην φυσική γλώσσα και συνεπώς καλύτερα κατανοητοί από τους αναλυτές. Ένας τέτοιος κανόνας θα έχει την μορφή:

Mισθός_ πελάτη υψηλός and Ήλικία_ πελάτη νέος \Rightarrow τιμή μεσαία

Ένα άλλο θέμα που προκύπτει κατά την προσέγγιση αυτή είναι ο τρόπος με τον οποίο θα γίνει η κατανομή των τιμών στα πεδία που έχουν οριστεί. Το πρόβλημα που πρέπει να αντιμετωπίσουμε είναι με πόση βεβαιότητα μπορούμε να κατατάξουμε σε μία κατηγορία την τιμή ενός γνωρίσματος όταν αυτή είναι κοντά στα όρια των πεδίων ορισμών δύο κατηγοριών. Για παράδειγμα, εάν έχουμε ορίσει την κατηγορία μέσος μισθός ως τους μισθούς από 100000 έως 250000 τότε ο μισθός 999000 δεν θα ανήκει στην κατηγορία αυτή παρά την μικρή απόκλιση από το κάτω όριο. Συνεπώς ένα άλλο θέμα που πρέπει να ληφθεί υπόψη είναι η απαίτηση για χρήση και αποκάλυψη της αβεβαιότητας κατά την διαδικασία της εξόρυξης γνώσης.

Με βάση το παραπάνω παράδειγμα μπορούμε να αντιληφθούμε την ασάφεια που κρύβεται στα σύνολα των δεδομένων που χειριζόμαστε καθημερινά και την επίδραση που μπορεί να έχει στην διαδικασία του data mining. Συνεπώς η χρήση της ασαφούς λογικής και ο συνδυασμός της με το data mining καθίσταται αναγκαίος προκειμένου να αντιμετωπίσουμε τον ασαφή χαρακτήρα των δεδομένων που αναλύουμε.

5.4 ΠΡΟΣΕΓΓΙΣΗ FUZZY DATA MINING

Η φύση των δεδομένων που διαχειριζόμαστε στο data mining και οι απαιτήσεις που έχουμε για κατανοητά και σαφή αποτελέσματα από την διαδικασία αυτή με βάση και όσα αναφέρθηκαν παραπάνω, μας οδηγούν στην ανάγκη για ορισμό ενός πλαισίου για κατηγοριοποίηση των δεδομένων και εξαγωγή γνώσης το οποίο θα ενσωματώνει στοιχεία ασαφούς λογικής. Σύμφωνα με την προσέγγιση για υποστήριξη της αβεβαιότητας στο data mining, που προτείνεται από τον κ. Βαζιργιάννη [Vaz98], οι βασικοί στόχοι κατά την διαδικασία εξόρυξης γνώσης από μεγάλες βάσεις δεδομένων είναι:

- ο ορισμός του σχήματος κατηγοριοποίησης των γνωρισμάτων σε λεκτικές κατηγορίες με βάση την ασαφή λογική και
- ο ορισμός ενός σχήματος για την εξαγωγή σχέσεων μεταξύ γνωρισμάτων με βάση το προηγούμενο σχήμα κατηγοριοποίησης

Στην παράγραφο αυτή θα περιγράψουμε τα βασικά χαρακτηριστικά της προσέγγισης αυτής, στην οποία βασίζεται και η παρούσα διπλωματική εργασία.

5.4.1 ΠΕΡΙΓΡΑΦΗ ΣΥΣΤΗΜΑΤΟΣ ΥΠΟΣΤΗΡΙΞΗΣ ΑΒΕΒΑΙΟΤΗΤΑΣ ΣΕ DATA MINING

Τα βασικά στάδια κατά την διαδικασία data mining σύμφωνα με την προτεινόμενη προσέγγιση [Vaz98] μπορούν να συνοψιστούν στα εξής:

- **Καθορισμός clusters για το σύνολο των δεδομένων.** Εφαρμόζοντας καλά ορισμένες μεθόδους clustering εξάγουμε/ορίζουμε clusters τα οποία θα μας δώσουν και τις αρχικές κατηγορίες. Έτσι έχοντας ένα αρχικό σύνολο δεδομένων "training data set" μπορούμε να εφαρμόσουμε ένα σύνολο μεθόδων clustering και να εξάγουμε τα αντίστοιχα μοντέλα.
- **Καθορισμός λεκτικών τιμών.** Για κάθε κατηγορία που προκύπτει με την εφαρμογή των μεθόδων clustering προσδιορίζουμε μία λεκτική τιμή. Με τον τρόπο αυτό ορίζεται ένα σύνολο λεκτικών τιμών $L=\{l_i\}$, όπου l_i είναι η λεκτική τιμή που αντιστοιχεί στην κατηγορία i (classification category).
- **Εξαγωγή συναρτήσεων συμμετοχής.** Τα clusters που προκύπτουν από τις περισσότερες μεθοδολογίες clustering (κλασικές μεθοδολογίες Κεφάλαιο 2) παράγουν crisp clusters (δηλ. κάθε αντικείμενο ανήκει σε ένα και μόνο ένα cluster). Καθώς όμως θέλουμε να λάβουμε υπόψη μας την ασάφεια που εμπεριέχουν τα δεδομένα από την φύση τους, θα θέλαμε να καθορίσουμε κάποιες συναρτήσεις συμμετοχής στα clusters. Οι συναρτήσεις αυτές βασίζονται σε μεθοδολογίες ασαφούς λογικής και στοχεύουν στην αντιστοίχηση των γνωρισμάτων A_i της βάσης δεδομένων στα clusters σύμφωνα με της λεκτικές τιμές L_i που έχουν οριστεί προηγουμένως. Το αποτέλεσμα αυτής της διαδικασίας είναι ένα σύνολο από βαθμούς πίστης (degree of belief) $M=\{\mu_{li}(t_k, A_i)\}$. Κάθε στοιχείο του συνόλου αναπαριστά το αποτέλεσμα της συνάρτησης συμμετοχής:

$$f(t_k, A_i) = \mu_{li}(t_k, A_i)$$

και δηλώνει την πίστη ότι η συγκεκριμένη τιμή t_k, A_i ανήκει στο cluster που δηλώνεται από την λεκτική τιμή l_i .

Τα δεδομένα που έχουμε για κάθε μοντέλο μπορούν να αναπαρασταθούν με την μορφή ενός κύβου C . Σε κάθε κελί του κύβου αποθηκεύεται ο βαθμός πίστης για κατηγοριοποίηση της τιμής του γνωρίσματος A_i στην κατηγορία l_i , $C[A_i, l_i, t_k] = \mu_{li}(t_k, A_i)$.



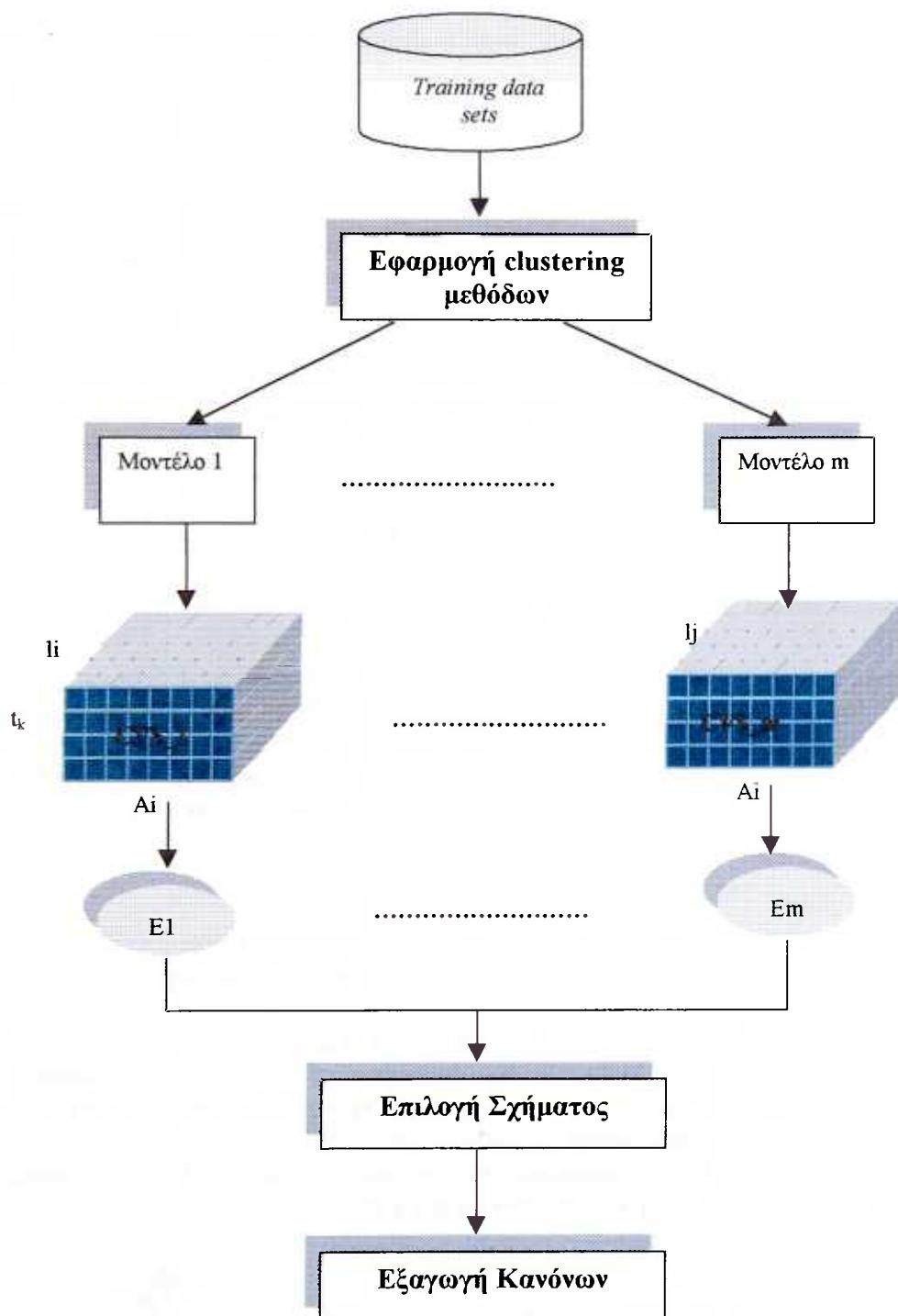
Ετσι τα σύνολα δεδομένων μετατρέπονται σε σύνολα CVS δηλαδή σε σύνολα της μορφής, $CVS = \{l_i, A_i, \mu_{li}(t_k, A_i)\}$.

- **Εκτίμηση της πληροφορίας που περιέχεται στο κύβο(σύνολο CVS).** Μετά την μετατροπή του συνόλου των δεδομένων σε σύνολα CVS (ένα για κάθε μοντέλο που προκύπτει από την εφαρμογή διαφορετικών clustering μεθόδων), θα πρέπει με βάση κάποιο κριτήριο ποιότητας να επιλέξουμε το καλύτερο μοντέλο. Καλύτερο μοντέλο για εξαγωγή γνώσης από την βάση δεδομένων μας χαρακτηρίζεται αυτό που περιέχει το μεγαλύτερο ποσοστό πληροφορίας ανά γνώρισμα ή συνολικά.

Εάν εφαρμοστούν τις μέθοδοι clustering τότε θα προκύψουν τα κύβοι. Οι κύβοι αυτοί αναφέρονται στην ίδια βάση δεδομένων δηλαδή στα ίδια γνωρίσματα και στις ίδιες τιμές για αυτά, στο μόνο που διαφοροποιούνται είναι οι λεκτικές μονάδες με βάση τις οποίες κατηγοριοποιούνται τα γνωρίσματα. Στο στάδιο αυτό το πρόβλημα που έχουμε να αντιμετωπίσουμε είναι η εκτίμηση της γνώσης που εμπεριέχεται σε κάθε κύβο (CVS Evaluation) ώστε να επιλεχθεί το καλύτερο μοντέλο. Στην παράγραφο 5.4.1.1 περιγράφονται συνοπτικά τα βασικά μέτρα εκτίμησης της πληροφορίας που περιέχεται στον κύβο που προκύπτει με βάση την μεθοδολογία αυτή.

- **Εξαγωγή κανόνων.** Βασιζόμενοι στο classification σχήμα που περιγράψαμε παραπάνω, προσπαθούμε να εξάγουμε τους κανόνες που υποστηρίζονται με συγκεκριμένη πίστη από το σύστημα μας. Ο στόχος δηλαδή είναι να προσδιορίσουμε συσχετίσεις μεταξύ των κατηγοριών (λεκτικών τιμών) των γνωρισμάτων, οι οποίες παρουσιάζουν κάποιο ενδιαφέρον για την εφαρμογή μας. Το βασικό στοιχείο της μεθοδολογίας είναι ο κατανοητός τρόπος με τον οποίο θα εμφανίζονται οι κανόνες, καθώς οι κανόνες παρουσιάζουν σχέσεις μεταξύ λεκτικών τιμών. Με τον τρόπο αυτό οι κανόνες προσεγγίζουν περισσότερο την φυσική γλώσσα και μπορούν να είναι καλύτερα κατανοητοί και από μη ειδικούς.

Στο Σχήμα 5.1 παρουσιάζονται διαγραμματικά τα βασικά βήματα της μεθοδολογίας για εισαγωγή της ασάφειας στη διαδικασία του data mining.



Σχήμα 5.1. Στάδια μεθοδολογίας για υποστήριξη αβεβαιότητας στο Data Mining

5.4.1.1 CVS EVALUATION

Με τον όρο *CVS Evaluation* αναφερόμαστε στην εκτίμηση της πληροφορίας που περιέχεται στον κύβο που προκύπτει με βάση την παραπάνω μεθοδολογία. Προσπαθούμε δηλαδή να απαντήσουμε στο ερώτημα: *Πόση γνώση εμπεριέχεται στον κύβο που δημιουργείται με την μεθοδολογία μας;*

Οι διαφορετικοί κύβοι που έχουν προκύψει από την εφαρμογή διαφόρων μεθοδολογιών clustering αναφέρονται όλοι στη ίδια βάση το μόνο που διαφέρει είναι οι λεκτικές μονάδες που αναφέρονται σε κάθε γνώρισμα. Ο στόχος μας είναι να αξιολογήσουμε το περιεχόμενο των συνόλων δεδομένων που αντιπροσωπεύει κάθε μοντέλο και να επιλέξουμε αυτό που παρέχει την πληροφορία μέγιστης σημαντικότητας για τις αποφάσεις μας. Για το σκοπό αυτό έχουν αναπτυχθεί κάποια μέτρα γνώσης, τα οποία μετρούν την γνώση που περιέχεται στα σύνολα δεδομένων (δηλ. στον κύβο που προκύπτει από κάθε μοντέλο) που προκύπτουν από την εφαρμογή κάθε μεθόδου clustering.

Τα μέτρα αυτά γνώσης είναι [AT98][Vaz98]:

- ◆ *H ενέργεια ανά λεκτική τιμή γνωρίσματος (Energy of each lexical value)*

Το μέτρο αυτό χρησιμοποιείται για να εκτιμηθεί πόσο σημαντική είναι η πληροφορία που περιέχεται στις τιμές ενός γνωρίσματος και πόσο καλά τα δεδομένα προσαρμόζονται σε αυτό το σχήμα κατηγοριοποίησης. Με βάση την συνάρτηση *energy metric function*, η συνολική γνώση ότι το εξεταζόμενο σύνολο δεδομένων που περιέχει τιμές l_i για το γνώρισμα A_i ορίζεται ως εξής:

$$E_{li}(A_i) = \sum_k [\mu_{li}(t_k, A_i)]^q \quad (\text{Εξισ. 5.1})$$

όπου l_i λεκτική μονάδα που αντιστοιχεί σε κάποια κατηγορία, A_i το γνώρισμα, t_k η k πλειάδα ενός πίνακα της βάσης μας.

Προκειμένου να υπολογίσουμε την πίστη ότι η βάση δεδομένων μας περιέχει για το γνώρισμα A_i την τιμή l_i , μπορούμε να ορίσουμε έναν συντελεστή πίστης(ή αβεβαιότητας) ανά λεκτική τιμή γνωρίσματος. Ο συντελεστής αυτός λαμβάνεται με την κατάλληλη κανονικοποίηση της συνολικής πίστης όπως υπολογίστηκε από την εξίσωση (Εξισ. 5.1) ώστε να κυμαίνεται στο διάστημα $[0,1]$. Έτσι, η πίστη ότι η βάση μας υποστηρίζει την λεκτική τιμή l_i για το γνώρισμα A_i μπορεί να εκφραστεί ως εξής



$$U_{li}(A_i) = \left(\frac{\sum_k [\mu_{li}(t_k, A_i)]^q}{N} \right)^{1/q}$$

όπου N συνολικός αριθμός των πλειάδων στην βάση μας.

- **Συνολική ενέργεια (Overall Energy)**

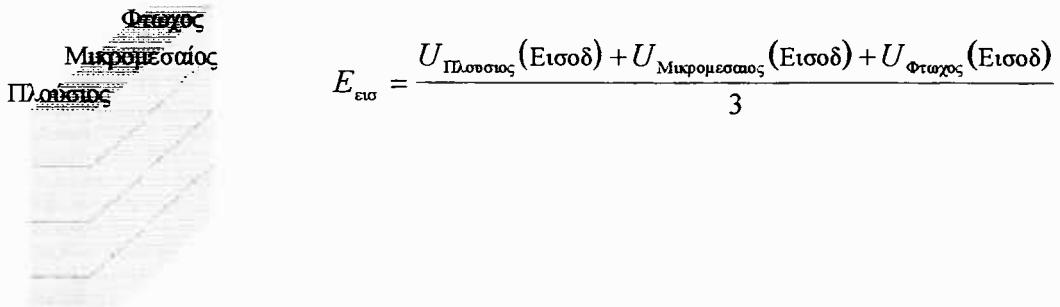
Η συνολική ενέργεια ενός γνωρίσματος A_i θα εκφράζει την συνολική πληροφορία για όλες τις κατηγορίες ενός χαρακτηριστικού. Προσδιορίζει δηλαδή την συνολική γνώση που υπάρχει στον κύβο μας σχετικά με το γνώρισμα A_i . Η ενέργεια αυτή δίνεται από την εξίσωση

$$E_{A_i} = \frac{\sum_{h_i} [U_{h_i}(A_i)]}{c} \quad (\text{Εξισ. 5.2})$$

όπου c ο αριθμός των κατηγοριών για το γνώρισμα A_i .

Παράδειγμα 2.

Πόση γνώση υπάρχει στον κύβο σχετικά με το "Εισόδημα";



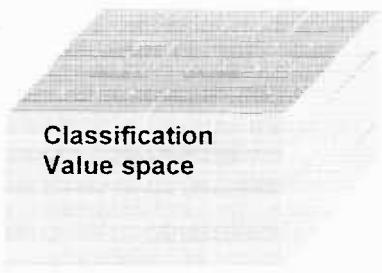
Εισόδημα

Μέσω λοιπόν της συνάρτησης εκτίμησης της συνολικής ενέργειας μπορούμε να υπολογίσουμε την συνολική γνώση που υπάρχει στον κύβο που προκύπτει για κάθε σχήμα κατηγοριοποίησης(classification schema) σχετικά με το γνώρισμα A_i . Το μέτρο αυτό μπορεί να χρησιμοποιηθεί για την σύγκριση των σχημάτων κατηγοριοποίησης πάνω σε ένα σύνολο δεδομένων, δηλαδή για την σύγκριση των κύβων που προκύπτουν από διαφορετικά μοντέλα.

Έτσι εάν κατά την διαδικασία data mining έχουν εφαρμοστεί τη διαφορετικές μέθοδοι clustering και έχουν προκύψει τη μοντέλα ($M_i, i=1\dots m$), τότε για κάθε μοντέλο M_j μπορούμε να υπολογίσουμε τη συνολική ενέργεια E_{A_i} . Το μοντέλο που θα περιέχει την μεγαλύτερη πίστη αναφορικά με το γνώρισμα A_i θα είναι και το μοντέλο που θα επιλεχθεί για την εξαγωγή συμπερασμάτων.

- **CVS ενέργεια**

Το μέτρο αυτό έχει σαν στόχο να εκφράσει την συνολική γνώση που έχει ο κύβος μας. Το μέτρο αυτό θα μπορούσε να υπολογίζεται σαν το άθροισμα της πίστης που έχει ο κύβος μας για κάθε γνώρισμα A_i .



$$E_{CVS} = \sum_{A_i} [E(A_i)] \quad (\text{Εξισ. 5.3})$$

Με βάση το μέγεθος αυτό θα μπορούσαμε να εκτιμήσουμε την ποιότητα του μοντέλου κατά την διάρκεια της ζωής του και έτσι να προσδιορίσουμε την ανάγκη για δημιουργία νέου μοντέλου για την βάση μας. Για το σκοπό αυτό μπορούμε να ορίσουμε κάποιο βάρος για τα γνωρίσματα ανάλογα με την σημαντικότητα που έχουν για το σύστημά μας. Το σημαντικό σε μία τέτοιου είδους προσέγγιση είναι πως ένας ειδικός μπορεί ανάλογα με τις ανάγκες του να καθοδηγήσει το αποτέλεσμα, ορίζοντας το κατάλληλο βάρος για κάθε γνώρισμα.

Ετσι εάν w_i ($0 \leq w_i \leq 1$) είναι το βάρος για το γνώρισμα A_i τότε η γνώση που θα περιέχει ο κύβος θα είναι

$$E_{CVS} = \sum_i w_i [E(A_i)] \quad (\text{Εξισ. 5.4})$$

Στον Πίνακα 5.1 συνοψίζονται τα βασικά χαρακτηριστικά κάθε ενός από τα παραπάνω μέτρα.

ΜΕΤΡΟ ΓΝΩΣΗΣ	ΠΕΡΙΓΡΑΦΗ
Συντελεστής πίστης ανά λεκτική τιμή	Εκφράζει ποία είναι η πίστη ότι η βάση δεδομένων περιέχει την λεκτική τιμή li για το γνώρισμα Ai.
Συνολική πίστη	Το μέτρο αυτό είναι χρήσιμο για σύγκριση του πληροφοριακού περιεχομένου διαφορετικών βάσεων που απεικονίζονται στο κύβο με τις ίδιες λεκτικές κατηγορίες για τα ίδια γνωρίσματα
CVS ενέργεια	Εκφράζει την συνολική γνώση που υπάρχει στον κύβο και η οποία σχετίζεται με το γνώρισμα Ai. Προσφέρεται για σύγκριση διαφορετικών μοντέλων που έχουν προκύψει από την ίδια βάση δεδομένων. Χρησιμοποιείται κατά την διαδικασία επιλογής του καλύτερου μοντέλου(δηλ. αυτού που περιέχει την περισσότερη γνώση)

Πίνακα 5.1. Συγκεντρωτικός πίνακας με τα μέτρα γνώσης

6^ο ΚΕΦΑΛΑΙΟ

ΥΠΟΣΤΗΡΙΞΗ ΤΗΣ ΑΣΑΦΕΙΑΣ ΣΤΟ CLUSTERING (FUZZY CLUSTERING)

6.1 ΕΙΣΑΓΩΓΗ

Το clustering είναι από τις βασικές διαδικασίες data mining, της οποία ο βασικός στόχος είναι να βρει ομάδες ομοίων αντικειμένων από ένα μεγάλο σύνολο δεδομένων. Περιγράφεται ως unsupervised learning, καθώς εφαρμόζεται πάνω σε δεδομένα χωρίς να υπάρχει η γνώση για το τι αποτελέσματα θα προκύψουν. Οι ομάδες δηλαδή στις οποίες θα κατηγοριοποιηθούν τα δεδομένα μας δεν είναι γνωστές εκ των προτέρων.

