FUZZY BAYES CLASSIFICATION

NECLA KAYAALP

MAY 2013

FUZZY BAYES CLASSIFICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
NATURAL AND APPLIED SCIENCES OF
IZMIR UNIVERSITY OF ECONOMICS

BY

NECLA KAYAALP

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE
IN THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

MAY 2013

M.S. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"FUZZY BAYES CLASSIFICATION"** completed by **Necla Kayaalp** under supervision of **Asst. Prof. Dr. Güvenç Arslan** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Güvenç Arslan
Supervisor

Thesis Committee Member

Thesis Committee Member

Director

ABSTRACT

FUZZY BAYES CLASSIFICATION

Necla Kayaalp
M.S. in Applied Statistics
Graduate School of Natural and Applied Sciences
Supervisor: Asst. Prof. Dr. Güvenç Arslan
May 2013

Recent developments show that Naive Bayesian Classifier (NBC) performs significantly better in applications although it is based on the assumption that all attributes are independent of each other. However, in the NBC each variable has a finite number of values which means that in large data sets NBC may not be so effective in classifications. For example, variables may take continous values. To overcome this issue many researchers used Fuzzy Naive Bayesian Classification (FNBC) for partitioning the continous values. On the other hand the choice of distance function is an important subject that should be taken into consideration in fuzzy partitioning or clustering.

In this thesis, a new Fuzzy Bayes Classification is proposed for numerical attributes without considering the independence assumption. In order to get high accuracy in classification membership functions are constructed by using Fuzzy C-Means Clustering (FCM). The main objective in using FCM is to obtain membership functions directly from the data set instead of consulting to an expert. The proposed method is demonstrated on two well-known data sets from the literature which consist of numerical attributes only. The results show that the proposed Fuzzy Bayes Classification is at least as well as comparable to other methods.

Keywords: Mahalanobis distance; Bayes classification; Fuzzy set theory.

ÖZ

BULANIK BAYES SINIFLANDIRICISI

Necla Kayaalp
Uygulamalı İstatistik, Yüksek Lisans
Fen Bilimleri Enstitüsü
Tez Yöneticisi: Yrd. Doç. Dr. Güvenç Arslan
Mayıs 2013

Son gelişmeler, tüm niteleyicilerin (attribute) birbirinden bağımsız olduğu varsayımına dayanmasına rağmen, Naive Bayes Sınıflamanın (NBS) uygulamalarda oldukça iyi bir biçimde işlediğini göstermektedir. Bununla birlikte, NBSde, her değişken, büyük veri kümelerinin sınıflamalarında çok da etkin olamayacağı anlamına gelen sınırlı sayıda değere sahiptir. Sözgelimi, değişkenler sürekli değerler alabilirler. Bu sorunun üstesinden gelmek için birçok araştırmacı sürekli değerleri bölüntülemek (kesikli hale getirmek) için Bulanık Naive Bayes Sınıflamasını (BNBS) kullanr. Öte yandan, uzaklık fonksiyonu seçeneği, bulanık bölüntülemede ya da kümelemede dikkate alınması gereken önemli bir konudur.

Bu tezde, bağımsızlık varsayımı dikkate alınmadan sayısal niteleyiciler için yeni bir Bulanık Bayes Sınıflaması önerilmiştir. Sınıflamada, yüksek doğruluğu elde etmek için, Bulanık C-Means Kümelemesi (BCM) kullanılarak üyelik fonksiyonları oluşturulmuştur. BCM kullanımındaki temel amaç, bir uzmana danışmak yerine üyelik fonksiyonlarını doğrudan veri setinden elde etmektir. Önerilen yöntem, yalnızca sayısal niteleyicileri içeren ve alanyazında iyi bilinen iki veri seti üzerinde gösterilmiştir. Sonuçlar, önerilen Bulanık Bayes Sınıflamasının en azından diğer yöntemlerle karşılaştırılabilir olduğunu ve genellemede daha iyi olduğunu göstermektedir.

Anahtar Kelimeler: Mahalanobis uzaklığı; Bayes sınıflandırıcısı; Bulanık küme teorisi.

ACKNOWLEDGEMENT

I would like to thank a number of people who helped and supported me. This work began under supervision of Assoc. Prof. Dr. G. Yazgı Tütüncü, continued and expanded with excellent guidance and encouragements of Yrd. Doc. Dr. Güvenç Arslan and completed under the supervision of Asst. Prof. Dr. Güvenç Arslan. I could not finish my thesis without them with I whom explored the ideas, organization, requirements and development of my thesis. Also I am grateful to Asst. Prof. Dr. Uğur MADRAN who gave technical support. I should also thank to Cihangir Kan, Murat Özgen, Sevim Ulusoy, Burçin Külahçıoğlu and Saygın Karabıyık for being there whenever I needed.

Last but not at least, I thank to my big family who I cannot count for giving me life in the first place, for educating me, and for their unconditional support. I hope that my thesis will always be useful and serve the mankind at its best.