Μία μεγάλη γκάμα από καλά τεκμηριωμένες μεθοδολογίες έχουν αναπτυχθεί για την διαδικασία clustering. Οι κλασικές μέθοδοι clustering (Κεφάλαιο 2) οδηγούν σε ομάδες με συγκεκριμένα όρια (crisp). Στην πραγματικότητα όμως οι περισσότερες εφαρμογές καθιστούν αναγκαία την εισαγωγή της ασάφειας. Για το σκοπό αυτό έχουν αναπτυχθεί κάποιοι fuzzy clustering αλγόριθμοι οι οποίοι λαμβάνουν υπόψη την ασάφεια των δεδομένων κατά την διαδικασία του clustering. Η ανάπτυξη μίας μεθοδολογίας η οποία θα υποστηρίζει την αβεβαιότητα στην διαδικασία clustering αποτελεί και το βασικό αντικείμενο της παρούσας εργασίας.

Στόχος του κεφαλαίου αυτού είναι η περιγραφή των βασικών στοιχείων της προσέγγισης που αναπτύχθηκε στα πλαίσια της διπλωματικής για την αντιμετώπιση του προβλήματος της ασάφειας στο clustering.

6.2 ΠΡΟΣΕΓΓΙΣΗ FUZZY CLUSTERING

Στο κεφάλαιο 5 αναφερθήκαμε στην ανάγκη για την ανάπτυξη ενός συστήματος το οποίο θα διαχειρίζεται την ασάφεια στις διαδικασίες clustering, classification και association rules extraction, με στόχο την εξαγωγή γνώσης που θα είναι χρήσιμη και κατανοητή για την εξαγωγή συμπερασμάτων. Οι βασικοί στόχοι της μεθοδολογίας που περιγράφαμε και η οποία βασίζεται στις ιδέες του κ. Βαζιργιάνη [Vaz98] συνοψίζονται στους εξής δύο:

- ο ορισμός ενός σχήματος κατηγοριοποίησης των γνωρισμάτων σε λεκτικές κατηγορίες με βάση την ασαφή λογική και

- ο ορισμός ενός σχήματος για την εξαγωγή σχέσεων μεταξύ γνωρισμάτων με βάση το προηγούμενο σχήμα κατηγοριοποίησης

Η μελέτη της διπλωματικής εργασίας εντάσσεται στον πρώτο από τους στόχους της διαδικασίας data mining και ειδικότερα στην εισαγωγή της ασάφειας στην διαδικασία clustering.

6.2.1 ΥΠΑΡΧΟΥΣΕΣ ΕΡΓΑΣΙΕΣ ΣΤΟ ΧΩΡΟ ΤΟΥ FUZZY CLUSTERING

Για την αντιμετώπιση της ασάφειας στο clustering έχουν αναπτυχθεί κάποιοι αλγόριθμοι οι οποίοι ενσωματώνουν στην υλοποίηση τους στοιχεία της Ασαφούς Λογικής. Οι αλγόριθμοι αυτοί οδηγούν σε επικαλυπτόμενες ομάδες δεδομένων και καθορίζουν τον βαθμό συμμετοχής ενός στοιχείου σε καθένα από τα clusters που έχουν προκύψει.

Οι κυριότεροι αλγόριθμοι fuzzy clustering είναι ο *Fuzzy C-Means*(Κεφ. 3) που αποτελεί επέκταση του κλασικού αλγορίθμου C-Means για fuzzy εφαρμογές και διατυπώθηκε από τους Bezdeck, J.C., Ehrlich R., Full W., "FCM: Fuzzy C-Means Algorithm", στο επιστημονικό περιοδικό Computers and Geoscience 1984, καθώς και ο αλγόριθμος *Fuzzy Kohonen Network* (Κεφ. 2) που αποτελεί προσαρμογή του αντίστοιχου αλγορίθμου Νευρωνικών Δικτύων.

6.2.2 ΠΡΟΣΕΓΓΙΣΗ ΠΑΡΟΥΣΑΣ ΕΡΓΑΣΙΑΣ

Εκτός όμως από τις περιπτώσεις που χρησιμοποιούμε κάποιον αλγόριθμο fuzzy clustering θα ήταν χρήσιμο να υπάρχει δυνατότητα εισαγωγής της ασάφειας και στην περίπτωση των clusters που προκύπτουν από κάποιον αλγόριθμο που οδηγεί σε crisp clusters. Για να διακρίνουμε τους αλγορίθμους που δεν λαμβάνουν από μόνοι τους υπόψη την ασάφεια σε αντίθεση με τους fuzzy clustering αλγορίθμους, θα τους αποκαλούμε στην συνέχεια crisp clustering αλγορίθμους.

Στα πλαίσια αυτής της διπλωματικής εργασίας έγινε μία προσπάθεια για την αντιμετώπιση του παραπάνω προβλήματος με την ανάπτυξη μίας μεθοδολογίας η οποία θα επιτρέπει την εισαγωγή της ασάφειας και στην περίπτωση που εφαρμόζονται κλασικοί αλγόριθμοι clustering. Στην συνέχεια περιγράφονται τα βασικά βήματα της προσέγγισης για fuzzy clustering:

1. *Εφαρμογή μεθόδων clustering για την εξαγωγή και τον ορισμό των clusters που θα δώσουν τις αρχικές κατηγορίες.* Έτσι έχοντας ένα αρχικό σύνολο δεδομένων μπορούμε να εφαρμόσουμε ένα σύνολο μεθόδων clustering και να εξάγουμε τα αντίστοιχα μοντέλα.
2. *Αξιολόγηση των clusters που προκύπτουν από την εφαρμογή κάθε μεθόδου και επιλογή των καλύτερων δυνατών μοντέλων.* Ο στόχος είναι τα clusters να είναι πυκνά δηλαδή η απόσταση μεταξύ των στοιχείων να είναι μικρή ενώ η απόσταση

μεταξύ τους να είναι μεγάλη. Για το σκοπό αυτό έχουν αναπτυχθεί διάφορες μέθοδοι εκτίμησης της ποιότητας του clustering (Κεφ. 4).

3. *Προσδιορισμός Βαθμών Συμμετοχής*. Εδώ έχουμε δύο περιπτώσεις ανάλογα με το είδος των αλγορίθμων clustering που εφαρμόζουμε. Στην περίπτωση που εφαρμόζουμε έναν αλγόριθμο fuzzy clustering οι βαθμοί συμμετοχής των στοιχείων στα clusters προκύπτουν κατά την διαδικασία. Όταν όμως εφαρμόζουμε διαδικασία κλασικών αλγορίθμων θα πρέπει να εισάγουμε την ασάφεια εκ των υστέρων. Στην τελευταία λοιπόν περίπτωση θα πρέπει να καθορίσουμε κάποιες συναρτήσεις συμμετοχής

Οι συναρτήσεις αυτές βασίζονται σε μεθοδολογίες ασαφούς λογικής και στοχεύουν στην αντιστοίχηση των γνωρισμάτων A_i της βάσης δεδομένων στα clusters σύμφωνα με της λεκτικές τιμές l_i που έχουν οριστεί. Το αποτέλεσμα αυτής της διαδικασίας είναι ένα σύνολο από βαθμούς πίστης (degree of belief) $M = \{\mu_{li}(t_k, A_i)\}$. Κάθε στοιχείο του συνόλου αναπαριστά το αποτέλεσμα της συνάρτησης κατηγοριοποίησης

$$f(t_k, A_i) = \mu_{li}(t_k, A_i)$$

και δηλώνει την πίστη ότι η συγκεκριμένη τιμή t_k, A_i ανήκει στο cluster που δηλώνεται από την λεκτική τιμή l_i .

6.2.2.1 ΣΥΝΔΕΣΗ ΤΩΝ ΠΑΡΑΠΑΝΩ ΒΗΜΑΤΩΝ ΜΕ ΤΗΝ ΔΙΑΔΙΚΑΣΙΑ CLASSIFICATION

Η εργασία αυτή όπως έχουμε προαναφέρει εντάσσεται μέσα στο γενικότερο πλαίσιο για την ανάπτυξη ενός συστήματος data mining το οποίο θα υποστηρίζει την ασάφεια. Έτσι μετά τον ορισμό των clusters από το σύνολο των δεδομένων και τον καθορισμό των βαθμών συμμετοχής κάθε στοιχείου στα clusters, μπορούμε να προχωρήσουμε στην διαδικασία του classification. Βέβαια τα θέματα του classification δεν αποτελούν αντικείμενο της εργασίας, ωστόσο θεωρούμε σκόπιμο να αναφέρουμε πως συνδέεται η διαδικασία clustering που περιγράφουμε με τα υπόλοιπα στάδια της μεθοδολογίας ώστε να έχουμε μία πιο ολοκληρωμένη εικόνα. Επίσης με βάση τα παρακάτω στάδια μπορούμε να κατανοήσουμε πως η προσέγγιση clustering που προτείνεται μπορεί να βοηθήσει στην γενικότερη διαδικασία data mining (Κεφ.5)[Vaz98][AT98].

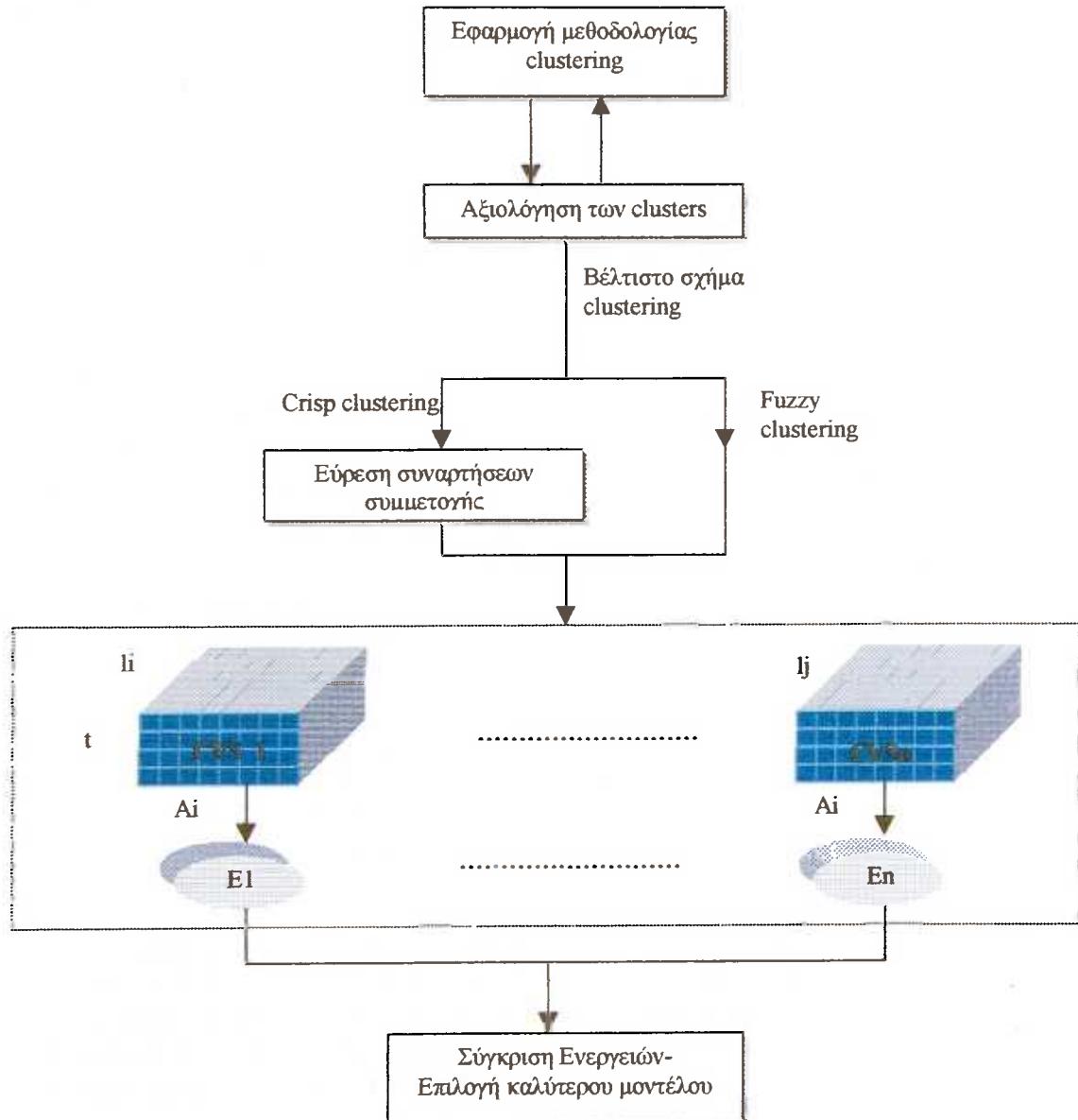
- *Μετατροπή των συνόλων των δεδομένων σε CVS (Classification Value Set).*

Τα δεδομένα που έχουμε για κάθε σχήμα clustering μπορούμε να αναπαρασταθούν με την μορφή ενός κύβου C . Σε κάθε κελί του κύβου αποθηκεύεται ο βαθμός πίστης για κατηγοριοποίηση της τιμής του γνωρίσματος A_i στην κατηγορία l_i , $C[A_i, l_i, t_k] = \mu_{li}(t_k, A_i)$. Έτσι τα σύνολα δεδομένων μετατρέπονται σε σύνολα CVS δηλαδή σε σύνολα της μορφής, $CVS = \{l_i, A_i, \mu_{li}(t_k, A_i)\}$.

- *Εκτίμηση της πληροφορίας που περιέχεται στο κύβο δηλαδή στο σύνολο CVS για κάθε μοντέλο που προκύπτει από την εφαρμογή διαφορετικών clustering μεθόδων.* Το καλύτερο μοντέλο για εξαγωγή γνώσης από την βάση μας είναι αυτό

που περιέχει το μεγαλύτερο ποσοστό πληροφορίας (ενέργειας) ανά γνώρισμα ή συνολικά.

Το σχήμα 6.1 παρουσιάζει τα βασικά βήματα της διαδικασίας fuzzy clustering και classification.



Σχήμα 6.1. Βήματα διαδικασίας Fuzzy Clustering

6.3 ΤΕΧΝΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΕΡΓΑΣΙΑΣ

Με βάση την μεθοδολογία που περιγράψαμε παραπάνω προχωρήσαμε σε μία πρώτη προσέγγιση υλοποίησης ενός συστήματος clustering που θα εμπεριέχει στοιχεία ασαφούς λογικής. Ειδικότερα, προσπαθήσαμε να εξετάσουμε την περίπτωση μίας διαδικασίας clustering η οποία θα βασιζόταν σε έναν crisp clustering αλγόριθμο ενώ θα λάμβανε υπόψη της θέματα ποιότητας clustering και ενσωμάτωσης της αβεβαιότητας.

6.3.1 ΑΛΓΟΡΙΘΜΟΣ CLUSTERING

Ο αλγόριθμος clustering που επιλέχθηκε να υλοποιηθεί στην εφαρμογή μας για την εκτέλεση του crisp clustering είναι ο *K-Means* αλγόριθμος. Οι λόγοι που μας οδήγησαν στην επιλογή του συγκεκριμένου αλγορίθμου είναι οι εξής:

- ◆ Αποτελεί έναν από τους πιο δημοφιλείς αλγορίθμους clustering, ο οποίος έχει εφαρμοστεί σε πολλές εργασίες εύρεσης και καθορισμού τμημάτων σε μεγάλα σύνολα δεδομένων. Επίσης πολλές μελέτες που έχουν γίνει σε θέματα clustering έχουν βασιστεί στον αλγόριθμο K-Means.
- ◆ Είναι κατανοητός και καλά τεκμηριωμένος καθώς βασίζεται στην κλασική θεωρία συνόλων.
- ◆ Η λογική του αλγορίθμου είναι ανάλογη αυτής του αλγορίθμου fuzzy clustering FCM. Θα μπορούσαμε να πούμε ότι ο FCM είναι μία επέκταση του K-Means για fuzzy clustering. Ετσι υλοποιώντας το K-Means και εφαρμόζοντας την μεθοδολογία μας για εισαγωγή ασάφειας σε αποτελέσματα που προέκυψαν από crisp clustering διαδικασία, θα μπορούσαμε να τα συγκρίνουμε μελλοντικά με αποτελέσματα που λαμβάνουμε από την εφαρμογή ενός fuzzy clustering αλγόριθμου όπως ο FCM. Η αναλογικότητα λοιπόν που παρουσιάζουν οι δύο αλγόριθμοι θα μπορούσε να μας βοηθήσει να καταλήξουμε σε χρήσιμα συμπεράσματα σχετικά με την ενσωμάτωση της ασάφειας σε μία διαδικασία crisp clustering.

Τα βασικά βήματα του αλγορίθμου που υλοποιήσαμε περιγράφονται στο **Κεφάλαιο 2** (παρ. 2.3.1). Ένα βασικό στοιχείο που θα πρέπει να επισημάνουμε είναι ότι ο αλγόριθμος υλοποιήθηκε για n -διάστατα δεδομένα ($x = (x_1, x_2, \dots, x_n)$). Μπορούμε δηλαδή να εφαρμόσουμε clustering σε Βάσεις Δεδομένων λαμβάνοντας υπόψη περισσότερα από ένα γνωρίσματα.

6.3.2 SCALING ΔΕΔΟΜΕΝΩΝ

Όταν εφαρμόζουμε clustering σε πολυδιάστατα δεδομένα έχουμε να διαχειριστούμε διαφορετικά γνωρίσματα που μετρώνται σε διαφορετικές μονάδες μέτρησης. Η διαφορά που παρατηρείται μεταξύ των γνωρισμάτων που αποτελούν τα δεδομένα μας μπορεί να επηρεάσουν σημαντικά τα αποτελέσματα του clustering. Για το λόγο αυτό υιοθετήσαμε στον αλγόριθμο μας μία διαδικασία κανονικοποίησης των δεδομένων. Η διαδικασία αυτή είναι γνωστή ως scaling (παρ. 2.9) και συμβάλλει στο να φέρουμε τις τιμές κάθε διάστασης (γνώρισμα) των δεδομένων μας σε συγκρίσιμα διαστήματα.

Με τον τρόπο αυτό προσπαθούμε να μειώσουμε την πιθανότητα μεταβολές μίας διάστασης να εμφανιστούν ως περισσότερο σημαντικές από ότι μεταβολές κάποιας άλλης.

6.3.3 ΕΠΙΛΟΓΗ ΚΑΛΥΤΕΡΟΥ ΣΧΗΜΑΤΟΣ CLUSTERING

Συνήθως οι αλγόριθμοι clustering μας δίνουν την τμηματοποίηση του συνόλου των δεδομένων για έναν καθορισμένο αριθμό clusters. Συνεπώς ορίζονται διαφορετικούς αριθμούς clusters μπορούμε να έχουμε διαφορετική τμηματοποίηση του ίδιου συνόλου των δεδομένων. Θα πρέπει επομένως να καθορίσουμε ποιος είναι ο καλύτερος αριθμός clusters στα οποία μπορούμε να διαχωρίσουμε τα δεδομένα μας, δεδομένου ότι ο στόχος μας είναι clusters πυκνά και καλά διαχωρισμένα. Υιοθετώντας κάποια μέτρα ποιότητας των clusters μπορούμε να προσδιορίσουμε τον αριθμό των clusters που θα μας δώσει το καλύτερο σχήμα για κάθε μεθοδολογία clustering.

6.3.3.1 ΟΡΙΣΜΟΣ ΣΥΝΑΡΤΗΣΗΣ ΕΚΤΙΜΗΣΗΣ ΠΟΙΟΤΗΤΑΣ CRISP CLUSTERING

Για τον ορισμό ενός αξιόπιστου μέτρο αξιολόγησης των clusters θα πρέπει να λάβουμε υπόψη μας την πυκνότητα των clusters, δηλαδή πόσο κοντά στο κέντρο του cluster είναι τα στοιχεία του καθώς και ότι τα clusters μεταξύ τους θα πρέπει να είναι καλά διαχωρισμένα. Με βάση τις διάφορες προσεγγίσεις που έχουν προταθεί κατά καιρούς για την εκτίμηση της ποιότητας των clusters προσπαθήσαμε να ορίσουμε ένα μέτρο ποιότητας clustering για το σύστημά μας. Ετσι στα πλαίσια της υλοποίησης του συστήματος θεωρήσαμε δύο προσεγγίσεις σχετικά με τα μέτρα ποιότητας τις οποίες περιγράφουμε στην συνέχεια.

1. Πρώτη προσέγγιση

Θεωρούμε ένα σύνολο δεδομένων $X = \{x_j: j=1,2,\dots,n\}$ τα οποία έχουν τμηματοποιηθεί σε c clusters με την εφαρμογή κάποιας προσέγγισης clustering.

Εστω s_i είναι η διακύμανση των δεδομένων μέσα στο i cluster η οποία μπορεί να υπολογιστεί ως η μέση τιμή του αθροίσματος των τετραγώνων των διαφορών των στοιχείων του cluster από το κέντρο του. Δηλαδή

$$\sigma_i = \frac{\sum_{k=1}^{n_i} \|x_k - v_i\|^2}{n_i} \quad (\text{Εξισ. 6.1})$$

όπου v_i είναι το κέντρο του cluster i και n_i ο αριθμός των στοιχείων του cluster i .

Η συνολική διακύμανση του συνόλου των δεδομένων σε σχέση με την τμηματοποίηση τους σε c clusters θα είναι

$$\sigma = \sum_{i=1}^c \sigma_i \quad (\text{Εξισ. 6.2})$$

Ένα καλό clustering του συνόλου των δεδομένων έχει σαν αποτέλεσμα μία μικρή τιμή για τη διακύμανση σ .

Το πηλίκο της συνολικής διακύμανσης με το συνολικό αριθμό των clusters ορίζει την μέση πυκνότητα των clusters για την συγκεκριμένη τμηματοποίηση

$$\text{πυκνότητα} = \sigma/c \quad (\text{Εξισ. 6.3})$$

Λαμβάνοντας υπόψη την συνολική διακύμανση του συνόλου δεδομένων μπορούμε να ορίσουμε την μέση διαφοροποίηση της πυκνότητας για τα c clusters. Δηλαδή

$$\text{comp_scat}(c) = \text{πυκνότητα} / \|\sigma(X)\| \quad (\text{Εξισ. 6.4})$$

όπου $\sigma(X)$ είναι η συνολική διακύμανση του συνόλου X . Η ρ διάσταση του $\sigma(X)$ δίνεται από τον τύπο

$$\sigma_x^p = \frac{1}{n} \sum_{k=1}^n \left(x_k^p - \bar{x}^p \right)^2$$

$$\text{με } \bar{x}^p \text{ η } \rho \text{ διάσταση του } \bar{X} = \frac{1}{n} \sum_{k=1}^n x_k, \forall x_k \in X$$

Η τιμή του $\text{comp_scat}(c)$ μετρά πόσο πυκνό είναι κάθε cluster. Όσο πιο πυκνά είναι τα clusters, τόσο πιο μικρό είναι το $\text{comp_scat}(c)$, το οποίο είναι μία συνάρτηση του πως θα κατανείμουμε τα δεδομένα σε clusters. Συνεπώς, για ένα δεδομένο σύνολο δεδομένων μία μικρή τιμή της διαφοροποίησης της πυκνότητας μέσα στα clusters δείχνει ότι έχουμε πετύχει μία τμηματοποίηση με αρκετά πυκνά clusters και συνεπώς έχουμε μία καλή τμηματοποίηση.

Η απόσταση (διαφοροποίηση) μεταξύ των clusters μπορεί να εκτιμηθεί ως η μέση τιμή των αποστάσεων μεταξύ των κέντρων όλων των clusters, δηλαδή

$$d = \frac{\sum_{i=1}^c \sum_{j=1}^c \|v_i - v_j\|}{c(c-1)} \quad (\text{Εξισ. 6.5})$$

όπου v_i, v_j είναι τα κέντρα των clusters i, j αντίστοιχα. Όσο πιο μεγάλη η τιμή του d τόσο πιο καλά διαχωρισμένα είναι τα clusters μας και συνεπώς τόσο πιο καλή τμηματοποίηση έχουμε επιτύχει.

Μία συνάρτηση για την εκτίμηση της ποιότητας του clustering η οποία θα λαμβάνει υπόψη της τόσο την πυκνότητα όσο και την διαφοροποίηση των clusters μπορεί να οριστεί ως εξής

$$CD = \text{comp_scat}(c)/d. \quad (\text{Εξισ. 6.6})$$

Η τιμή της συνάρτησης μπορεί να χρησιμοποιηθεί για να εκτιμήσουμε την ποιότητα του clustering που εφαρμόστηκε στα δεδομένα μας. Η καλύτερη τιμηματοποίηση του συνόλου των δεδομένων μας θα μας δώσει την ελάχιστη τιμή για το CD.

2. Δεύτερη προσέγγιση

Το δεύτερο μέτρο που θεωρήσαμε βασίζεται στο V_{CWB} [PP98] το οποίο περιγράψαμε στην παράγραφο 4.4.2.1. Ο υπολογισμός όμως της διασποράς $Scat(c)$ τροποποιήθηκε ώστε να βασίζεται στην διακύμανση crisp clusters. Ήτοι το μέτρο ποιότητας ορίζεται ως εξής:

$$SD(c) = a^* Scat(c) + Disc(c), \quad (\text{Εξισ. 6.7})$$

Η παράμετρος a αποσκοπεί στην εξομάλυνση της κλίμακας στην οποία κυμαίνονται οι δύο όροι. Για το σκοπό αυτό λαμβάνεται $a = Disc(c_{max})$ όπου c_{max} ένας αρκετά μεγάλος αριθμός clusters.

Διακύμανση Συνόλου Δεδομένων. Η διακύμανση των στοιχείων του συνόλου των δεδομένων X καλείται $\sigma(X) \in \mathbb{R}^p$. Η τιμή της p διάστασης του $\sigma(X)$ ορίζεται από την εξίσωση (Εξισ. 6.8):

$$\sigma_x^p = \frac{1}{n} \sum_{k=1}^n \left(x_k^p - \bar{x}^p \right)^2 \quad (\text{Εξισ. 6.8})$$

$$\text{με } \bar{X} = \frac{1}{n} \sum_{k=1}^n x_k, \quad \forall x_k \in X$$

Διακύμανση cluster i . Η διακύμανση του cluster i καλείται $\sigma(v_i) \in \mathbb{R}^p$. Η τιμή της p διάστασης του $\sigma(v_i)$ ορίζεται από την εξίσωση (Εξισ. 6.9):

$$\sigma_{v_i}^p = \frac{1}{n_i} \sum_{k=1}^{n_i} \left(x_k^p - v_i^p \right)^2 \quad (\text{Εξισ. 6.9})$$

όπου n_i ο αριθμός των στοιχείων που ανήκουν στο cluster i .

Μέση Διασπορά clusters. Η μέση διασπορά για c clusters ορίζεται από την εξίσωση:

$$Scat(c) = \frac{\frac{1}{c} \sum_{i=1}^c \|\sigma(v_i)\|}{\|\sigma(X)\|} \quad (\text{Εξισ. 6.10})$$

$$\text{οπου } \|x\| = (x^T x)^{1/2}$$

Συνολική διαφοροποίηση αποστάσεων μεταξύ των clusters. Η συνάρτηση για την εκτίμηση της συνολικής διαφοροποίησης της απόστασης μεταξύ των clusters, $Disc(c)$, ορίζεται ως εξής

$$Disc(c) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^c \left(\sum_{z=1}^c \|v_k - v_z\| \right)^{-1} \quad (\text{Εξισ. 6.11})$$

όπου $D_{\max} = \text{maximum} \|v_i - v_j\| \quad \forall i, j \in \{2, 3, \dots, c\}$ είναι η μέγιστη απόσταση ανάμεσα στα κέντρα των clusters. Η $D_{\min} = \text{minimum} \|v_i - v_j\| \quad \forall i, j \in \{2, 3, \dots, c\}$ είναι η ελάχιστη απόσταση ανάμεσα στα κέντρα των clusters.

Ο αριθμός των clusters, ο οποίος ελαχιστοποιεί το δείκτη εγκυρότητας μπορεί να ληφθεί σαν βέλτιστη τιμή για τον αριθμό των κλάσεων(clusters) των δεδομένων.

6.3.3.1.1 ΣΥΓΚΡΙΣΗ ΜΕΤΡΩΝ

Τα δύο μέτρα ποιότητας που περιγράψαμε παραπάνω λαμβάνουν υπόψη τους τα δύο βασικά κριτήρια αξιολόγησης των clusters. Βασίζονται δηλαδή και τα δύο τόσο στην διασπορά των στοιχείων μέσα στα clusters όσο και την απόσταση μεταξύ των διαφορετικών clusters. Προκειμένου όμως να επιλέξουμε εκείνο το μέτρο που θα οδηγούσε σε καλύτερα και πιο αξιόπιστα αποτελέσματα για την εφαρμογή μας προχωρήσαμε σε συγκριτική μελέτη αυτών. Η μελέτη αυτή οδήγησε στα εξής συμπεράσματα:

Πρώτη προσέγγιση

- Η τιμή της συνάρτησης έχει μονότονη εξάρτηση από τον αριθμό των clusters. Ειδικότερα παρατηρείται σημαντική μείωση της τιμής της με την αύξηση του αριθμού των clusters. Στο συμπέρασμα αυτό μπορούμε να καταλήξουμε εάν λάβουμε υπόψη μας τον ορισμό του μέτρου και τις μεταβολές της πυκνότητας και των αποστάσεων μεταξύ clusters καθώς μεταβάλλεται ο αριθμός clusters. Η θεωρητική απόδειξη του συμπεράσματος αυτού παρουσιάζεται στο Παράρτημα B'.
- Το μέτρο βασίζεται στην μέση απόσταση μεταξύ των clusters, με αποτέλεσμα σε αρκετές περιπτώσεις να οδηγεί σε σχήματα clustering στα οποία παρατηρείται μεγάλη απόκλιση μεταξύ των αποστάσεων των clusters.

Δεύτερη προσέγγιση

- Το μέτρο της δεύτερης προσέγγισης βασίζεται στην διαφοροποίηση απόστασης μεταξύ των clusters. Με τον τρόπο αυτό αποτρέπει κατά κάποιο τρόπο την επιλογή σχημάτων clustering στα οποία θα υπάρχουν clusters τα οποία θα απέχουν αρκετά μεταξύ τους ενώ κάποια άλλα θα είναι πολύ κοντά. Συμβάλλει δηλαδή στην επιλογή clusters των οποίων οι αποστάσεις δεν παρουσιάζουν πολύ μεγάλες αποκλίσεις.
- Ο παράγοντας εξομάλυνση α που χρησιμοποιείται στο μέτρο εξαρτάται από την μέγιστη τιμή που έχει οριστεί για τον αριθμό των clusters. Αυτό έχει σαν αποτέλεσμα οι τιμές του μέτρου ποιότητας να επηρεάζονται από το άνω όριο που ορίζεται για τον αριθμό των clusters.

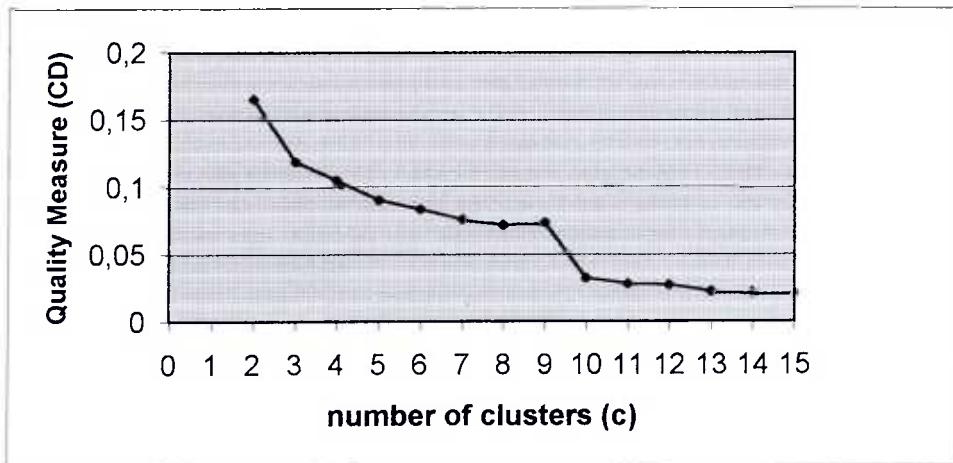
6.3.3.1.2 ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ

Προκειμένου να επιβεβαιώσουμε και πειραματικά τις παραπάνω παρατηρήσεις εφαρμόσαμε σε ενδεικτικά σύνολα δεδομένων τα παραπάνω μέτρα. Συγκεκριμένα εφαρμόστηκαν σε σύνολα δεδομένων με στοιχεία που αφορούσαν τις Χρηματιστηριακές συναλλαγές που πραγματοποιήθηκαν κατά τις ημέρες 12/1/98 έως 16/1/98.

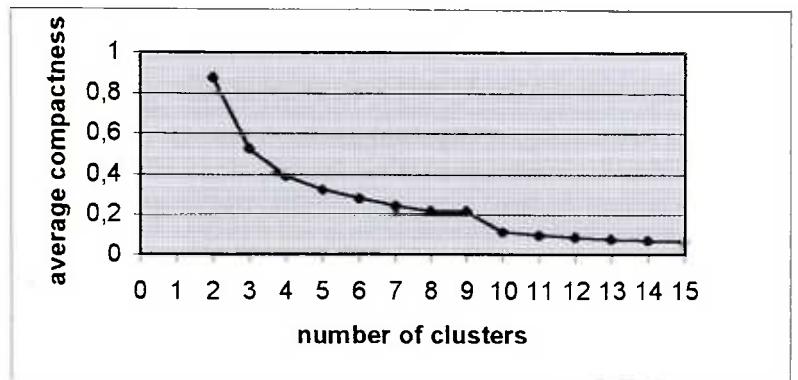
- *To μέτρο ποιότητας CD, 1^η προσέγγιση, μειώνεται με την αύξηση του αριθμού των clusters.*

Τα Διαγράμματα 6.1, 6.2 παρουσιάζουν τα αποτελέσματα εκτίμησης της ποιότητας με βάση το μέτρο ποιότητας CD σε σύνολα δεδομένων με διδιάστατα και μονοδιάστατα στοιχεία αντίστοιχα. Από τα διαγράμματα αυτά προκύπτει ότι η τιμή του μέτρου ποιότητας μειώνεται σημαντικά με την αύξηση του αριθμού των clusters.

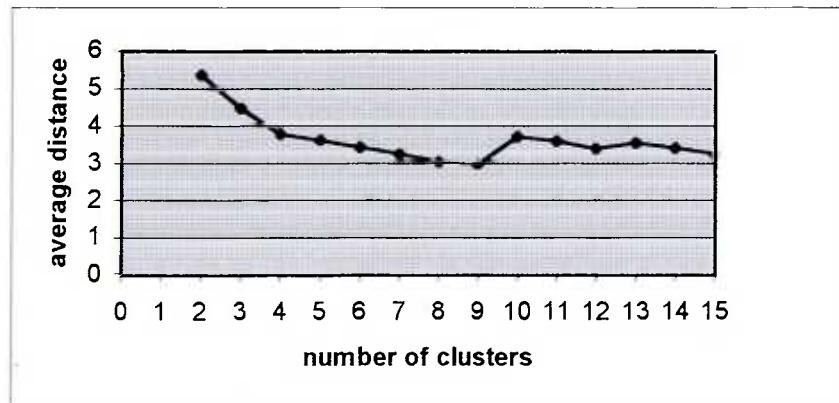
Επίσης, τα Διαγράμματα 6.1α, β και 6.2α, β παρουσιάζουν τις μεταβολές της διασποράς και της μέσης απόστασης μεταξύ των clusters σε σχέση με τον αριθμό των clusters. Παρατηρώντας τα διαγράμματα μπορούμε να καταλήξουμε στο συμπέρασμα ότι η μέση πυκνότητα παρουσιάζει μονότονη μείωση με την αύξηση του αριθμού των clusters. Η μέση απόσταση παρουσιάζει επίσης κάποια μείωση με την αύξηση του αριθμού των clusters μικρότερη όμως σε σχέση με την πυκνότητα και με κάποιες αυξομειώσεις. Θα μπορούσαμε λοιπόν να πούμε ότι η μείωση της μέσης πυκνότητας των clusters είναι ιδιαίτερα σημαντική σε σχέση με αυτή της μέσης απόστασης, γεγονός που επηρεάζει και το μέτρο ποιότητας.



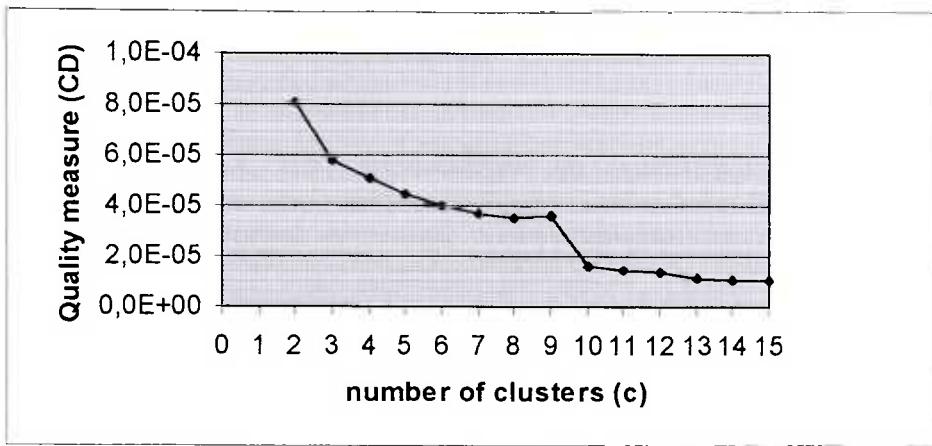
Διάγραμμα 6.1. Διάγραμμα μεταβολής του μέτρου ποιότητας clustering σε σχέση με την μεταβολή των clusters. Το clustering αφορά σε διδιάστατα στοιχεία με την μορφή (τιμή κλεισίματος μετοχής, υψηλότερη τιμή μετοχής).



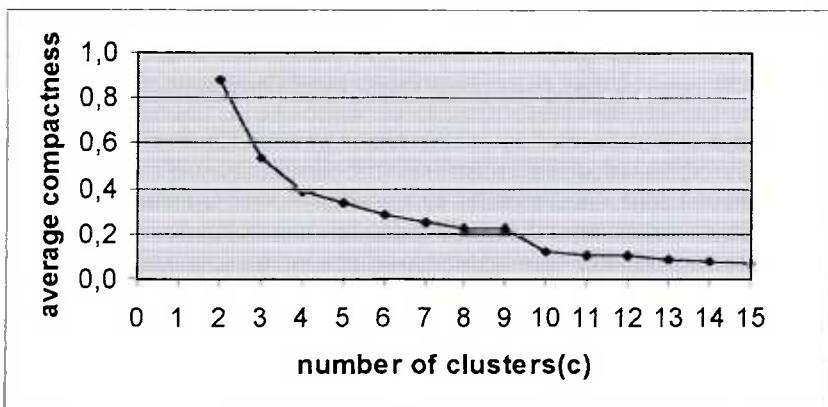
Διάγραμμα 6.1α. Μεταβολή της μέσης πυκνότητας μέσα στα clusters σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε σύνολο δεδομένων του οποίου τα στοιχεία είχαν την μορφή (τιμή κλεισίματος μετοχής, μέγιστη τιμή μετοχής).



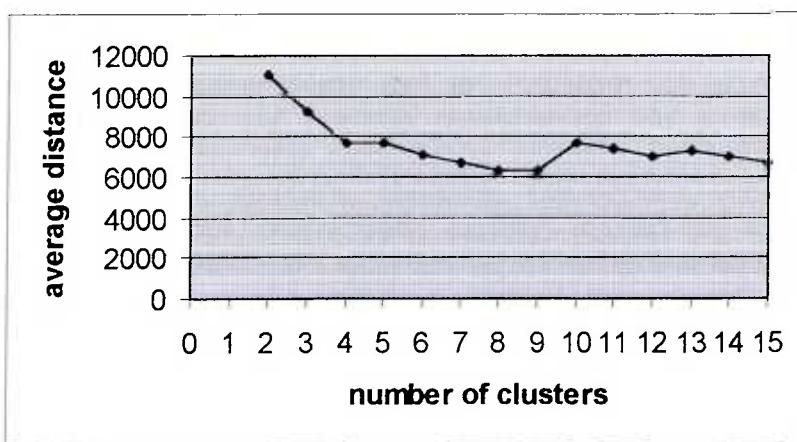
Διάγραμμα 6.1β. Διάγραμμα που παρουσιάζει την μεταβολή της μέσης απόστασης μεταξύ των clusters σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε σύνολο δεδομένων του οποίου τα στοιχεία είχαν την μορφή (τιμή κλεισίματος μετοχής, μέγιστη τιμή μετοχής).



Διάγραμμα 6.2. Διάγραμμα που παρουσιάζει την μεταβολή του μέτρου ποιότητας clustering σε σχέση με την μεταβολή των clusters. Το clustering αφορά σε μονοδιάστατα στοιχεία με την μορφή (τιμή κλεισίματος μετοχής).



Διάγραμμα 6.2α. Μεταβολή της μέσης πυκνότητας μέσα στα clusters σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε σύνολο δεδομένων του οποίου τα στοιχεία είχαν την μορφή (τιμή κλεισίματος μετοχής).



Διάγραμμα 6.2β. Διάγραμμα που παρουσιάζει την μεταβολή της μέσης απόστασης μεταξύ των clusters σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε σύνολο δεδομένων του οποίου τα στοιχεία είχαν την μορφή (τιμή κλεισίματος μετοχής).

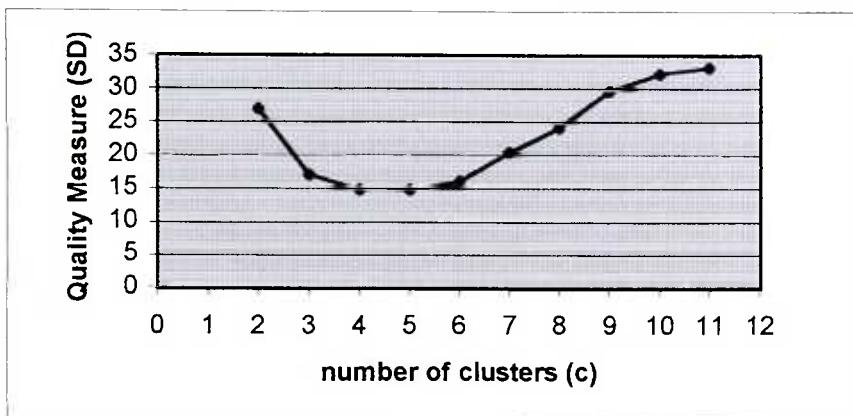
- *To μέτρο ποιότητας SD, 2nd προσέγγιση, αξιολογεί τα σχήματα clustering εξισορροπώντας την διαφοροποίηση των αποστάσεων μεταξύ των clusters με την διασπορά μέσα στα clusters.*

Τα Διαγράμματα 6.4, 6.5 παρουσιάζουν τα αποτελέσματα εκτίμησης της ποιότητας με βάση το μέτρο ποιότητας SD σε σύνολα δεδομένων με διδιάστατα και μονοδιάστατα στοιχεία αντίστοιχα. Από τα διαγράμματα αυτά προκύπτει ότι το μέτρο ποιότητας δεν επηρεάζεται από τον αριθμό των clusters με τρόπο ώστε να εμφανίζει μονότονη μείωση της τιμής του με την αύξηση του αριθμού των clusters, όπως συμβαίνει στο μέτρο της 1st προσέγγισης. Ήα μπορούσαμε να πούμε ότι γίνεται συνεκτίμηση των δύο κριτηρίων ποιότητας, της πυκνότητας των clusters και της απόστασης μεταξύ αυτών.

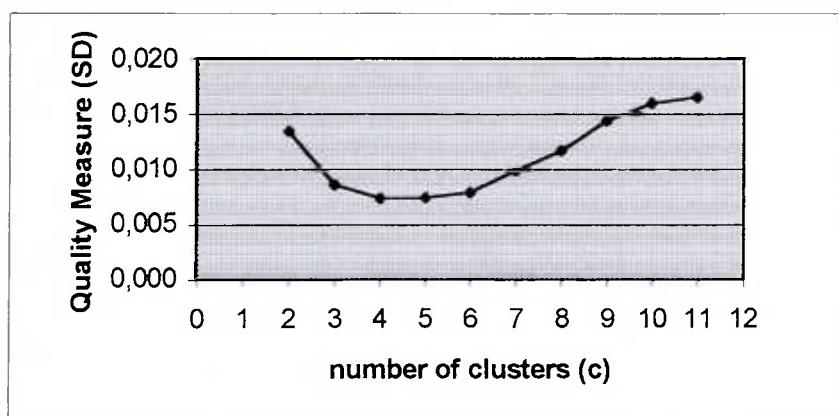
- *To μέτρο ποιότητας SD, επηρεάζεται από την μέγιστη τιμή που ορίζεται για τον αριθμό των clusters.*

Τα Διαγράμματα 6.6, 6.7 παρουσιάζουν την επίδραση που έχει στις τιμές του μέτρου ποιότητας η μεταβολή της μέγιστης τιμής για τον αριθμό των clusters. Από τα διαγράμματα αυτά προκύπτει ότι η τιμή του μέτρου SD μειώνεται με την μείωση του άνω ορίου του αριθμού των clusters. Η μείωση της μεταβολής των τιμών έχει σαν αποτέλεσμα σε πολλές περιπτώσεις την μείωση και της βέλτιστης τιμής για τον αριθμό των clusters. Για παράδειγμα, στο Διάγραμμα 6.6 όταν το άνω όριο για τον αριθμό των clusters είναι 11 η βέλτιστη τιμή είναι 5, ενώ όταν το άνω όριο μειώνεται σε 10 η βέλτιστη τιμή για τον αριθμό των clusters ορίζεται στο 4. Επίσης, παρατηρούμε στο ίδιο διάγραμμα ότι η μεταβολή της τιμής του άνω ορίου των clusters από 10 έως 8 δεν επηρεάζει την βέλτιστη τιμή για τον αριθμό των clusters, ενώ για τις τιμές 7 και 6 η βέλτιστη τιμή καθορίζεται σε 3. Το φαινόμενο αυτό οφείλεται στο γεγονός ότι το SD κλιμακώνει τις τιμές των δύο μέτρων που το αποτελούν, την πυκνότητα και την διαφοροποίηση των αποστάσεων, με βάση την μέγιστη τιμή που ορίζεται για τον αριθμό των clusters. Ήα μπορούσαμε όμως να αντιμετωπίσουμε το πρόβλημα αυτό εάν καθορίσουμε ως μέγιστη τιμή για τον αριθμό των clusters μία σχετικά μεγάλη τιμή(ανάλογα με την εφαρμογή μας).

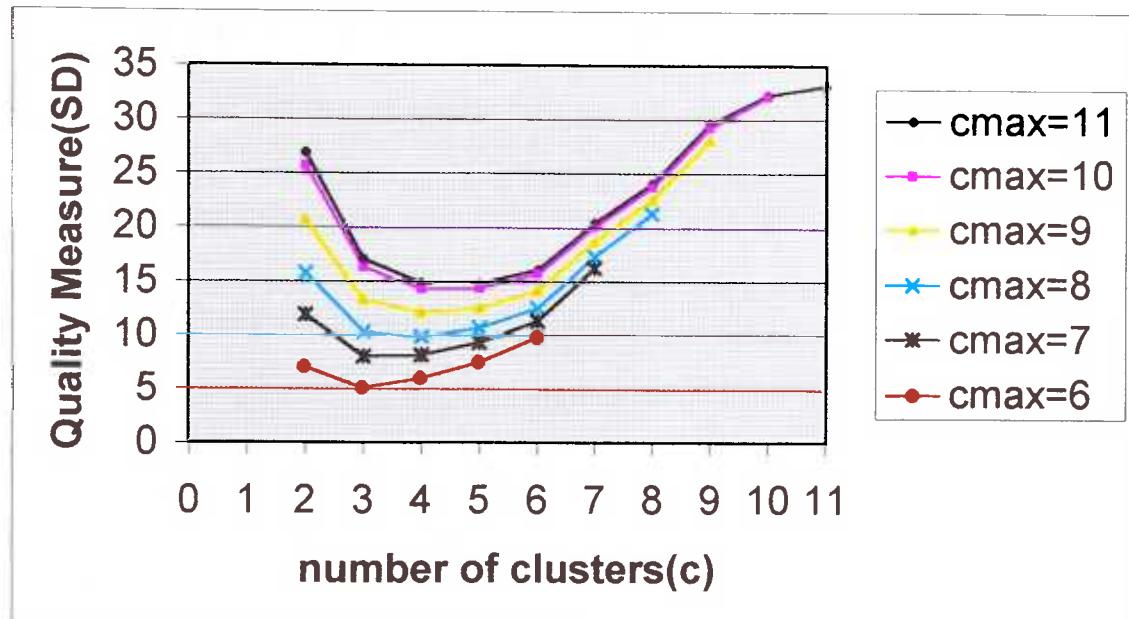
Με βάση τις παραπάνω παρατηρήσεις από την εφαρμογή των μέτρων ποιότητας καταλήξαμε στην επιλογή της δεύτερης προσέγγισης. Ήα πρέπει βέβαια να σημειώσουμε ότι η μελέτη που περιγράψαμε παραπάνω αποτελεί μία πρώτη προσέγγιση στο θέμα της ποιότητας clustering. Οπωσδήποτε υπάρχουν κάποια θέματα τα οποία απαιτούν εκτενέστερης μελέτης σχετικά με τα μέτρα που περιγράψαμε όπως για παράδειγμα η επίδραση της παραμέτρου α , θέμα το οποίο δεν έχει καθοριστεί και στο αρχικό μέτρο VCwb [RR98] στο οποίο βασίζεται και η προσέγγισή μας.



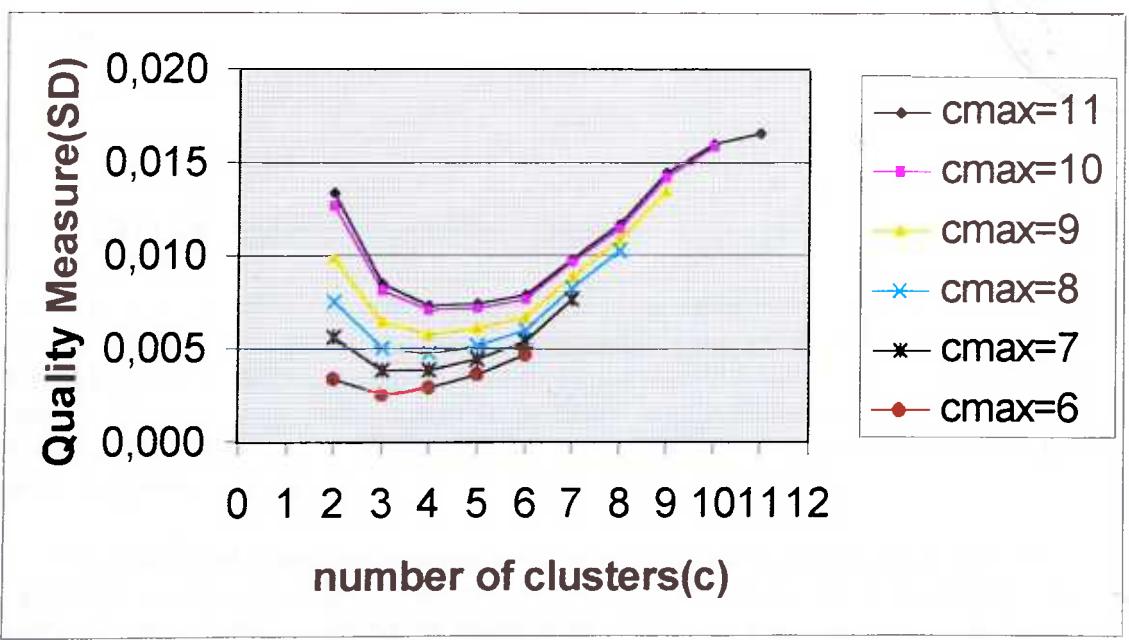
Διάγραμμα 6.4. Μεταβολή του μέτρου ποιότητας SD σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε ένα σύνολο δεδομένων με διδιάστατα στοιχεία (τιμή κλεισίματος μετοχής, υψηλότερη τιμή μετοχής).



Διάγραμμα 6.5. Μεταβολή του μέτρου ποιότητας SD σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε ένα σύνολο δεδομένων με μονοδιάστατα στοιχεία (τιμή κλεισίματος μετοχής).



Διάγραμμα 6.6. Επίδραση της μέγιστης τιμής του αριθμού των clusters στην τιμή του μέτρου ποιότητας. Το clustering αφορά στοιχεία διδιάστατα (τιμή κλεισίματος μετοχής, υψηλότερη τιμή μετοχής).



Διάγραμμα 6.7. Επίδραση της μέγιστης τιμής του αριθμού των clusters στην τιμή του μέτρου ποιότητας. Το clustering αφορά στοιχεία μονοδιάστατα (τιμή κλεισίματος μετοχής).

6.3.3.2 ΕΥΡΕΣΗ ΒΕΛΤΙΣΤΟΥ ΑΡΙΘΜΟΥ CLUSTERS ΠΟΥ ΠΡΟΚΥΠΤΟΥΝ ΑΠΟ ΑΛΓΟΡΙΘΜΟΥΣ CRISP CLUSTERING.

Εάν ορίσουμε τη μέγιστη, ελάχιστη τιμή για τον αριθμό των clusters μπορούμε εφαρμόζοντας τον παρακάτω αλγόριθμο να προσδιορίσουμε το βέλτιστο αριθμό clusters (c_{opt}) για το σύνολο των δεδομένων μας χρησιμοποιώντας ως κριτήριο ένα μέτρο αξιολόγησης clustering. Στην περίπτωση μας χρησιμοποιούμε το SD.

Τα βασικά βήματα που θα πρέπει να ακολουθηθούν κατά την διαδικασία clustering ενός συνόλου δεδομένων μπορούν να συνοψιστούν στα εξής:

1. Ορίζουμε την μέγιστη τιμή c_{max} και ελάχιστη τιμή c_{min} του αριθμού των clusters.
2. Αρχικοποίηση: $c = c_{max}$, $c_{opt} = c$;
3. Εφαρμογή αλγορίθμου clustering στο σύνολο των δεδομένων, με σκοπό να καθοριστούν τα clusters.
4. If ($c=c_{max}$)
 { a = Disc(c), indexValue=SD(c) }
 else if (Μέτρο_ποιότητας(c)<indexValue)
 { $c_{opt}=c$; indexValue = SD(c); }
5. $c=c-1$,
 if ($c=c_{min}-1$)
 stop
 else
 goto 3.

6.3.4 ΣΥΝΑΡΤΗΣΕΙΣ ΣΥΜΜΕΤΟΧΗΣ

Οι γενικοί αλγόριθμοι clustering (crisp clustering) στοχεύουν στην διαίρεση ενός συνόλου δεδομένων σε clusters με συγκεκριμένα όρια με βάση κάποια κριτήρια εκτίμησης της ομοιότητας μεταξύ των στοιχείων του συνόλου δεδομένων. Τα όρια μεταξύ των clusters που προκύπτουν από τους αλγορίθμους αυτούς είναι συγκεκριμένα. Αυτό σημαίνει ότι κάθε δείγμα από τα δεδομένα ενός cluster ανήκει σε αυτό και μόνο το cluster.

Ένα πρόβλημα όμως που προκύπτει κατά την προσέγγιση αυτή είναι κατά πόσο μπορούμε να είμαστε βέβαιοι για την κατανομή των στοιχείων στα clusters που έχουν οριστεί κυρίως όταν αυτά βρίσκονται στα όρια των πεδίων ορισμού των clusters. Για το λόγο αυτό θέλουμε να εισάγουμε την αβεβαιότητα στην διαδικασία κατανομής των στοιχείων στα clusters, ορίζοντας συναρτήσεις συμμετοχής για κάθε ένα από τα clusters που έχουν οριστεί. Οι συναρτήσεις συμμετοχής που θα χρησιμοποιηθούν για την αντιστοίχηση των στοιχείων στα clusters θα βασίζονται στις *hypertrapezoidal fuzzy membership functions*. Τα κύρια στοιχεία των συναρτήσεων αυτών, τα οποία συνετέλεσαν στην υιοθέτηση τους για τον υπολογισμό των βαθμών συμμετοχής στην εφαρμογή μας είναι ότι:

- ◆ μπορούν να χρησιμοποιηθούν επιτυχώς στον υπολογισμό βαθμών συμμετοχής σε πολυδιάστατα σύνολα και
- ◆ οι συναρτήσεις αυτές χρησιμοποιούν για τον ορισμό τους τα πρότυπα(κέντρα) των clusters καθώς και έναν συντελεστή επικάλυψης, σ , ο οποίος δηλώνει το βαθμό επικάλυψης μεταξύ γειτονικών clusters.

6.3.4.1 HYPERTRAPEZOIDAL FUZZY MEMBERSHIP FUNCTIONS

Οι *hypertrapezoidal fuzzy membership functions* έχουν προταθεί[KP96] ως ένας κατάλληλος μηχανισμός για την αναπαράσταση και τον υπολογισμό πολυδιάστατων ασαφών συνόλων. Μία από τις βασικές υποθέσεις στην ανάπτυξη της τεχνικής αυτής είναι η απαίτηση ότι το άθροισμα των πολυδιάστατων ασαφών συνόλων ισούται με την μονάδα, σύμφωνα με την παρακάτω ισότητα:

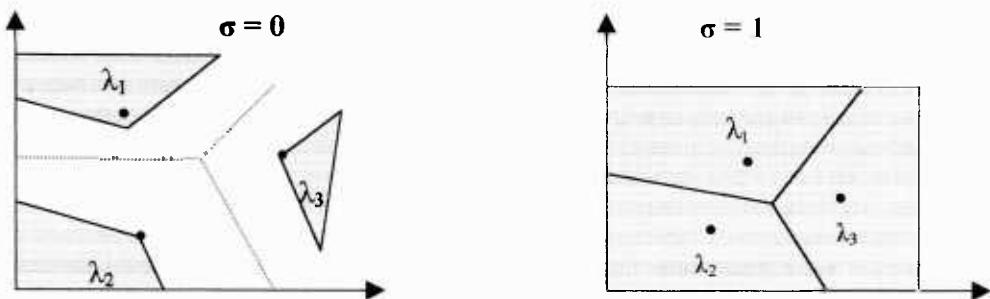
$$\sum_i \mu_i(\bar{x}) = 1, \quad \forall \bar{x} \quad (\text{Εξισ. 6.12})$$

Οι συναρτήσεις συμμετοχής ορίζονται με έναν τρόπο που χαρακτηρίζεται ως *fuzzy partitioning* σε καθορισμένο χώρο και είναι συνεπείς με την Bayesian ερμηνεία των ασαφών συνόλων. Ένα ακόμα σημαντικό στοιχείο είναι ότι οι n -διάστατες συναρτήσεις συμμετοχής ορίζονται με λίγες μόνο παραμέτρους. Μία μονοδιάστατη ασαφής συνάρτηση συμμετοχής έχει την μορφή τ_{λ_i} και μπορούμε να την ορίσουμε με τέσσερα μόνο σημεία, εάν επεκτείνουμε όμως τον ορισμό της συνάρτησης συμμετοχής σε δύο διαστάσεις βλέπουμε ότι ο ορισμός της γίνεται μη πρακτικός. Επίσης προσπαθώντας να ορίσουμε τις γωνίες ενός n -διάστατου ασαφούς συνόλου με τρεις ή περισσότερες εισόδους βλέπουμε ότι η διαδικασία αυτή είναι πρακτικά αδύνατο να επιτευχθεί, ειδικά εάν πρέπει τα σύνολα μας να ικανοποιούν την ισότητα (1).

Μία εναλλακτική λύση την οποία υιοθετούν οι *hypertrapezoidal membership functions* είναι να χρησιμοποιήσουμε κάποια σημεία στο χώρο τα οποία θα αποτελούν τους αντιροσώπους (πρότυπα σημεία) των n -διάστατων ασαφών συνόλων (clusters στην περίπτωση μας). Τα σημεία αυτά αποτελούν παραμέτρους της ασαφούς τμηματοποίησης. Εάν λ_i είναι ο αντιρόσωπος για ένα σύνολο S_i , με συνάρτηση συμμετοχής $\mu_i(x)$, τότε θα ισχύουν οι εξής ισότητες:

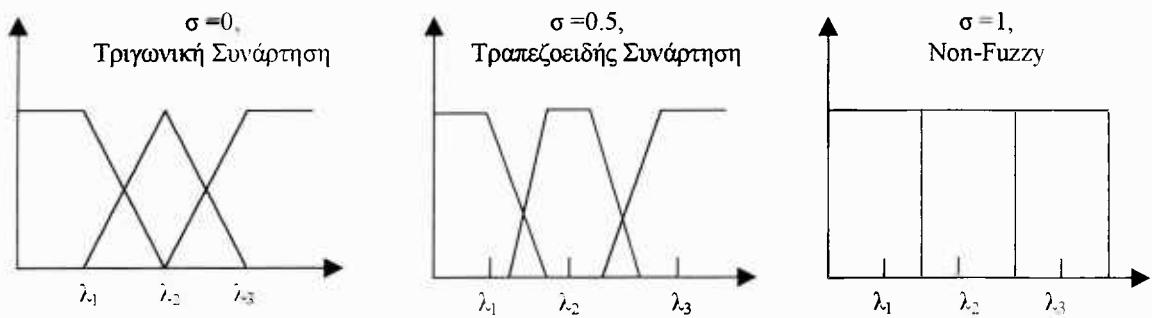
$$\begin{aligned} \mu_i(\lambda_i) &= 1 \\ \mu_j(\lambda_i) &= 0, \quad j \neq i \end{aligned} \quad (\text{Εξισ. 6.13})$$

Μία επιπρόσθετη παράμετρος η οποία χρησιμοποιείται για τον ορισμό μίας n -διάστατης ασαφούς τμηματοποίησης είναι ο *crispness factor*, ο οποίος καθορίζει πόση επικάλυψη υπάρχει ανάμεσα στα πρότυπα σημεία δύο γειτονικών ασαφών συνόλων. Ο παράγοντας αυτός, σ , κυμαίνεται στο διάστημα $[0, 1]$. Για $\sigma = 1$, δεν υπάρχει επικάλυψη ανάμεσα στα σύνολα. Το σχήμα 6.2 παρουσιάζει τα αποτελέσματα της τμηματοποίησης για τις δύο ακραίες τιμές του σ ($\sigma = 0$ και $\sigma = 1$).



Σχήμα 6.2 . Επίδραση του *crispness factor*(σ) σε διδιάστατα σύνολα.

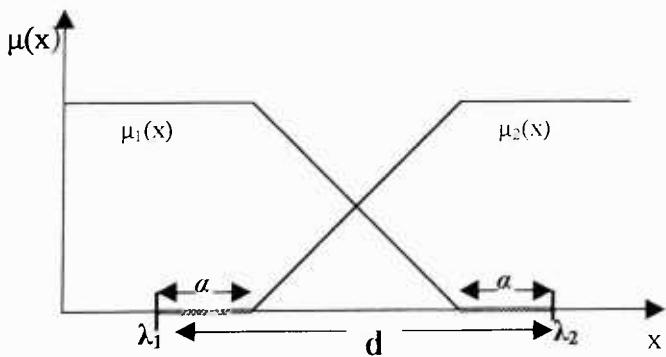
Ένα τέτοιο σχήμα μπορεί να χρησιμοποιηθεί για να ορίσουμε και μονοδιάστατα ασαφή σύνολα. Το σχήμα 6.3 παρουσιάζει πως μεταβάλλοντας τις τιμές του σ σε ένα σύνολο μονοδιάστατων δεδομένων επηρεάζονται οι συναρτήσεις συμμετοχής.



Σχήμα 6.3. Επίδραση παράγοντα σ σε μονοδιάστατα σύνολα

Ο παράγοντας επικάλυψης, σ , ορίζεται από τους Kelly, Painter σύμφωνα με το σχήμα 6.4 και την εξίσωση (Εξισ. 6.14):

$$\sigma = \frac{2a}{d} \quad (\text{Εξισ. 6.14})$$



Σχήμα 6.4. Ορισμός επικάλυψης ασαφών τμημάτων

Με βάση όσο αναφέρθηκαν παραπάνω και τον τρόπο με τον οποίο ορίζονται οι *hypertrapezoidal* συναρτήσεις συμμετοχής [KP96] μπορούμε να υπολογίσουμε την τιμή της συνάρτησης συμμετοχής κάθε στοιχείου στα clusters που ορίστηκαν κατά την εφαρμογή μίας μεθοδολογίας clustering. Ετσι εάν έχουμε τα κέντρα των clusters στα οποία θέλουμε να κατανείμουμε τα δεδομένα μας και τον παράγοντα επικάλυψης σ μπορούμε να υπολογίσουμε το βαθμό συμμετοχής ενός στοιχείου x στα clusters. Τα βήματα που θα ακολουθήσουμε είναι:

- Καθορίζουμε το πρότυπο(κέντρο) κάθε cluster ως την μέση τιμή των στοιχείων που περιέχονται στο cluster.

$$\lambda_i = \frac{\sum_{i=1}^{n_i} x_i}{n_i} \quad (\text{Εξισ. 6.15})$$

όπου n_i ο αριθμός των στοιχείων που περιέχονται στο cluster i .

- Για κάθε στοιχείο x του συνόλου των δεδομένων μας υπολογίζουμε την απόσταση, ρ_{ij} , για κάθε ζεύγος clusters
-

$$\rho_{ij} = \frac{d^2(x, \lambda_i) - d^2(x, \lambda_j)}{d^2(\lambda_i, \lambda_j)} \quad (\text{Εξισ. 6.16})$$

όπου $d(x,y)$ είναι η Ευκλείδεια απόσταση μεταξύ του x και y .

- Υπολογίζουμε τις υποσυνθήκη συναρτήσεις συμμετοχής για κάθε ζεύγος προτύπων. Εάν v_{ji} είναι το διάνυσμα από το λ_i στο λ_j , v_{jx} είναι το διάνυσμα από το λ_j στο x τότε έχουμε:

$$\mu_{i/j}(x) = \begin{cases} 0; & \rho_{ij} \geq 1 - \sigma \\ 1; & \rho_{ij} \leq \sigma - 1 \\ \frac{v_{ji} - \frac{\sigma}{2} d^2(\lambda_j, \lambda_i)}{(1 - \sigma) d^2(\lambda_j, \lambda_i)} & \text{otherwise} \end{cases} \quad (\text{Εξισ. 6.17})$$

Τελικά, ο βαθμός συμμετοχής, $\mu_i(x)$, για ένα στοιχείο x στο ασαφές σύνολο i , δίνεται από την εξίσωση (Εξισ. 6.18):

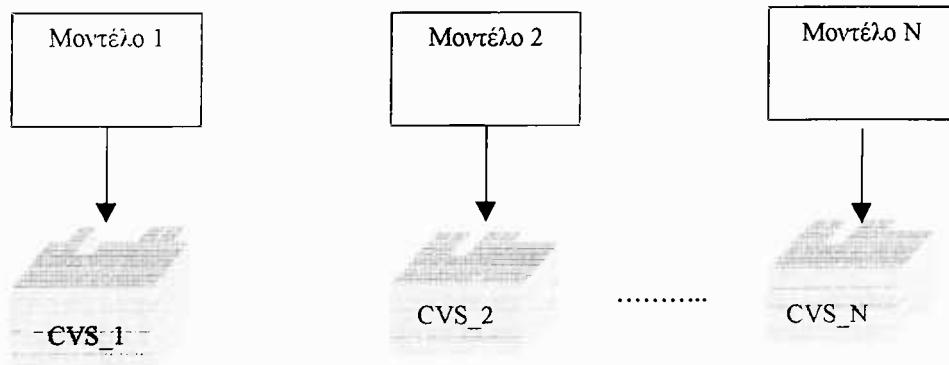
$$\mu_i(x) = \frac{\prod_{j=1, j \neq i}^c \mu_{i,j}(x)}{\sum_{k=1}^c \left(\prod_{j=1, j \neq k}^c \mu_{k,j}(x) \right)} \quad (\text{Εξισ. 6.18})$$

όπου c ο αριθμός των ασαφών συνόλων.

6.3.5 ΕΚΤΙΜΗΣΗ ΠΛΗΡΟΦΟΡΙΑΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ ΣΧΗΜΑΤΩΝ CLUSTERING

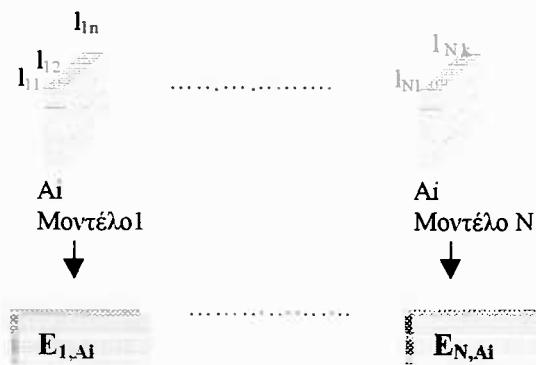
Εάν εφαρμόσουμε όσα αναφέρθηκαν στις προηγούμενες παραγράφους για διαφορετικούς αλγορίθμους clustering θα έχουμε σαν αποτέλεσμα την εύρεση των καλύτερων σχημάτων clustering για κάθε μεθοδολογία clustering που εφαρμόσαμε. Θα πρέπει όμως να επιλέξουμε εκείνο το σχήμα (τμηματοποίηση των δεδομένων) που ανταποκρίνεται καλύτερα στις απαιτήσεις μας και το οποίο περιέχει το μεγαλύτερο πληροφοριακό περιεχόμενο. Συνεπώς, θα πρέπει να εκτιμήσουμε την γνώση ανά γνώρισμα που μας παρέχει το κάθε σχήμα.

Για κάθε σχήμα clustering προσδιορίζουμε τις συναρτήσεις συμμετοχής με βάση των οποίων μπορεί να γίνει classification των δεδομένων μας στα clusters κάθε σχήματος. Από την εφαρμογή της διαδικασίας classification για κάθε σχήμα clustering προκύπτουν διαφορετικά μοντέλα τμηματοποίησης της βάσης μας. Κάθε μοντέλο μπορεί να αναπαρασταθεί με την μορφή ενός κύβου C , $C[A_i, l_i, t_k] = \mu_{li}(t_k, A_i)$.



Σχήμα 6.5. Αντιστοίχηση clustering σχημάτων σε CVS

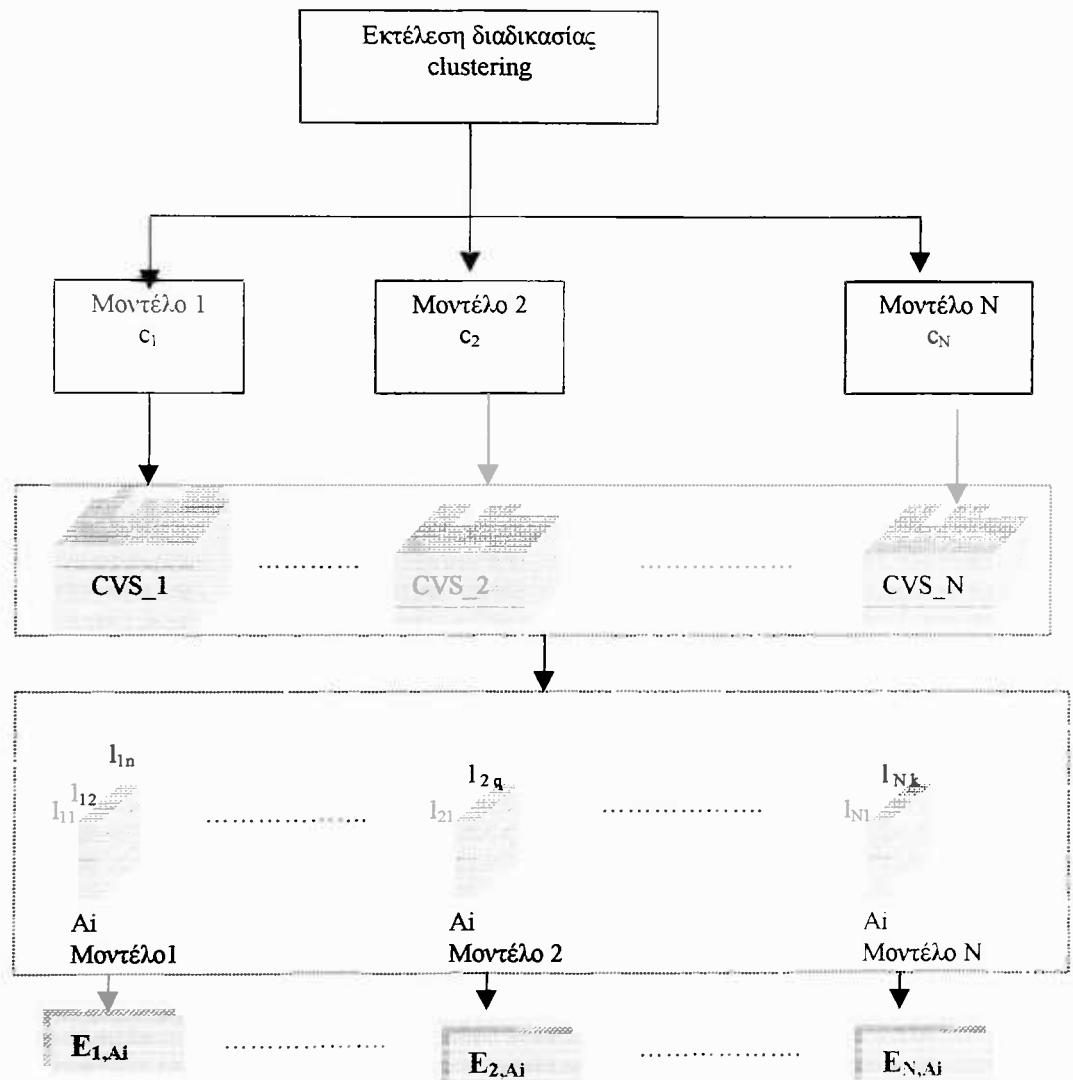
Υπολογίζοντας την ενέργεια ανά γνώρισμα σε κάθε κύβο μπορούμε να επιλέξουμε την καλύτερη δυνατή τμηματοποίηση της βάσης για κάθε γνώρισμα. Για παράδειγμα, έστω ότι θέλουμε να βρούμε την καλύτερη τμηματοποίηση του γνωρίσματος A_i . Για κάθε μοντέλο υπολογίζουμε την συνολική γνώση (E_j, A_i) που υπάρχει σε κάθε κύβο σχετικά με το A_i .



Σχήμα 6.6. Γνώση που περιέχει ο κύβος ανά γνώρισμα

Με βάση τις ενέργειες για το χαρακτηριστικό A_i που προκύπτουν από κάθε μοντέλο μπορούμε να επιλέξουμε αυτό που παρέχει την περισσότερη γνώση αναφορικά με το γνώρισμα.

Το σχήμα 6.7 παρουσιάζει διαγραμματικά την διαδικασία επιλογής του καλύτερου σχήματος clustering που προέρχονται από διαφορετικές μεθοδολογίες με βάση το πληροφοριακό περιεχόμενο που περιέχουν για την εφαρμογή μας.



Σχήμα 6.7. Επιλογή του καλύτερου clustering σχήματος

7^ο ΚΕΦΑΛΑΙΟ

ΕΦΑΡΜΟΓΗ FUZZY CLUSTERING

7.1 ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό θα περιγράψουμε την εφαρμογή που αναπτύχθηκε στα πλαίσια της διπλωματικής εργασίας. Η εφαρμογή βασίζεται στην μεθοδολογία που περιγράψαμε στο κεφάλαιο 6. Συγκεκριμένα, αφορά την υλοποίηση ενός συστήματος clustering σε συνδυασμό με αρχές ασαφούς λογικής.

7.2 ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ

Ο σκοπός της εφαρμογής είναι η υλοποίηση μίας clustering διαδικασίας η οποία θα υποστηρίζει την ασάφεια. Στοχεύει δηλαδή στην εξαγωγή συνόλων ομοίων δεδομένων από μεγάλα σύνολα δεδομένων, τα οποία θα επικαλύπτονται (κάθε στοιχείο θα μπορεί να συμμετέχει σε περισσότερα από ένα clusters).

Στην προσπάθεια μας να προσεγγίσουμε τον παραπάνω στόχο προχωρήσαμε στην υλοποίηση ενός συστήματος clustering για έναν συγκεκριμένο αλγόριθμο. Η υλοποίηση έγινε σε Java ενώ η επικοινωνία με το σύνολο των δεδομένων (βάση δεδομένων) γίνεται μέσω ODBC. Γενικά, τα βασικά στοιχεία που αποτελούν την εφαρμογή είναι:

- ο K-Means αλγόριθμος για πολυδιάστατα δεδομένα,
- ένα μέτρο αξιολόγησης του clustering, για την επιλογή του clustering σχήματος που προκύπτει από την εφαρμογή του K-Means για διαφορετικές τιμές του αριθμού των clusters,
- οι hypertrapezoidal membership functions για τον υπολογισμό των βαθμών συμμετοχής των στοιχείων του συνόλου δεδομένων μας στα clusters που προέκυψαν από την εφαρμογή του K-Means αλγορίθμου.

Η θεωρητική περιγραφή των βασικών στοιχείων που υλοποιήθηκαν για την εφαρμογή μας έγινε στο Κεφάλαιο 6. Στο κεφάλαιο 7 θα επικεντρωθούμε στον τρόπο με τον οποίο έγινε η υλοποίηση.

7.3 ΥΛΟΠΟΙΗΣΗ

Στην παράγραφο αυτή περιγράφονται οι βασικές δομές που υλοποιήθηκαν για την εφαρμογή μας.

7.3.1 CLUSTERING

Το τμήμα της εφαρμογής που αφορά στο clustering εφαρμόζει τον K-Means αλγόριθμο για πολυδιάστατα δεδομένα. Για παράδειγμα, μπορούμε να εφαρμόσουμε clustering σε σύνολα δεδομένων με στοιχεία που αποτελούνται ένα γνώρισμα (π.χ. *τιμή μετοχής*) ή από τα δύο γνωρίσματα (π.χ. *τιμή κλεισίματος μετοχής, υψηλότερη τιμή μετοχής*) ή από τρία γνωρίσματα (π.χ. *τιμή κλεισίματος, υψηλότερη τιμή, όγκος συναλλαγών*) κ.ο.κ.

Στην διαδικασία του clustering συμπεριλαμβάνεται και το *scaling*. Αυτό σημαίνει ότι οι μεταβλητές που συνθέτουν τα πολυδιάστατα δεδομένα κανονικοποιούνται ώστε να συμμετέχουν ισότιμα στην διαδικασία του clustering. Στην εφαρμογή μας επιλέχθηκε το *scaling* να γίνεται με διαίρεση των δεδομένων κάθε διάστασης με τον μέσο όρο των τιμών που λαμβάνει.

Οι κλάσεις που υλοποιήθηκαν για το clustering παρουσιάζονται στην συνέχεια, Πίνακας 7.1, Πίνακας 7.2, Πίνακας 7.3.

Κλάση:

scale_avg

Περιγραφή:

Η κλάση αυτή υλοποιεί τις διαδικασίες για scaling των στοιχείων ενός συνόλου δεδομένων.

Μεταβλητές

Vector avg_x = new Vector()

Vector με τις μέσες τιμές κάθε διάστασης των στοιχείων του συνόλου των δεδομένων. Δηλαδή το i στοιχείο του avg_x αντιστοιχεί στην μέση τιμή της I διάστασης του συνόλου δεδομένων ($X=\{x : x = (x_1, \dots, x_i, \dots, x_n)\}$)

int d

Διάσταση στοιχείων του συνόλου δεδομένων

Μέθοδοι

scale_avg(Connection Ex1Con, String attr[], String table, int dim)

Constructor της κλάσης scale

Παράμετροι

Ex1Con: Αντικείμενο Connection για την ΒΔ που χρησιμοποιείται.

attr: Γνωρίσματα που αποτελούν τα δεδομένα που θα χρησιμοποιήσουμε στο clustering.

name_table: όνομα πίνακα από τον οποίο θα ανακτηθούν δεδομένα για το clustering

dim: διάσταση δεδομένων

public double scaling(double stoixeio, int dim)	Scalling της τιμής stoixeio της dim διάστασης(γνωρίσματος)
---	--

public double unscaling(double stoixeio, int dim)	Unscaling της τιμής stoixeio της dim διάστασης (γνωρίσματος)
---	--

Πίνακας 7.1 Κλάση για την υλοποίηση του scaling

Κλάση:

Περιγραφή:

Euclidean_dist

Η κλάση αυτή υλοποιεί τις διαδικασίες για τον υπολογισμό τις Ευκλείδειας απόστασης. Χρησιμοποιείται τόσο στην εφαρμογή του αλγορίθμου K-Means (class Km_dob) όσο και στον υπολογισμό των βαθμού συμμετοχής (class HMF).

Mέθοδοι

Euclidean_dist()

Constructor της κλάσης *Euclidean_dist*

public double distance(Vector x, Vector m_i)

Υπολογισμός της Ευκλείδεια απόσταση μεταξύ των d-διάστατων στοιχείων x και m_i.

Παράμετροι

x, m_i: d-διάστατα στοιχεία από το σύνολο δεδομένων για το οποία θέλουμε να μετρήσουμε την απόσταση.

Πίνακας 7.2 Κλάση για την υλοποίηση της Ευκλείδειας αποστάσεως

Κλάση:

Περιγραφή:

Km_dob

Η κλάση αυτή υλοποιεί τον αλγόριθμο K-means για πολυδιάστατα δεδομένα, εφαρμόζοντας παράλληλα και scaling στα δεδομένα.

Μεταβλητές

public int c :

public int d:

private Vector cluster =
new Vector();

private Vector m :

private Vector mant :

private Connection Ex1Con

αριθμός στοιχείων

διάσταση δεδομένων στα οποία εφαρμόζουμε clustering

Vector του οποίου το στοιχείο i αντιστοιχεί στο αριθμό του cluster στο οποίο ανήκει το i στοιχείο του συνόλου των δεδομένων .

κέντρα των clusters

Περιέχει τα κέντρα των clusters που προέκυψαν στην προηγούμενη τμηματοποίηση του συνόλου δεδομένων.

Αντικείμενο Connection για την ΒΔ που χρησιμοποιείται.

public String str :
scale_avg scale_cl
int min_size;

έκφραση SELECT για την ανάκτηση των στοιχείων από το σύνολο
Κλάση για τον ορισμό των μεθόδων σχετικά με το scaling που θα εφαρμοστεί στα δεδομένα μας.
Ελάχιστος αριθμός στοιχείων κάθε cluster.

Mέθοδοι

Km_dob(int size, Connection Con)

Constructor για την κλάση Km_dob.

Παράμετροι

size: ελάχιστος αριθμός στοιχείων που μπορεί να έχει κάθε cluster.

connection: Αντικείμενο Connection για την ΒΔ που χρησιμοποιείται.

public void init_param(int dim, String name_table, String attr[])

Αρχικοποίηση παραμέτρων για τον K_means αλγόριθμο.

Παράμετροι

dim: διάσταση δεδομένων

name_table: όνομα πίνακα από τον οποίο θα ανακτηθούν δεδομένα για το clustering

attr: Γνωρίσματα που αποτελούν τα δεδομένα που θα χρησιμοποιήσουμε στο clustering.

Mέθοδοι για K-means

void calculate_dist()

Προσδιορίζει το cluster στο οποίο ανήκει κάθε στοιχείο από την ΒΔ που εφαρμόζουμε clustering με βάση την απόσταση του από τα κέντρα των clusters.

boolean calculate_mesos()

Υπολογίζει τα κέντρα των clusters. Το κέντρο σε κάθε επανάληψη του αλγορίθμου είναι η μέση τιμή των στοιχείων που ανήκουν στο συγκεκριμένο cluster.

double calculate_error()

Προσδιορίζει το κριτήριο σταματήματος. Υπολογίζει την απόκλιση μεταξύ των κέντρων των clusters, τα οποία προκύπτουν από την εφαρμογή δύο διαδοχικών εκτιμήσεων των clusters (επαναλήψεων του K-means).

Γενικοί Μέθοδοι

String printmesous()

Επιστρέφει ένα String με τα κέντρα των clusters που προέκυψαν από τον αλγόριθμο K-Means ώστε να εμφανιστεί η περιγραφή των clusters στο χρήστη.

public boolean exec_alg(int num_cluster)	Εκτελεί τον αλγόριθμο K-Means για num_cluster με βάση τις μεθόδους που περιγράφηκαν παραπάνω.
public Vector get_centers()	Επιστρέφει τα κέντρα των clusters αφού γίνει unscaling στις τιμές που αντιπροσωπεύουν τα κέντρα.
public Vector getscl_clusters()	Επιστρέφει τα clusters στα οποία ανήκει κάθε στοιχείο που συμμετέχει στο clustering. Πρόκειται για τα crisp clusters τα οποία προκύπτουν από τον K-Means.

Πίνακας 7.3 Κλάση υλοποίησης αλγορίθμου K-Means

7.3.2 ΑΞΙΟΛΟΓΗΣΗ CLUSTERING (CLUSTERING VALIDATION)

Το κομμάτι αυτό της εφαρμογής όπως το περιγράψαμε στην παράγραφο 6.3.3 αποσκοπεί στην εύρεση του καλύτερου αριθμού clusters για το σύνολο δεδομένων μας (εύρεση καλύτερου σχήματος clustering). Ως κριτήριο για την αξιολόγηση χρησιμοποιήθηκε το μέτρο που περιγράψαμε στην 6.3.3.1.

Ειδικότερα, στην εφαρμογή μας για την πραγματοποίηση της αξιολόγησης του clustering υλοποιήθηκε η κλάση Validation (Πίνακας 7.4).

Κλάση:	Validation
Περιγραφή:	Η κλάση αυτή προσδιορίζει τον δείκτη αξιολόγησης για ένα clusters σχήμα που προέκυψε από την εφαρμογή ενός αλγορίθμου clustering(για την εφαρμογή μας K-Means). Ο δείκτης αξιολόγησης είναι το μέτρο της παρ. 6.3.3.1

Μεταβλητές

private Vector var_c = new Vector()	Vector με την διασπορά κάθε cluster του σχήματος clustering για το οποίο θέλουμε να υπολογίσουμε το δείκτη ποιότητας.
private Vector m	κέντρα clusters
private int d	διαστάσεις των στοιχείων του συνόλου δεδομένων
private Connection Con	Αντικείμενο σύνδεσης στη ΒΔ
private String str	έκφραση SELECT για την ανάκτηση των στοιχείων του συνόλου δεδομένων που εφαρμόστηκε το clustering
private Vector cluster	Vector με το cluster στο οποίο ανήκει κάθε στοιχείο της Β.Δ.

Mέθοδοι

Validation(Connection connection)	Constructor για την κλάση Validation. Παράμετροι Connection: Αντικείμενο Connection για την ΒΔ που χρησιμοποιείται.
public double init_param(Vector mesoi, String s)	Υπολογίζει και επιστρέφει την τιμή της παραμέτρου εξομάλυνσης α. α = Disc(cmax), όπου cmax είναι ένας μεγάλος αριθμός clusters. Η παράμετρος δηλαδή αντιστοιχεί στην απόσταση των κέντρων των clusters για την χειρότερη περίπτωση . Παράμετροι
	mesoi: κέντρα των clusters του χειρότερου σχήματος clustering. s: Select string για την ανάκτηση στοιχείων από την ΒΔ τα οποία θα χρησιμοποιηθούν για clustering.
private void var_cluster(int i)	Υπολογίζει την διασπορά για το cluster i. Ορίζει έναν Vector του οποίου το στοιχείο j αντιστοιχεί στην j διάσταση της διασποράς για το i cluster.
private double Scatt(c)	Εκτίμηση της μέσης διασποράς στα clusters του τρέχοντος σχήματος
double Disc(c)	Εκτίμηση της διαφοροποίησης της απόστασης μεταξύ των κέντρων των clusters
public double validity(double a, Vector cl_ind, Vector mesoi)	Προσδιορίζει δείκτη αξιολόγησης (validation index) για το τρέχον σχήμα <i>Validation Index = $\alpha^* Scatt(c) + Disc(c)$</i> Παράμετροι
	a : παράμετρος εξομάλυνσης. cl_ind: Vector του οποίου το i στοιχείο αντιστοιχεί στο cluster που ανήκει το i στοιχείο της ΒΔ. [Τα clusters προκύπτουν από τον K-Means(crisp clusters)]. mesoi : κέντρα των clusters για τα οποία θα υπολογιστή ο δείκτης αξιολόγησης

Πίνακας 7.4. Κλάση για την υλοποίηση των μέτρων ποιότητας

7.3.3 ΣΥΝΑΡΤΗΣΕΙΣ ΣΥΜΜΕΤΟΧΗΣ

Τα clusters που προκύπτουν από την εφαρμογή του K-means είναι crisp cluster(δεν επικαλύπτονται). Προκειμένου να εισάγουμε την ασάφεια στο σχήμα clustering που προέκυψε από την εφαρμογή του αλγορίθμου προσδιορίζουμε έναν παράγοντα επικάλυψης, ο οποίος δηλώνει την επικάλυψη μεταξύ γειτονικών clusters και υιοθετούμε την λογική των hypertrapezoidal functions (παρ. 6.4.4) για να υπολογιστούν οι βαθμοί συμμετοχής των στοιχείων της ΒΔ στα clusters.

Οι συναρτήσεις συμμετοχής υλοποιούνται στην εφαρμογή μας με την κλάση HMF (Πίνακας 7.5).

Κλάση:	HMF
Περιγραφή:	Η κλάση αυτή υλοποιεί την διαδικασία υπολογισμού των hypertrapezoidal membership functions.
Παράμετροι	
public int c	αριθμός clusters
private double s	crispness factor
public int d	διάσταση δεδομένων στα οποία εφαρμόζουμε clustering
private Vector m	κέντρα των clusters
public String str_attr	έκφραση SELECT για την ανάκτηση των στοιχείων του συνόλου δεδομένων που εφαρμόστηκε το clustering
Μέθοδοι	
HMF(int num_cluster, double crisp, Vector centers, String st_select)	Constructor για την κλάση HMF. Η κλάση ορίζει τους βαθμούς συμμετοχής
Παράμετροι	
num_clusters:	αριθμός clusters
crisp:	crispness factor
centers :	κέντρα των clusters για τα οποία θα υπολογιστή ο δείκτης αξιολόγησης.
st_select:	έκφραση SELECT για την ανάκτηση των στοιχείων για τα οποία θα υπολογιστούν οι βαθμοί συμμετοχής.
private Vector mi_jcalculate(Vector x)	Υπολογισμός των υποσυνθήκη βαθμών συμμετοχής του n-διάστατου στοιχείου x για τα clusters i, j
public Vector calculate_dob(Vector x)	Υπολογισμός των βαθμών συμμετοχής για το n-διάστατο στοιχείο x hypertrapezoidal memberships functions

Πίνακας 7.5 Κλάση υλοποίησης των Hypertrapezoidal Membership Functions

7.3.4 INTERFACE CLASSES

Εκτός από τις παραπάνω κλάσεις, οι οποίες υλοποιούν και τα βασικά στοιχεία του συστήματος fuzzy clustering, αναπτύχθηκε και μία ομάδα κλάσεων που υλοποιούν το interface της εφαρμογής. Οι κλάσεις αυτές είναι:

- *class Connection_Dlg.* Υλοποιεί το παράθυρο διαλόγου για σύνδεση μέσω ODBC στην βάση δεδομένων που θα χρησιμοποιηθεί για clustering.
- *class table_dlg.* Υλοποιεί το παράθυρο διαλόγου για την επιλογή του πίνακα από τον οποίο θα γίνει η ανάκτηση των δεδομένων στα οποία θα εφαρμοστεί clustering.
- *class Cluster_Dlg.* Υλοποιεί το βασικό παράθυρο διαλόγου για την εφαρμογή μας. Στο παράθυρο αυτό καθορίζονται οι βασικοί παράμετροι για την υλοποίηση του clustering, την επιλογή του καλύτερου σχήματος clustering και τον καθορισμό του crispness factor για τις συναρτήσεις συμμετοχής.

Στην κλάση αυτή ορίζονται και οι μέθοδοι εκτέλεσης του clustering με βάση τις παραμέτρους που ορίστηκαν σε αυτό. Οι μέθοδοι αυτοί είναι:

```
public void exec_clust(int
c_max, int c_min, double
f_crisp, String table)
```

Εκτελεί τον αλγόριθμο εύρεσης του καλύτερου clustering σχήματος.
 c_max: μέγιστος αριθμός clusters
 c_min: ελάχιστος αριθμός clusters
 f_crisp: crispness factor
 table: πίνακας από τον οποίο γίνεται η ανάκτηση των δεδομένων για clustering.

```
int min_clsize()
```

Προσδιορίζει τον ελάχιστο αριθμό στοιχείων που μπορεί να έχει κάθε cluster.

```
public void actionPerformed
(ActionEvent event)
```

Ελέγχει τα συμβάντα του παραθύρου, και εκτελεί τις κατάλληλες ενέργειες ανάλογα με τον συμβάν.

```
public void exec_clust(int
c_max, int c_min, double
f_crisp, String table)
```

Εφαρμογή του αλγορίθμου εύρεσης του καλύτερου σχήματος crisp clustering σύμφωνα με την παρ. 6.3.3.2. Ειδικότερα, εκτελεί τον αλγόριθμο K-Means για όλες τις τιμές του αριθμού των clusters που ορίζονται από τον χρήστη και προσδιορίζει το καλύτερο σχήμα clustering. Επίσης ορίζει την κλάση HMF για το βέλτιστο σχήμα clustering που προκύπτει.

```
Public HMF get_cluster()
```

Επιστρέφει την κλάση HMF για το βέλτιστο σχήμα clustering.

```
public void paint(String str)
```

Εισάγει τα αποτελέσματα που προκύπτουν κάθε φορά από την εκτέλεση του αλγορίθμου clustering.

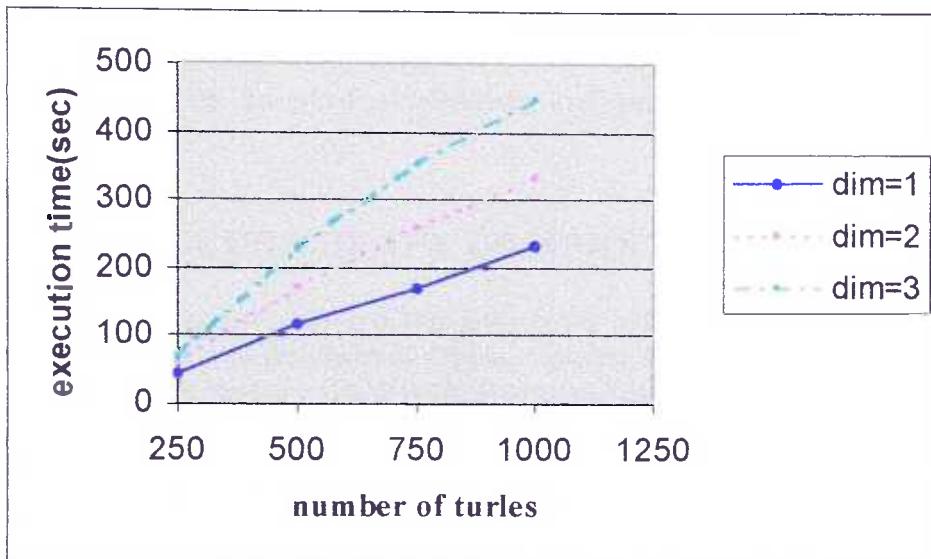
7.4 ΠΟΛΥΠΛΟΚΟΤΗΤΑ ΕΦΑΡΜΟΓΗΣ

Με βάση τον αλγόριθμο που περιγράψαμε στην παρ. 6.3.3.2, ο οποίος αποτελεί και τον βασικό αλγόριθμο για την εφαρμογή μας μπορούμε να υπολογίσουμε την πολυπλοκότητα για την εφαρμογή μας.

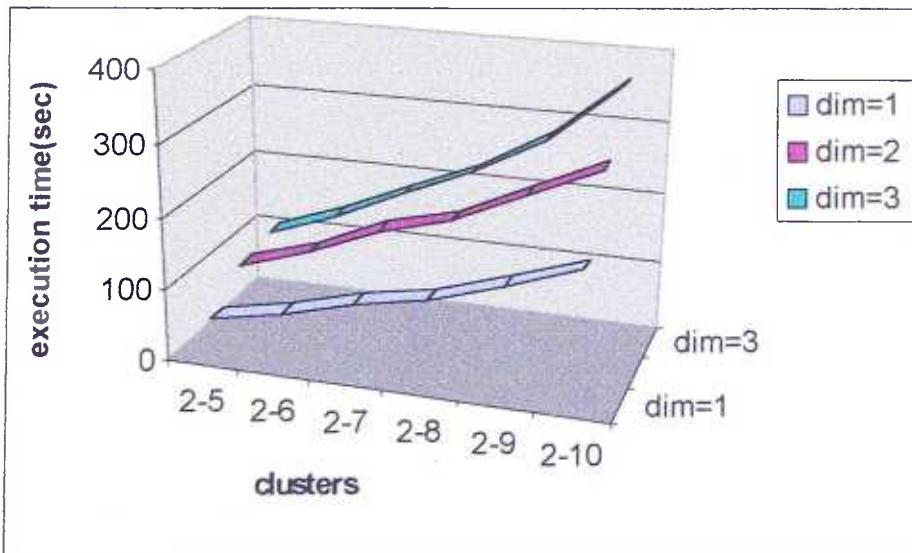
Ο K-means αλγόριθμος έχει πολυπλοκότητα $O(tndc)$, όπου d είναι ο αριθμός των γνωρισμάτων, c είναι ο αριθμός των clusters, n είναι ο αριθμός των δεδομένων και t είναι ο αριθμός των επαναλήψεων σε όλο το σύνολο των δεδομένων. Συνήθως $d, c, t \ll n$. Για τον υπολογισμό του μέτρου αξιολόγησης η πολυπλοκότητα είναι $O(ndc + c^2d)$, όπου τα n, c, d ορίζονται όπως παραπάνω. Η εκτέλεση του αλγορίθμου K-Means και ο υπολογισμός του μέτρου αξιολόγησης επαναλαμβάνει για κάθε τιμή του αριθμού των clusters, $c \in [c_{\min}, c_{\max}]$. Συνεπώς, η συνολική πολυπλοκότητα για τον υπολογισμό του καλύτερου clustering σχήματος για το σύνολο των δεδομένων μας είναι $O((c_{\max}-c_{\min}+1)*(tncd + c^2d))$.

Η πολυπλοκότητα για τον υπολογισμό των βαθμών συμμετοχής ενός στοιχείου από το σύνολο των δεδομένων στα clusters που προέκυψαν από την διαδικασία του clustering, είναι $O(c^2d)$. Συνεπώς η συνολική πολυπλοκότητα για τον υπολογισμό των βαθμών συμμετοχής n στοιχείων θα είναι $O(nc^2d)$.

Το σύστημα μας εφαρμόστηκε σε σύνολα δεδομένων με 250-1000 εγγραφές και για στοιχεία με διαστάσεις 1-3. Στο Διάγραμμα 7.1 παρουσιάζεται η μεταβολή του χρόνου εκτέλεσης σε σχέση με την μεταβολή του αριθμού των δεδομένων, όταν ο αριθμός των clusters κυμαίνεται στο διάστημα [2, 10]. Από το διάγραμμα αυτό παρατηρούμε ότι ο χρόνος εκτέλεσης εξαρτάται από το αριθμό και τις διαστάσεις των στοιχείων στα οποία εφαρμόζεται το clustering. Η αύξηση του χρόνου εκτέλεσης μπορούμε να πούμε ότι είναι ανάλογη περίπου του αριθμού των στοιχείων. Επίσης στο Διάγραμμα 7.2 παρουσιάζεται η μεταβολή του χρόνου εκτέλεσης σε σχέση με την μεταβολή των ορίων στα οποία κυμαίνεται ο αριθμός των clusters κατά την διαδικασία αναζήτησης του βέλτιστου σχήματος clustering.



Διάγραμμα 7.1. Μεταβολή του χρόνου εκτέλεσης σε σχέση με τον αριθμό των στοιχείων του συνόλου δεδομένων.



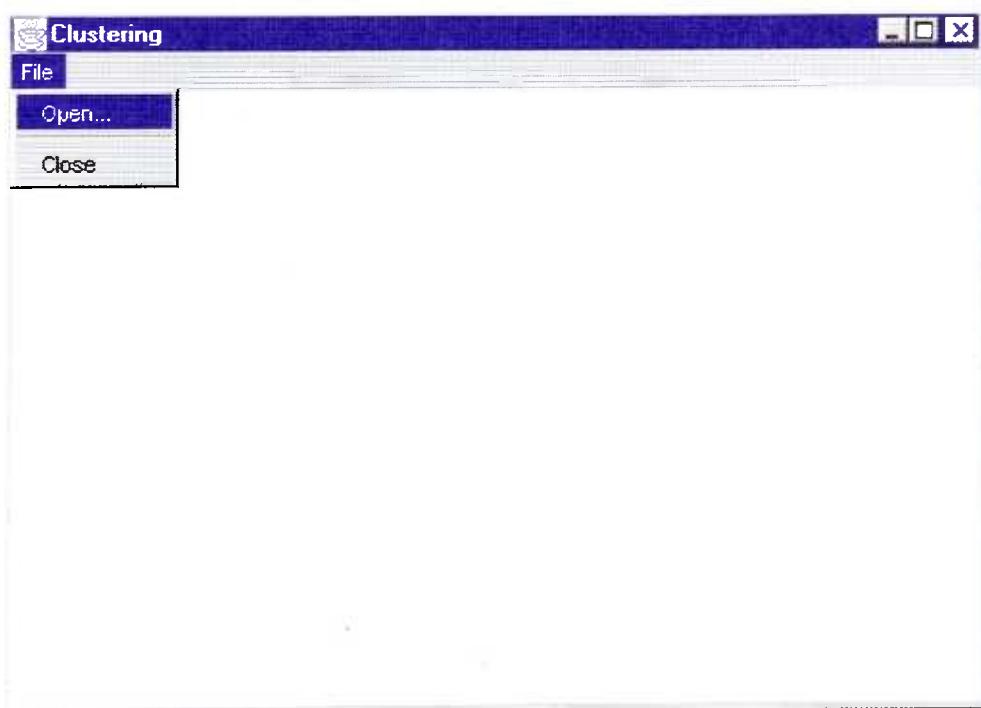
Διάγραμμα 7.2. Μεταβολή του χρόνου εκτέλεσης σε σχέση με τα όρια στα οποία κυμαίνεται ο αριθμός των clusters στην διαδικασία αναζήτησης βέλτιστου σχήματος clustering.

7.5 ΠΑΡΟΥΣΙΑΣΗ ΕΦΑΡΜΟΓΗΣ

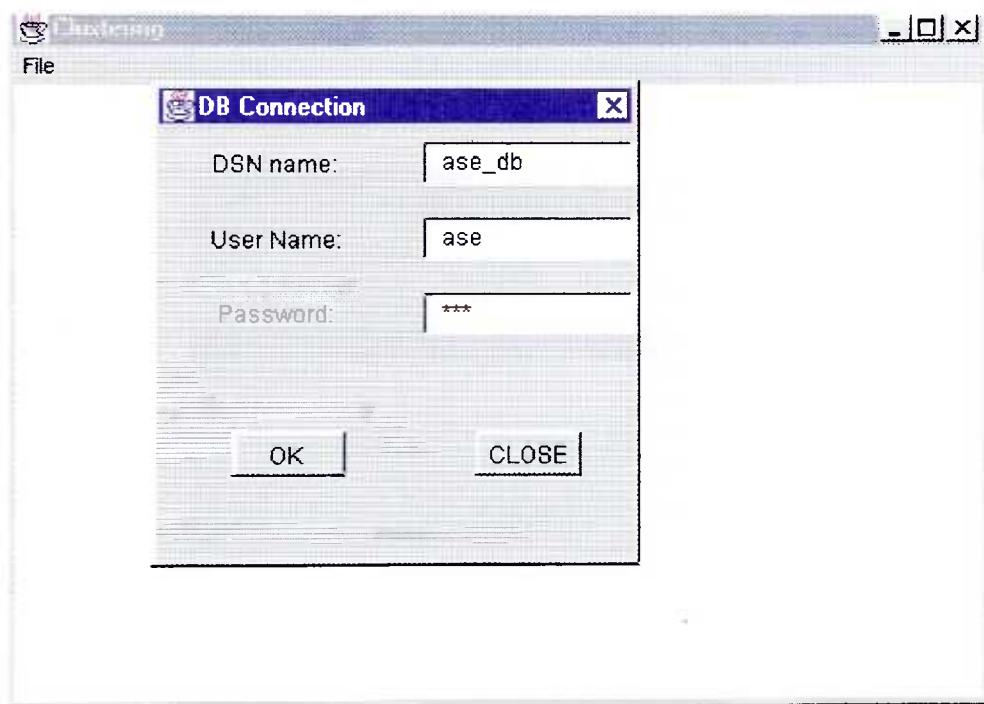
Στην παράγραφο αυτή θα γίνει παρουσίαση της εφαρμογής και του τρόπου λειτουργία της.

7.5.1 ΟΘΟΝΗ ΕΚΚΙΝΗΣΗΣ ΤΗΣ ΕΦΑΡΜΟΓΗΣ

Στην εκκίνηση της εφαρμογής εμφανίζεται ένα κενό παράθυρο (Εικόνα 1). Στο μενού File του παραθύρου αυτού επιλέγοντας Open... ανοίγει ένα παράθυρο διαλόγου για σύνδεση στην βάση δεδομένων (Εικόνα 2). Συγκεκριμένα ζητάει το όνομα σύνδεσης με ODBC, το όνομα και τον κωδικό του χρήστη που πρόκειται να συνδεθεί στην συγκεκριμένη βάση δεδομένων. Το όνομα και ο κωδικός καθορίζονται από το ίδιο το σύστημα διαχείρισης της βάσης δεδομένων και μπορούν να αφεθούν κενά εάν επιτρέπεται ελεύθερη πρόσβαση.



Εικόνα 1. Οθόνη Εκκίνησης



Εικόνα 2. Οθόνη Σύνδεσης

7.5.2 ΟΘΟΝΗ ΕΠΙΛΟΓΗΣ ΠΙΝΑΚΑ

Μετά την σύνδεση στην βάση δεδομένων μέσω του παραπάνω παραθύρου διαλόγου εμφανίζεται στην οθόνη μας μία λίστα με τους πίνακες που υπάρχουν στην συγκεκριμένη βάση δεδομένων (Εικόνα 3). Ο χρήστης επιλέγει από την λίστα τον πίνακα που θα ήθελε να χρησιμοποιήσει στην διαδικασία του clustering. Αφού γίνει η επιλογή του πίνακα ο χρήστης επιλέγοντας "OK" προχωράει στην βασική οθόνη του clustering.



Εικόνα 3. Οθόνη Επιλογής Πίνακα

7.5.3 ΚΕΝΤΡΙΚΗ ΟΘΟΝΗ CLUSTERING

Η οθόνη αυτή αποτελεί την κύρια οθόνη της εφαρμογής μας (Εικόνα 4). Μέσω του παραθύρου αυτού καθορίζονται οι παράμετροι για την εκτέλεση του clustering.

Ένα από τα κύρια στοιχεία που θα πρέπει να προσδιορίσει ο χρήστης είναι η μορφή των δεδομένων που θα χρησιμοποιηθούν στο clustering. Με βάση τον πίνακα που επιλέχθηκε στο προηγούμενο παράθυρο διαλόγου εμφανίζεται μία λίστα με τα γνωρίσματα του πίνακα. Ο χρήστης καλείται να επιλέξει ένα ή περισσότερα από τα γνωρίσματα της λίστας. Τα γνωρίσματα καθορίζουν τις διαστάσεις των δεδομένων που θα χρησιμοποιηθούν στο clustering.

Άλλα στοιχεία τα οποία καλείται να καθορίσει ο χρήστης είναι:

- **μέγιστη και η ελάχιστη τιμή για τον αριθμό των clusters.** Οι τιμές αυτές καθορίζουν το διάστημα $[c_{\min}, c_{\max}]$ στο οποίο θα αναζητηθεί ο βέλτιστος αριθμός των clusters για το σύνολο των δεδομένων μας.

Παρατήρηση: Ο χρήστης μπορεί να επιλέξει ώστε η μέγιστη και ελάχιστη τιμή να ταυτίζονται. Στην περίπτωση αυτή ως βέλτιστος αριθμός λαμβάνεται η τιμή $c = (c_{\min} = c_{\max})$.

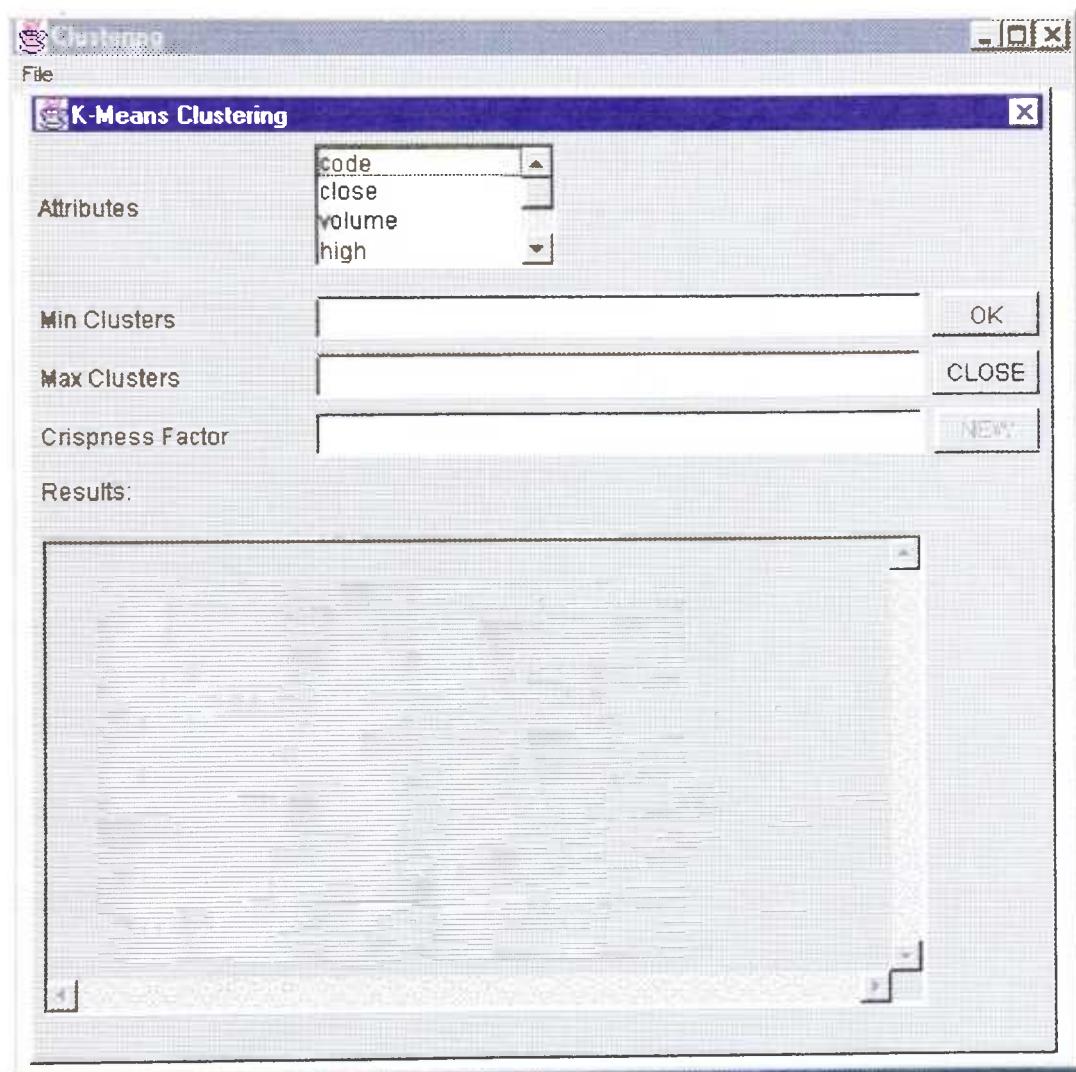
- **Ο παράγοντας επικάλυψης(crispness factor), $\sigma \in [0, 1]$.** Με βάση τον παράγοντα αυτό και τα κέντρα των clusters που θα προκύψουν από την εφαρμογή του K-means αλγορίθμου για τον βέλτιστο αριθμό clusters, θα προσδιοριστούν οι συναρτήσεις συμμετοχής.

Στο δεξί μέρος του παραθύρου εμφανίζονται δύο κουμπιά:

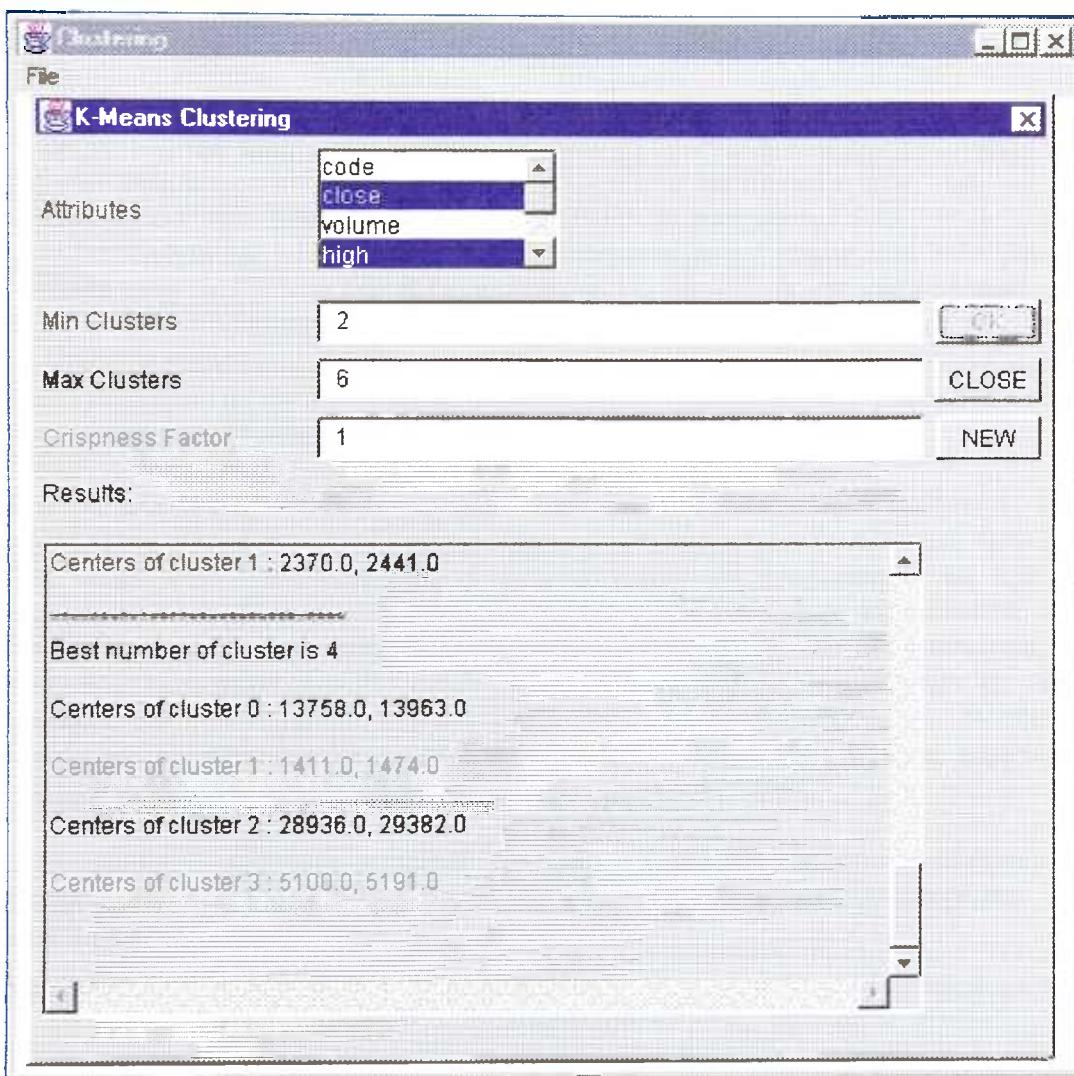
- ◆ **OK:** Ο χρήστης έχοντας καθορίσει τις παραπάνω παραμέτρους μπορεί να επιλέξει το κουμπί OK για να προχωρήσει στην εκτέλεση της διαδικασίας clustering. Τα δεδομένα στα οποία εφαρμόζεται το clustering ανακτώνται από την ΒΔ που έχει επιλεχθεί κατά την διαδικασία της σύνδεσης και αποτελούνται από τα γνωρίσματα που επέλεξε ο χρήστης. Το σύστημα εφαρμόζει τον K_means αλγόριθμο για όλες τις τιμές του αριθμού των clusters, μεταξύ του c_{\min} και c_{\max} και επιλέγει τον βέλτιστο αριθμό χρησιμοποιώντας ως κριτήριο το δείκτη αξιολόγησης που αναφέραμε σε προηγούμενα κεφάλαια.

Τα αποτελέσματα της εκτέλεσης του αλγορίθμου για κάθε τιμή $c \in [c_{\min}, c_{\max}]$ καθώς και τα αποτελέσματα για το καλύτερο αριθμό clusters εμφανίζονται στην περιοχή κειμένου(TextArea) που υπάρχει στο κάτω μέρος του παραθύρου (Εικόνα 5).

- ◆ **CLOSE:** Εάν ο χρήστης επιλέξει CLOSE τότε κλείνει το παράθυρο και ακυρώνεται η διαδικασία clustering.
- ◆ **NEW:** Η επιλογή NEW δίνει την δυνατότητα επανορισμού των παραμέτρων για την εκτέλεση του clustering.



Εικόνα 4. Κεντρική Οθόνη Clustering



Εικόνα 5. Οθόνη Αποτελεσμάτων Clustering

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η διαδικασία της εξόρυξης γνώσης (data mining) έχει βασικό σκοπό την αναζήτηση και εξαγωγή προτύπων γνώσης που παρουσιάζουν ενδιαφέρον μέσα από μεγάλα σύνολα δεδομένων. Είναι κοινά όμως αποδεκτό ότι τα πρότυπα αυτά θα πρέπει να είναι κατανοητά και εύκολα ερμηνεύσιμα από μη ειδικούς. Επίσης ένα άλλο βασικό στοιχείο στην διαδικασία του data mining είναι η διαχείριση της αβεβαιότητας. Τα δεδομένα στην πραγματικότητα εμπεριέχουν στοιχεία ασάφειας τα οποία θα πρέπει να ληφθούν υπόψη στις διάφορες εργασίες του data mining (clustering, classification, association rules extraction) προκειμένου να οδηγηθούμε σε μορφές γνώσεις χρήσιμες για την εξαγωγή συμπερασμάτων και την λήψη αποφάσεων.

Στα πλαίσια της εργασίας αυτής παρουσιάστηκε μία μεθοδολογία ανάπτυξης ενός συστήματος clustering το οποίο υποστηρίζει την αβεβαιότητα. Το σύστημα αυτό έχει σαν στόχο την εξαγωγή/ορισμό clusters τα οποία θα αποτελέσουν τις αρχικές κατηγορίες. Με βάση τις κατηγορίες αυτές μπορεί να γίνει classification των στοιχείων ενός συνόλου δεδομένων και στην συνέχεια με κατάλληλες διαδικασίες να προχωρήσουμε στην εξαγωγή κανόνων και άλλων μορφών γνώσης που θα είναι χρήσιμες για την εξαγωγή συμπερασμάτων και την λήψη αποφάσεων. Βλέπουμε λοιπόν ότι το σύστημα fuzzy clustering αποτελεί βασικό κομμάτι ενός ευρύτερου συστήματος data mining το οποίο υποστηρίζει την αβεβαιότητα.

Η προτεινόμενη μεθοδολογία συνδυάζει στοιχεία κλασικών τεχνικών clustering με στοιχεία ασαφούς λογικής. Γενικά, τα βασικά στοιχεία που την χαρακτηρίζουν της συνοψίζονται στα εξής:

- *Μπορεί να εφαρμοστεί για οποιονδήποτε αλγόριθμο clustering.* Ο αλγόριθμος που θα χρησιμοποιηθεί για την διαδικασία μπορεί να είναι είτε ένας αλγόριθμος crisp clustering, ο οποίος οδηγεί σε crisp clusters, είτε ένας αλγόριθμος ο οποίος λαμβάνει εξ ορισμού υπόψη του την ασάφεια (fuzzy clustering αλγόριθμος).
- *Υιοθετεί κριτήρια ποιότητας για την αξιολόγηση των clusters.* Σύμφωνα με την μεθοδολογία, κάθε αλγόριθμος clustering που θα επιλεχθεί εφαρμοζόμενος κάτω από διαφορετικές υποθέσεις μπορεί να οδηγήσει σε διαφορετικά σχήματα clustering. Προκειμένου όμως να επιλεχθεί το καλύτερο σχήμα (δηλ. το σχήμα με τα πιο πυκνά και καλά διαχωρισμένα clusters), γίνεται αξιολόγηση αυτών με βάση κάποιο μέτρο ποιότητας.
- *Καθορίζει συναρτήσεις συμμετοχής για τα clusters.* Σύμφωνα με την μεθοδολογία ορίζονται κάποιες συναρτήσεις συμμετοχής για τα clusters που προκύπτουν με την εφαρμογή του clustering στο σύνολο δεδομένων. Με βάση τις συναρτήσεις αυτές κατά την διαδικασία του classification παράγονται οι βαθμοί πίστης των στοιχείων στα clusters.

Με βάση την μεθοδολογία έγινε μία πρώτη προσπάθεια ανάπτυξης ενός συστήματος clustering που θα διαχειρίζεται την αβεβαιότητα κατά την διαδικασία του data mining σε σύνολα αριθμητικών δεδομένων. Τα βασικά χαρακτηριστικά και οι

υποθέσεις που έγιναν για το συστήμα που υλοποιήθηκε στα πλαίσια της εργασίας μπορούν να συνοψιστούν στα εξής:

- Το σύστημα clustering υλοποιήθηκε για αριθμητικά δεδομένα μόνο. Συνεπώς δεδομένα με λεκτικές τιμές ή/και δεδομένα που περιέχουν λεκτικές και αριθμητικές τιμές δεν υποστηρίζονται.
- Στην παρούσα υλοποίηση η αρχικοποίηση της διαδικασίας clustering γίνεται με τυχαία επιλογή σημείων. Ο αλγόριθμος clustering (K-Means) λαμβάνει ως αρχικά κέντρα των clusters που προσπαθεί να προσδιορίσει οποιοδήποτε στοιχεία από το σύνολο των δεδομένων. Με την υιοθέτηση όμως κάποιας μεθοδολογίας η οποία θα οδηγούσε σε επιλογή σημείων που θα ήταν αντιτροσωπευτικά των κατηγοριών που θέλουμε να παράγουμε θα μπορούσαμε να συγκλίνουμε γρηγορότερα σε καλύτερα σχήματα clustering.
- Η εφαρμογή του συστήματος clustering σε πολύ μεγάλα σύνολα δεδομένων καθίσταται απαγορευτική λόγω του μεγάλου κόστους. Οι αλγόριθμοι clustering απαιτούν πολλαπλές προσπελάσεις στα δεδομένα μέχρι να καταλήξουν στο τελικό σχήμα clustering. Για μεγάλα όμως σύνολα δεδομένων, οι πολλαπλές προσπελάσεις καθίστανται απαγορευτικά ακριβές. Μέθοδοι οι οποίοι θα βοηθήσουν στην τμηματοποίηση(partitioning) του συνόλου των δεδομένων και την επιλογή αντιτροσωπευτικών δειγμάτων (sampling) θα μπορούσαν να βοηθήσουν στην αντιμετώπιση του προβλήματος εφαρμογής clustering σε μεγάλες βάσεις δεδομένων.

Μερικά θέματα σχετικά με το σύστημα clustering, τα οποία θα μπορούσαν να αποτελέσουν αντικείμενο μελλοντικής μελέτης είναι:

- Υλοποίηση fuzzy clustering αλγορίθμων ώστε να έχουμε μία ολοκληρωμένη εικόνα για το clustering. Επίσης με σύγκριση των αποτελεσμάτων που προκύπτουν από την εφαρμογή κλασικών και fuzzy αλγορίθμων μπορούμε να οδηγηθούμε σε χρήσιμα συμπεράσματα σε ότι αφορά την επίδραση του fuzzy clustering στα δεδομένα.
- Σύνδεση της διαδικασίας clustering με το classification σύστημα[Vaz98] το υποστηρίζει την αβεβαιότητα. Έτσι, η έξοδος του προτεινόμενου σχήματος clustering (δηλαδή clusters, membership functions) θα χρησιμοποιείται για την παραγωγή των βαθμών πίστης στην διαδικασία του classification (classification beliefs). Ο βασικός σκοπός αυτής της σύνδεσης είναι η δημιουργία ενός data mining συστήματος το οποίο θα διαχειρίζεται την αβεβαιότητα.
- Ένα επίσης σημαντικό στοιχείο είναι η επέκταση της προσέγγισης μας για την υποστήριξη incremental clustering. Οι βάσεις δεδομένων συνήθως έχουν συχνές ενημερώσεις και έτσι τα πρότυπα που εξάγουμε από τα σύνολα δεδομένων μέσω των data mining μεθόδων θα πρέπει επίσης να ενημερώνονται.

ΠΑΡΑΡΤΗΜΑ Α'

ΠΑΡΟΥΣΙΑΣΗ ΜΟΡΦΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΣΥΣΤΗΜΑΤΟΣ FUZZY CLUSTERING.

Το σύστημα fuzzy clustering που αναπτύχθηκε στα πλαίσια της εργασίας εφαρμόστηκε πειραματικά σε τυχαία σύνολα δεδομένων που επιλέχθηκαν από το χώρο του Χρηματιστηρίου. Τα σύνολα αυτά περιέχουν πραγματικά δεδομένα και αφορούν χρηματιστηριακές συναλλαγές που πραγματοποιήθηκαν συγκεκριμένες ημέρες της εβδομάδας. Η δομή ενός τέτοιου συνόλου παρουσιάζεται στο Πίνακα B, και αποτελεί ένα υποσύνολο των συναλλαγών που πραγματοποιήθηκαν στις 12/1/98 έως 15/1/98. Ειδικότερα, τα πεδία του πίνακα της βάσης δεδομένων που χρησιμοποιήθηκαν στο πείραμα μας είναι:

- *code*, το όνομα της μετοχής,
- *date*, η ημέρα στην οποία αναφέρονται οι συναλλαγές,
- *close*, η τιμή κλεισίματος της μετοχής την συγκεκριμένη μέρα,
- *volume*, ο όγκος των συναλλαγών της μετοχής για την συγκεκριμένη μέρα,
- *high*, η υψηλότερη τιμή που πήρε η μετοχή της συγκεκριμένη μέρα,
- *low*, η χαμηλότερη τιμή της μετοχής την συγκεκριμένη μέρα,
- *value*, η αξία των συναλλαγών της μετοχής,
- *trans*, ο αριθμός συναλλαγών της μετοχής για την συγκεκριμένη μέρα.

Με βάση τα παραπάνω πεδία δημιουργήσαμε τα σύνολα δεδομένων στα οποία εφαρμόσαμε clustering με την βοήθεια του συστήματός μας. Στην συνέχεια παρουσιάζονται τα αποτελέσματα από την εφαρμογή clustering σε σύνολα που αποτελούνται από μονοδιάστατα, διδιάστατα και τρισδιάστατα στοιχεία. Επίσης στην παράγραφο 2 παρουσιάζονται ενδεικτικά αποτελέσματα κατανομής των δεδομένων στα clusters που προέκυψαν με βάση της συναρτήσεις συμμετοχής. Σε κάθε περίπτωση το διάστημα στο οποίο αναζητήθηκε ο βέλτιστος αριθμός clustering ήταν [2,10].

1. ΕΠΙΛΟΓΗ ΒΕΛΤΙΣΤΟΥ ΣΧΗΜΑΤΟΣ CLUSTERING

Πείραμα 1. Το σύνολο δεδομένων που επιλέχθηκε για την διαδικασία clustering αποτελείται από μονοδιάστατα στοιχεία με γνώρισμα το *close*.

Παράμετροι

Attributes: close

Min Clusters: 2

Max Clusters: 10

Crispness Factor: 1

Αποτελέσματα

*For 10 clusters, Validation index : 0.015873042807715784

Centers of cluster 0 : 1719.0

Centers of cluster 1 : 3996.0

Centers of cluster 2 : 6594.0

Centers of cluster 3 : 2486.0

Centers of cluster 4 : 588.0

Centers of cluster 5 : 893.0

Centers of cluster 6 : 1252.0

Centers of cluster 7 : 255.0

Centers of cluster 8 : 13176.0

Centers of cluster 9 : 24602.0

*For 9 clusters, Validation index : 0.014179456926330668

Centers of cluster 0 : 2165.0

Centers of cluster 1 : 5599.0

Centers of cluster 2 : 10645.0

Centers of cluster 3 : 3347.0

Centers of cluster 4 : 602.0

Centers of cluster 5 : 961.0

Centers of cluster 6 : 1411.0

Centers of cluster 7 : 261.0

Centers of cluster 8 : 18114.0

*For 8 clusters, Validation index : 0.011466879059987278

Centers of cluster 0 : 3032.0

Centers of cluster 1 : 9830.0

Centers of cluster 2 : 16915.0

Centers of cluster 3 : 5369.0

Centers of cluster 4 : 720.0

Centers of cluster 5 : 1205.0

Centers of cluster 6 : 1852.0

Centers of cluster 7 : 304.0

*For 7 clusters, Validation index : 0.009611253295093432

Centers of cluster 0 : 2878.0

Centers of cluster 1 : 9617.0

Centers of cluster 2 : 16738.0

Centers of cluster 3 : 5331.0
Centers of cluster 4 : 353.0
Centers of cluster 5 : 874.0
Centers of cluster 6 : 1573.0

*For 6 clusters, Validation index : 0.007569358762122419

Centers of cluster 0 : 2684.0
Centers of cluster 1 : 9215.0
Centers of cluster 2 : 16487.0
Centers of cluster 3 : 5124.0
Centers of cluster 4 : 553.0
Centers of cluster 5 : 1379.0

*For 5 clusters, Validation index : 0.007095915455757884

Centers of cluster 0 : 1957.0
Centers of cluster 1 : 8645.0
Centers of cluster 2 : 16406.0
Centers of cluster 3 : 4395.0
Centers of cluster 4 : 746.0

*For 4 clusters, Validation index : 0.006982660153717767

Centers of cluster 0 : 952.0
Centers of cluster 1 : 6278.0
Centers of cluster 2 : 15158.0
Centers of cluster 3 : 2794.0

*For 3 clusters, Validation index : 0.00804179198712987

Centers of cluster 0 : 1333.0
Centers of cluster 1 : 5739.0
Centers of cluster 2 : 15158.0

*For 2 clusters, Validation index : 0.012650319887178352

Centers of cluster 0 : 1862.0
Centers of cluster 1 : 12847.0

Best number of cluster is 4
Centers of cluster 0 : 952.0
Centers of cluster 1 : 6278.0
Centers of cluster 2 : 15158.0
Centers of cluster 3 : 2794.0

Πείραμα 2. Το σύνολο δεδομένων που επιλέχθηκε για την διαδικασία clustering αποτελείται από διδιάστατα στοιχεία με γνωρίσματα το *close*, *high*.

Παράμετροι

Attributes : close, high

Min Clusters: 2

Max Clusters: 10

Crispness Factor : 1

Αποτελέσματα

*For 10 clusters, Validation index : 32.128114937959225

Centers of cluster 0 : 1715.0, 1742.0

Centers of cluster 1 : 2482.0, 2534.0

Centers of cluster 2 : 6417.0, 6540.0

Centers of cluster 3 : 3901.0, 3985.0

Centers of cluster 4 : 891.0, 904.0

Centers of cluster 5 : 578.0, 590.0

Centers of cluster 6 : 1251.0, 1275.0

Centers of cluster 7 : 242.0, 247.0

Centers of cluster 8 : 13111.0, 13425.0

Centers of cluster 9 : 24602.0, 25176.0

*For 9 clusters, Validation index : 29.281675889618647

Centers of cluster 0 : 2148.0, 2189.0

Centers of cluster 1 : 3299.0, 3357.0

Centers of cluster 2 : 10345.0, 10582.0

Centers of cluster 3 : 5495.0, 5621.0

Centers of cluster 4 : 958.0, 972.0

Centers of cluster 5 : 619.0, 630.0

Centers of cluster 6 : 1402.0, 1430.0

Centers of cluster 7 : 280.0, 286.0

Centers of cluster 8 : 17734.0, 18120.0

*For 8 clusters, Validation index : 23.744087635961368

Centers of cluster 0 : 3032.0, 3089.0

Centers of cluster 1 : 5369.0, 5492.0

Centers of cluster 2 : 16825.0, 17199.0

Centers of cluster 3 : 9768.0, 9989.0

Centers of cluster 4 : 1189.0, 1211.0

Centers of cluster 5 : 714.0, 726.0

Centers of cluster 6 : 1835.0, 1867.0

Centers of cluster 7 : 303.0, 310.0

*For 7 clusters, Validation index : 20.027393465228158

Centers of cluster 0 : 2847.0, 2900.0

Centers of cluster 1 : 5295.0, 5417.0

Centers of cluster 2 : 16652.0, 17024.0

Centers of cluster 3 : 9555.0, 9768.0

Centers of cluster 4 : 856.0, 869.0

Centers of cluster 5 : 348.0, 355.0
Centers of cluster 6 : 1552.0, 1580.0

*For 6 clusters, Validation index : 15.645578235715622

Centers of cluster 0 : 2651.0, 2699.0
Centers of cluster 1 : 5025.0, 5139.0
Centers of cluster 2 : 16487.0, 16852.0
Centers of cluster 3 : 9050.0, 9256.0
Centers of cluster 4 : 1373.0, 1398.0
Centers of cluster 5 : 554.0, 564.0

*For 5 clusters, Validation index : 14.271494497170789

Centers of cluster 0 : 1836.0, 1870.0
Centers of cluster 1 : 4296.0, 4383.0
Centers of cluster 2 : 15997.0, 16361.0
Centers of cluster 3 : 8243.0, 8432.0
Centers of cluster 4 : 679.0, 691.0

*For 4 clusters, Validation index : 14.226618485530977

Centers of cluster 0 : 952.0, 970.0
Centers of cluster 1 : 2787.0, 2839.0
Centers of cluster 2 : 15228.0, 15587.0
Centers of cluster 3 : 6299.0, 6431.0

*For 3 clusters, Validation index : 16.237223525128606

Centers of cluster 0 : 1333.0, 1357.0
Centers of cluster 1 : 5739.0, 5863.0
Centers of cluster 2 : 15158.0, 15511.0

*For 2 clusters, Validation index : 25.692830517228565

Centers of cluster 0 : 1868.0, 1905.0
Centers of cluster 1 : 12904.0, 13193.0

Best number of cluster is 4

Centers of cluster 0 : 952.0, 970.0
Centers of cluster 1 : 2787.0, 2839.0
Centers of cluster 2 : 15228.0, 15587.0
Centers of cluster 3 : 6299.0, 6431.0

Πείραμα 3. Το σύνολο δεδομένων που επιλέχθηκε για την διαδικασία clustering αποτελείται από τρισδιάστατα στοιχεία με γνωρίσματα το *close*, *high*, *low*.

Παράμετροι

Attributes : close, high, low

Min Clusters: 2

Max Clusters: 10

Crispness Factor : 1

Αποτελέσματα

*For 10 clusters, Validation index : 28.22966154608347

Centers of cluster 0 : 1718.0, 1746.0, 1681.0

Centers of cluster 1 : 3830.0, 3915.0, 3759.0

Centers of cluster 2 : 6275.0, 6392.0, 6187.0

Centers of cluster 3 : 2482.0, 2534.0, 2436.0

Centers of cluster 4 : 609.0, 620.0, 594.0

Centers of cluster 5 : 901.0, 914.0, 884.0

Centers of cluster 6 : 1257.0, 1282.0, 1228.0

Centers of cluster 7 : 278.0, 285.0, 273.0

Centers of cluster 8 : 12590.0, 12936.0, 12407.0

Centers of cluster 9 : 21070.0, 21422.0, 20718.0

*For 9 clusters, Validation index : 23.68260546548602

Centers of cluster 0 : 2151.0, 2192.0, 2104.0

Centers of cluster 1 : 5519.0, 5644.0, 5429.0

Centers of cluster 2 : 10290.0, 10523.0, 10137.0

Centers of cluster 3 : 3319.0, 3379.0, 3266.0

Centers of cluster 4 : 624.0, 636.0, 609.0

Centers of cluster 5 : 965.0, 979.0, 945.0

Centers of cluster 6 : 1408.0, 1435.0, 1376.0

Centers of cluster 7 : 283.0, 290.0, 278.0

Centers of cluster 8 : 17615.0, 18004.0, 17355.0

*For 8 clusters, Validation index : 19.26559321031757

Centers of cluster 0 : 3032.0, 3089.0, 2980.0

Centers of cluster 1 : 9768.0, 9989.0, 9614.0

Centers of cluster 2 : 16825.0, 17199.0, 16586.0

Centers of cluster 3 : 5369.0, 5492.0, 5283.0

Centers of cluster 4 : 721.0, 733.0, 707.0

Centers of cluster 5 : 1196.0, 1219.0, 1169.0

Centers of cluster 6 : 1839.0, 1871.0, 1799.0

Centers of cluster 7 : 308.0, 314.0, 300.0

*For 7 clusters, Validation index : 16.309465620971032

Centers of cluster 0 : 2859.0, 2912.0, 2808.0

Centers of cluster 1 : 9617.0, 9834.0, 9459.0

Centers of cluster 2 : 16738.0, 17109.0, 16503.0

Centers of cluster 3 : 5307.0, 5429.0, 5223.0

Centers of cluster 4 : 348.0, 355.0, 340.0

Centers of cluster 5 : 860.0, 874.0, 841.0
Centers of cluster 6 : 1559.0, 1587.0, 1524.0

*For 6 clusters, Validation index : 12.663327918076387

Centers of cluster 0 : 2654.0, 2703.0, 2606.0
Centers of cluster 1 : 9031.0, 9234.0, 8881.0
Centers of cluster 2 : 16406.0, 16763.0, 16172.0
Centers of cluster 3 : 5040.0, 5157.0, 4959.0
Centers of cluster 4 : 554.0, 564.0, 542.0
Centers of cluster 5 : 1375.0, 1399.0, 1344.0

*For 5 clusters, Validation index : 11.760318476378142

Centers of cluster 0 : 1923.0, 1959.0, 1883.0
Centers of cluster 1 : 8527.0, 8720.0, 8374.0
Centers of cluster 2 : 16322.0, 16684.0, 16093.0
Centers of cluster 3 : 4367.0, 4459.0, 4302.0
Centers of cluster 4 : 726.0, 738.0, 710.0

*For 4 clusters, Validation index : 11.624708290733924

Centers of cluster 0 : 2836.0, 2891.0, 2781.0
Centers of cluster 1 : 6378.0, 6510.0, 6281.0
Centers of cluster 2 : 15298.0, 15659.0, 15075.0
Centers of cluster 3 : 962.0, 979.0, 941.0

*For 3 clusters, Validation index : 13.257144988275234

Centers of cluster 0 : 1336.0, 1360.0, 1309.0
Centers of cluster 1 : 5752.0, 5877.0, 5661.0
Centers of cluster 2 : 15158.0, 15511.0, 14931.0

*For 2 clusters, Validation index : 20.918821794433036

Centers of cluster 0 : 1874.0, 1911.0, 1838.0
Centers of cluster 1 : 12962.0, 13252.0, 12770.0

Best number of cluster is 4

Centers of cluster 0 : 2836.0, 2891.0, 2781.0
Centers of cluster 1 : 6378.0, 6510.0, 6281.0
Centers of cluster 2 : 15298.0, 15659.0, 15075.0
Centers of cluster 3 : 962.0, 979.0, 941.0

2. ΒΑΘΜΟΙ ΣΥΜΜΕΤΟΧΗΣ ΣΤΟΙΧΕΙΩΝ ΣΤΑ CLUSTERS

Μετά την επιλογή του βέλτιστου σχήματος clustering μπορούμε με την βοήθεια των Hypertrapezoidal Membership Functions να κατανείμουμε τα στοιχεία της Βάσης Δεδομένων μας στα clusters με βάση κάποιο βαθμό πίστης.

Για παράδειγμα, εάν εφαρμόζουμε τις συναρτήσεις συμμετοχής για να κατανήμουμε τα δεδομένα μας στα clusters που προκύπτουν από το δεύτερο πείραμα της παραγράφου 2, θα λαμβάναμε σαν αποτέλεσμα το βαθμό πίστης με τον οποίο ένα στοιχείο ανήκει σε κάθε cluster. Στους Πίνακα A1, Πίνακα A2 παρουσιάζονται ενδεικτικά αποτελέσματα υπολογισμού των βαθμών συμμετοχής στοιχείων του Πίνακα A στα clusters που προκύπτουν από το πείραμα 2, όταν ο παράγοντας επικάλυψης ορίζεται σε 1 και 0,5 αντίστοιχα.

cd	attr	clu	dob
196301	1210.0 1270.0	0	1
196302	1210.0 1270.0	1	0
196303	1210.0 1270.0	2	0
196304	1210.0 1270.0	3	0
196305	1220.0 1220.0	0	1
196306	1220.0 1220.0	1	0
196307	1220.0 1220.0	2	0
196308	1220.0 1220.0	3	0
196309	1215.0 1250.0	0	1
196310	1215.0 1250.0	1	0
196311	1215.0 1250.0	2	0
196312	1215.0 1250.0	3	0
196313	2240.0 2285.0	0	0
196314	2240.0 2285.0	1	1
196315	2240.0 2285.0	2	0
196316	2240.0 2285.0	3	0
196317	2335.0 2335.0	0	0
196318	2335.0 2335.0	1	1
196319	2335.0 2335.0	2	0
196320	2335.0 2335.0	3	0
196321	2295.0 2365.0	0	0
196322	2295.0 2365.0	1	1
196323	2295.0 2365.0	2	0
196324	2295.0 2365.0	3	0
196325	2330.0 2330.0	0	0
196326	2330.0 2330.0	1	1
196327	2330.0 2330.0	2	0
196328	2330.0 2330.0	3	0
196329	2530.0 2680.0	0	0
196330	2530.0 2680.0	1	1
196331	2530.0 2680.0	2	0
196332	2530.0 2680.0	3	0
196333	2695.0 2700.0	0	0
196334	2695.0 2700.0	1	1
196335	2695.0 2700.0	2	0
196336	2695.0 2700.0	3	0

Πίνακας A1. Βαθμοί πίστης όταν ο παράγοντας επικάλυψης ορίζεται ίσος με 1.

cd	attr	clust	dob
1061	1210.0 1270.0	0	1
1062	1210.0 1270.0	1	0
1063	1210.0 1270.0	2	0
1064	1210.0 1270.0	3	0
1065	1220.0 1220.0	0	1
1066	1220.0 1220.0	1	0
1067	1220.0 1220.0	2	0
1068	1220.0 1220.0	3	0
1069	1215.0 1250.0	0	1
1070	1215.0 1250.0	1	0
1071	1215.0 1250.0	2	0
1072	1215.0 1250.0	3	0
1073	2240.0 2285.0	0	9,4451171819
1074	2240.0 2285.0	1	0,9055488281
1075	2240.0 2285.0	2	0
1076	2240.0 2285.0	3	0
1077	2335.0 2335.0	0	1,6266534548
1078	2335.0 2335.0	1	0,9837334654
1079	2335.0 2335.0	2	0
1080	2335.0 2335.0	3	0
1081	2295.0 2365.0	0	2,1299384910
1082	2295.0 2365.0	1	0,9787006150
1083	2295.0 2365.0	2	0
1084	2295.0 2365.0	3	0
1085	2330.0 2330.0	0	2,1660732325
1086	2330.0 2330.0	1	0,9783392676
1087	2330.0 2330.0	2	0
1088	2330.0 2330.0	3	0
1089	2530.0 2680.0	0	0
1090	2530.0 2680.0	1	1
1091	2530.0 2680.0	2	0
1092	2530.0 2680.0	3	0
1093	2695.0 2700.0	0	0
1094	2695.0 2700.0	1	1
1095	2695.0 2700.0	2	0
1096	2695.0 2700.0	3	0

Πίνακας Α2. Βαθμοί πίστης όταν ο παράγοντας επικάλυψης ορίζεται ίσος με 0,5.

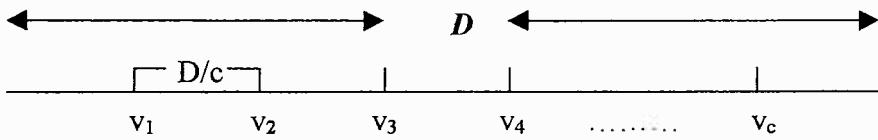
code	date	close	volum	high	low	value	trans
AEGEK	12	1225	34100	1285	1220	42437050	111
AEGEK	13	1245	24810	1275	1240	31106750	78
AEGEK	14	1290	24924	1325	1250	316623100	163
AEGEK	15	1240	35040	1285	1205	43255800	78
AEGEP	12	850	1800	890	850	1564900	9
AEGEP	13	860	4230	870	860	3648300	11
AEGEP	14	876	6780	898	866	5931230	19
AEGEP	15	840	4940	857	820	4144530	15
AFAIN	12	7900	950	7950	7900	7526750	10
AFAIN	13	7960	3240	7960	7900	25697000	16
AFAIN	14	7950	1590	7960	7880	12617350	11
AFAIN	15	7940	6280	7950	7710	49511200	21
AFAIN	16	8000	1790	8000	7880	14162850	11
AIOLC	12	1200	7180	1230	1195	8689400	20
AIOLC	13	1230	4600	1250	1200	5603500	13
AIOLC	14	1210	4850	1270	1210	5988000	17
AIOLC	15	1220	6050	1220	1190	7290000	15
AIOLC	16	1215	10600	1250	1200	12952000	27
AKTOR	12	2240	18610	2285	2220	42167600	74
AKTOR	13	2335	32485	2335	2250	73985600	89
AKTOR	14	2295	37430	2365	2280	86993250	121
AKTOR	15	2330	22855	2330	2200	51579760	87
ALATK	12	2530	13530	2680	2525	34889650	75
ALATK	13	2695	16850	2700	2550	43883650	104
ALATK	14	2910	75110	2910	2755	216801800	282
ALATK	15	3090	69950	3095	2970	212561700	318
ALCO	12	3300	17599	3400	3280	58755705	150
ALCO	13	3290	7881	3300	3235	25721770	76
ALCO	14	3255	14974	3290	3200	48933630	79
ALCO	15	3275	18438	3340	3170	59563130	104
ALEK	12	15880	3510	15950	15445	54789750	19
ALEK	13	15690	1060	15850	15600	16716300	7
ALEK	14	16490	10	16490	16490	164900	1
ALEK	15	15750	7850	15750	15400	121628200	26
ALEPA	12	15000	160	15000	14515	2370900	2
ALEPA	13	14775	560	15000	14715	8286300	9
ALEPA	14	14920	1350	15100	14915	20258650	16
ALEPA	15	14410	5300	14700	14350	76277650	15
ALEPM	12	16000	100	16000	16000	1600000	1
ALEPM	13	16400	60	16400	16400	984000	1
ALEPM	15	15965	1100	15965	15100	17165750	10
ALKΑ	12	2500	7990	2570	2455	20054500	42
ALKΑ	13	2550	5390	2645	2465	13908750	28
ALKΑ	14	2570	20770	2620	2505	53345600	102
ALKΑ	15	2510	2390	2550	2510	6028400	11
ALKAR	12	415	16770	429	400	7059830	40

Πίνακας Β. Ενδεικτικά δεδομένα συναλλαγών που πραγματοποιήθηκαν στις 12-15 Ιανουαρίου 1998

ΠΑΡΑΡΤΗΜΑ Β'

Θεωρούμε ένα μονοδιάστατο σύνολο δεδομένων X του οποίου τα στοιχεία ακόλουθούν την ομοιόμορφη κατανομή. Έστω D είναι η απόσταση μεταξύ των πιο απομακρυσμένων στοιχείων του συνόλο και c ο αριθμός των clusters. Η απόσταση μεταξύ των κέντρων γειτονικών clusters θα είναι D/c . Ετσι μπορούμε να ορίσουμε την απόσταση μεταξύ των κέντρων των clusters με βάση το σχήμα B.1 και την εξίσωση (1):

$$dist = \frac{2D \sum_{j=1}^{c-1} \sum_{i=1}^{c-j} i}{c} \quad (1)$$



Σχήμα B.1: Ομοιόμορφη κατανομή, απόσταση μεταξύ των κέντρων των clusters

Η μέση απόσταση μεταξύ των κέντρων των clusters ορίζεται από την εξίσωση (2):

$$d = \frac{2D \sum_{j=1}^{c-1} \sum_{i=1}^{c-j} i}{c^2(c-1)} \quad (2)$$

Λαμβάνοντας υπόψη την ομοιόμορφη κατανομή, η διακύμανση για όλα τα clusters του συνόλου των δεδομένων θα είναι η ίδια. Η ακόλουθη εξίσωση (3) ορίζει την διακύμανση για το cluster i .

$$\sigma_i = \frac{1}{12} \left(\frac{D}{c} \right)^2 \quad (3)$$

Η συνολική διακύμανση του συνόλου δεδομένων, λαμβάνοντας υπόψη την ομοιόμορφη κατανομή, ορίζεται από την εξίσωση (4):

$$\sigma = \frac{1}{12} D^2 \quad (4)$$

Σύμφωνα με τους παραπάνω ορισμούς και την εξίσωση (Εξισ. 6.4) η μέση διαφοροποίηση της πυκνότητας των c clusters ισούται με:

$$comp_scat = \frac{1}{c^2} \quad (5)$$

Βασιζόμενοι στην εξίσωση (Εξισ. 6.6), η τιμή του μέτρου ποιότητας CD για την ομοιόμορφη κατανομή ορίζεται ως εξής:

$$CD = \frac{1/c^2}{2D \sum_{j=1}^{c-1} \sum_{i=1}^{c-j} i / c^2 (c-1)} = \frac{(c-1)}{2D \sum_{j=1}^{c-1} \sum_{i=1}^{c-j} i} \quad (6)$$

Σύμφωνα με την παραπάνω εξίσωση, ο παρανομαστής αυξάνεται γρηγορότερα σεσχέση με τον αριθμητή όταν αυξάνεται ο αριθμός των clusters, c . Συνεπώς, είναι φανερό ότι το μέτρο CD παρουσιάζει σημαντική τάση μείωσης με την αύξηση του αριθμού των clusters c .

A. ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ-ΕΙΚΟΝΩΝ

Σχήμα 1.1. Χρήση απλού ορίου για την μεταβλητή "εισόδημα" προκειμένου να κατηγοριοποιήσουμε το σύνολο δεδομένων για τα δάνεια.	5
Σχήμα 1.2. Τα στάδια που αποτελούν την διαδικασία KDD.	8
Σχήμα 1.3. Δέντρα Αποφάσεων.	17
Σχήμα 1.4. Δομή ενός νευρωνικού δικτύου.	18
Σχήμα 1.5. Η μονάδα επεξεργασίας(κόμβος) του νευρωνικού δικτύου.	19
Σχήμα 2.1. Δενδρογράφημα.	21
Σχήμα 2.2. Δίκτυο Kohonen.	24
Σχήμα 2.3. Υπολογισμός activation level σε ένα δίκτυο Kohonen.	25
Σχήμα 2.4. Αρχικοποίηση K-means.	27
Σχήμα 2.5. Διαδικασία Εκλέπτυνσης Αρχικών Σημείων	31
Σχήμα 4.1. Προσεγγίσεις μέτρησης της απόστασης μεταξύ των clusters.	63
Σχήμα 5.1. Στάδια μεθοδολογίας για υποστήριξη αβεβαιότητας στο Data Mining.	78
Σχήμα 6.1. Βήματα διαδικασίας Fuzzy Clustering.	86
Σχήμα 6.2. Επίδραση του <i>crispness factor</i> (σ) σε διδιάστατα σύνολα.	100
Σχήμα 6.3. Επίδραση παράγοντα σ σε μονοδιάστατα σύνολα.	100
Σχήμα 6.4. Ορισμός επικάλυψης ασαφών τμημάτων.	100
Σχήμα 6.5. Αντιστοίχηση clustering σχημάτων σε CVS.	102
Σχήμα 6.6. Γνώση που περιέχει ο κύβος ανά γνώρισμα.	103
Σχήμα 6.7. Επιλογή του καλύτερου clustering σχήματος.	104
Εικόνα 1. Οθόνη Εκκίνησης.	115
Εικόνα 2. Οθόνη Σύνδεσης.	116
Εικόνα 3. Οθόνη Επιλογής Πίνακα.	117
Εικόνα 4. Κεντρική Οθόνη Clustering.	119
Εικόνα 5. Οθόνη Αποτελεσμάτων Clustering.	120
Σχήμα B.1: Ομοιόμορφη κατανομή, απόσταση μεταξύ των κέντρων των clusters.	133

B. ΕΥΡΕΤΗΡΙΟ ΔΙΑΓΡΑΜΜΑΤΩΝ

Διάγραμμα 6.1. Διάγραμμα μεταβολής του μέτρου ποιότητας clustering σε σχέση με την μεταβολή των clusters. Το clustering αφορά σε διδιάστατα στοιχεία με την μορφή (τιμή κλεισίματος μετοχής, υψηλότερη τιμή μετοχής).	92
Διάγραμμα 6.1α. Μεταβολή της μέσης διαφοροποίησης της πυκνότητας μέσα στα clusters σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε σύνολο δεδομένων του οποίου τα στοιχεία είχαν την μορφή τιμή κλεισίματος μετοχής, μέγιστη τιμή μετοχής).	93
Διάγραμμα 6.1β. Διάγραμμα που παρουσιάζει την μεταβολή της μέσης απόστασης μεταξύ των clusters σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε σύνολο δεδομένων του οποίου τα στοιχεία είχαν την μορφή (τιμή κλεισίματος μετοχής, μέγιστη τιμή μετοχής).	93
Διάγραμμα 6.2. Διάγραμμα που παρουσιάζει την μεταβολή του μέτρου ποιότητας clustering σε σχέση με την μεταβολή των clusters. Το clustering αφορά σε μονοδιάστατα στοιχεία με την μορφή (τιμή κλεισίματος μετοχής).	94
Διάγραμμα 6.2α. Μεταβολή της μέσης διαφοροποίησης της πυκνότητας μέσα στα clusters σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε σύνολο δεδομένων του οποίου τα στοιχεία είχαν την μορφή (τιμή κλεισίματος μετοχής).	94
Διάγραμμα 6.2β. Διάγραμμα που παρουσιάζει την μεταβολή της μέσης απόστασης μεταξύ των clusters σε σχέση με τον αριθμό των clusters. Το clustering έγινε σε σύνολο δεδομένων του οποίου τα στοιχεία είχαν την μορφή (τιμή κλεισίματος μετοχής).	94



Διάγραμμα 6.4. Μεταβολή του μέτρου ποιότητας SD σε σχέση με τον αριθμό των clusters.	
Το clustering έγινε σε ένα σύνολο δεδομένων με διδιάστατα στοιχεία (τιμή κλεισίματος μετοχής, υψηλότερη τιμή μετοχής).....	96
Διάγραμμα 6.5. Μεταβολή του μέτρου ποιότητας SD σε σχέση με τον αριθμό των clusters.	
Το clustering έγινε σε ένα σύνολο δεδομένων με μονοδιάστατα στοιχεία (τιμή κλεισίματος μετοχής).....	96
Διάγραμμα 6.6. Επίδραση της μέγιστης τιμής του αριθμού των clusters στην τιμή του μέτρου ποιότητας. Το clustering αφορά στοιχεία διδιάστατα (τιμή κλεισίματος μετοχής, υψηλότερη τιμή μετοχής).....	97
Διάγραμμα 6.7. Επίδραση της μέγιστης τιμής του αριθμού των clusters στην τιμή του μέτρου ποιότητας. Το clustering αφορά στοιχεία μονοδιάστατα (τιμή κλεισίματος μετοχής).....	97
Διάγραμμα 7.1. Μεταβολή του χρόνου εκτέλεσης σε σχέση με τον αριθμό των στοιχείων του συνόλου δεδομένων.....	114
Διάγραμμα 7.2. Μεταβολή του χρόνου εκτέλεσης σε σχέση με τα όρια στα οποία κυμαίνεται ο αριθμός των clusters στην διαδικασία αναζήτησης βέλτιστου σχήματος clustering ..	114

Γ. ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

Πίνακας 4.1. Κλασικά μέτρα Ποιότητας <i>Fuzzy Clustering</i>	69
Πίνακα 5.1. Συγκεντρωτικός πίνακας με τα μέτρα γνώσης.....	82
Πίνακας 7.1 Κλάση για την υλοποίηση του scaling	107
Πίνακας 7.2 Κλάση για την υλοποίηση της Ευκλείδειας αποστάσεως	107
Πίνακας 7.3 Κλάση υλοποίησης αλγορίθμου K-Means	109
Πίνακας 7.4. Κλάση για την υλοποίηση του μέτρου ποιότητας	110
Πίνακας 7.5 Κλάση υλοποίησης των Hypertrapezoidal Membership Functions	111
ΠίνακαςΑ1. Βαθμοί πίστης όταν ο παράγοντας επικάλυψης ορίζεται ίσος με 1.....	130
ΠίνακαςΑ2. Βαθμοί πίστης όταν ο παράγοντας επικάλυψης ορίζεται ίσος με 0,5.....	131



ΒΙΒΛΙΟΓΡΑΦΙΑ

- [AGGR98] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopoulos, Prabhakar, Raghavan. "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications." IBM Almaden Research Center.
- [AF96] K.S. Al-Sultan, C.A.Fedjki, "A Tabu Search-Based Algorithm For The Fuzzy Clustering Problem", *Pattern Recognition*. Vol 30, No12, pp. 2023-2030, 1997.
- [AT98] Αμανατίδης Χρήστος, Γιαλίδης - Τζούρης Μενέλαος. *Πρακτική Άσκηση Φοιτητών: Data Mining*. Οικ. Πανεπιστήμιο Αθηνών, Σεπτέμβριος 1998.
- [Berry 97] Michael J. A. Berry, Gordon Linoff. *Data Mining Techniques For marketing, Sales and Customer Support*. John Wiley & Sons, Inc, 1996.
- [BF98] P. S. Bradley, Usama M. Fayyad. "Refinining Initial Points for K-Means Clustering".
<http://www.research.microsoft.com/research/dtg/fayyad/papers/icml.htm>
- [BFR98] P. S. Bradley, Usama Fayyad, Cory Reina. " Scaling Clustering Algormths to Large Databases".
<http://www.research.microsoft.com/research/dtg/fayyad/papers/>
- [CHY96] Ming-Syan Chen, Jiawei Han, Philip S. Yu. "Data Mining: An Overview from a Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*. Vol8, No6, Dec 96.
- [Dave96] Rajesh N. Dave. "Validating fuzzy partitions obtained through c-shells clustering", *Pattern Recognition Letters*, Vol. 17, pp 613-623, 1996
- [EKSWX98] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Michael Wimmer, Xiaowei Xu. "Incremental Clustering for Mining in a Data Warehousing Environment", *Proceedings of 24th VLDB Conference, New York, USA*, 1998.
- [FU96] Usama Fayyad, Ramasamy Uthurusamy. "Data Mining and Knowledge Discovery in Databases", *Communications of the ACM*. Vol 39, No11, Nov 1996.
- [FPSU96] Usama M. Fayyad, Gregory Piatesky-Shapiro, Padhraic Smuth and Ramasamy Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press 1996

- [GG89] I. Gath, B. Geva. "Unsupervised Optimal Fuzzy Clustering". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 11, No7, July 1989.
- [GRK98] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim. "CURE: An Efficient Clustering Algorithm for Large Databases", *Published in the Proceedings of the ACM SIGMOD Conference*, 1998.
<http://www.bell-labs.com/project/serendip/>
- [GRK99] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim. "ROCK: A Robust Clustering Algorithm for Categorical Attributes", *Published in the Proceedings of the IEEE Conference on Data Engineering*, 1999.
<http://www.bell-labs.com/project/serendip/>
- [H96] Jiawei Han. Data Mining Techniques. *ACM-SIGMOD '96 CONFERENCE TUTORIAL*. <http://fas.sfu.ca/cs/research/groups/DB/>
- [HBD96] Richard J. Hathaway, James C. Bezdek, John W. Davenport. "On relational data versions of c-means algorithm", *Pattern Recognition Letters*, Vol 17, pp. 607-612, 1996.
- [HB94] Richard J. Hathaway, James C. Bezdek. "NERF c-Means: Non-Euclidean Relational Fuzzy Clustering", *Pattern Recognition Letters*, Vol 27, No 3, pp. 428-437, 1994.
- [Huang97] Zhexue Huang. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining", *DMKD*, 1997.
- [JWeber] Joseph Weber, et al, *Sopecial Edition: Using JAVA*, QUE
- [Ksam 97] Sami Kaski. *Clustering methods*.
<http://www.cis.hut.fi/~sami/thesis/node9.html>
- [KNN1] *Kohonen Neural Network*.
<http://mmlin1.pha.unc.edu/%7Ejin/QSAR/KNN2/som.html>
- [KNN2] *Kohonen Neural Network*.
http://ppl.mines.colorado.edu/bill/bak/subsection3_6_1.html
- [KANN] *Kohonen artificial neural networks*.
<http://gepasi.dbs.aber.ac.uk/roy/koho/kohonen.htm>
- [KOP97]. Krzysztof (Kris) Koperski. *Methods Using Clustering*.
<http://db.cs.sfu.ca/GeoMiner/survey/html/node9.html>.
- [KP96] Wallace E. Kelly, John H. Painter. *Hypertrapezoidal Membership Function for Decision Aiding*.
<http://joshua.tamu.edu/astra/Library/WSC2/index.html>
- [LLM96] Laura Lemay, Charles Lperkins, Michael Morrison. *Εγχειρίδιο της JAVA*. 1996 Sams Net Publishing.
- [Pedrycz95] Witold Pedrycz. "Conditional Fuzzy C-Means", *Pattern Recognition Letters*, Vol 17, pp625-631, 1996.

- [RJ94] Raymond T.Ng, Jiawei Han. "Efficient and Effective Clustering Methods for Spatial Data Mining", *Proceedings of 20th VLDB Conference*. Santiago, Chile, 1994.
- [RR98] Ramze Rezaee, B.P.F. Lelieveldt, J.H.C Reiber. "A new cluster validity index for the fuzzy c-mean", *Pattern Recognition Letters*, Vol19, pp237-246, 1998.
- [Raat93] Kimmo E.E. Raatikainen. "Cluster Analysis and Workload Classification" *Performance Evaluation Review*, Vol 20, No4, May 1993.
- [Kaski97] Sami Kaski. *Clustering Methods*. (1997)
<http://www.cis.hut.fi/~sami/thesis/node9.html>
- [Vaz98] M. Vazirgiannis, "A classification and relationship extraction scheme for relational databases based on fuzzy logic", *in the proceedings of the Pacific -Asian Knowledge Discovery & Data Mining '98 Conference*, Melbourne, Australia.
- [XB91] Xunali Lisa Xie, Genardo Beni. "A Validity measure for Fuzzy Clustering", *IEEE Transactions on Pattern Analysis and machine Intelligence*, Vol13, No4, August 1991.