TABLE OF CONTENTS

Chapter 1

Introduction

A main objective of data analysis is to process and organize data in order to extract useful information from it. It is known that there has been a huge amount of information pollution in recent years due to the fact that there is too much data to be processed in data analysis because of the increasing use of computers. In addition to the increasing use of computers and the internet, easy access to data and increasing capacities of data storage led to the potential of misuse of data. Another problem is how to deal with such large data sets. Moreover, there may be redundant data, which should be avoided, if possible. In fact, today it can be said that traditional approaches to data analysis and information processing are not effective or even not appropriate at all. Another important aspect to be considered nowadays is to guard against poor quality of data to be processed. Various approaches have been proposed to eliminate such handicaps and to overcome similar problems during data analysis. One of the most common approaches is, classification, which is a data analysis technique especially used in machine learning, pattern recognition, decision making problems, and communication networks. By using classification procedures, we can discover new information from significant amount of data. Therefore, the classification speed becomes an important task in classification in order to save time and such procedures help to make classification as soon as possible. Moreover, because of overwhelming developments in science and technology, information or data is changing in a rapid

fashion as well. Thus, there exists abundance of data which also has to be processed by appropriate and sometimes new approaches in classification procedures, which may help to update changes and make accurate classifications.

Classification has a very broad field of study. The main task of classification is to group instances by using relevant features. For this reason a classifier can be defined as a function that maps a class label to instances by some of the available features [21]. A data set may sometimes consist of pre-classified instances. This is called supervised classification. In supervised classification, we have a set of observations with given classes and a new observation is classified into one of the given classes. More briefly, observations are pre-classified or class labels are known so that it can be predictive. Unlike the supervised classification, in unsupervised classification, all given observations are unclassified. That is, unsupervised learning occurs when observations have not been previously classified. Therefore, it is not predictive in contrast to supervised classification.

One of the widely used classifiers is the Naive Bayesian Classifier (NBC) which was proposed by Duda and Hart in 1973 [15]. This classifier is constructed with the strong independence assumption of conditional probabilities. This means that all attributes are conditionally independent given the label of the class C. It can be observed that naive Bayesian classifier has been widely used and gives surprisingly high performance in classification despite those strong assumptions. However, many researchers stressed their concern with those assumptions and tried to improve this classifier by considering various extensions.

In recent years researchers also considered additional aspects such as uncertainty, quality of data, and effects of distance functions used in the algorithms, that needs further investigation. When naive Bayesian classifier is faced with uncertain (imperfect) data, its performance may decrease [21]. During the past several years, researchers have been attracted by uncertainty in classification problems. Randomness, which is related to probability theory, is used for eliminating uncertainty. However, it is well known that not all types of uncertainties can be handled by randomness. On the other hand they may sometimes be expressed by human knowledge which can be handled with fuzzy set theory. In

such situations, attributes are expressed by fuzzy membership functions within the perspective of fuzzy logic. These membership functions then can be used in classification. Uncertainty problems may arise in knowledge based systems since usually experts are consulted when membership functions are constructed. Although expert systems began to attract researchers interests, it still may reduce the performance of classification and this may lead to misclassification. In order to get an efficient performance in classification, distance functions are taken into consideration in order to obtain membership functions instead of consulting to an expert. In recent years one can find many studies which investigate their use and performance in various algorithms such as clustering or classification. At the same time, the choice of distance functions is another topic of research within this methodology.

One of the main goals in this study is to use fuzzy set theory in the framework of Bayesian classification to deal with uncertainty related to membership functions. It is known from the literature that constructing appropriate membership functions is not an obvious task. Different approaches have been proposed in order to construct membership functions from experts or training sets. For instance, Wu and Chen used α-cuts of fuzzy equivalence relations for generating rules from training set [37]. In [37], one can find other methods studied in the past. However, most of them are related to fuzzy if-then rules, that is if-then rules are generalized in those studies. Moreover, Hasuike et al., developed a method based on Shannon entropy and Human's interval estimation in order to construct membership functions [38]. Recently, researchers became interested in fuzzy functions in order to generalize membership functions. For details see [41]. In the proposed approach the fuzzy membership functions are constructed directly from the data set by applying Fuzzy C-Means (FCM) algorithm. Namely, a **learned** Mahalanobis distance is used in FCM algorithm after a Mahalanobis distance is learned.

The thesis is organized as follows. In chapter 2, some preliminaries are presented and some basic concepts are overviewed. In chapter 3, a special type of classification, Bayesian classification is summarized. The emergence and development of this type of classification is described as well. In chapter 4, a new

approach to naive Bayesian classification which is called Fuzzy Naive Bayesian Classification, is given. Besides, not only distance metric learning but also FCM clustering algorithm is explained. In chapter 5, the proposed method is applied on two data sets, the Fisher Iris data set and Seed data set from the literature [39]. In chapter 6, the results are analyzed and further directions for research of this study are given.

Chapter 2

Classification Problem

In this chapter, the classification problem is outlined and different classification methods are shortly given. Classification is one of the most commonly applied data mining methods used to obtain information from data. In other words, classification is a process of learning or data mining. In this thesis, one of our goal basically is data classification. Data classification is a process consisting of two steps; the first step is the learning step where a model is established. The second step is the classification step where new examples are classified by using the model established in the first step.

In general a classifciation task uses a training set, in which there is a set of observations and each observation is described by a set of attributes with known class values. Observations are also known as examples, tuples or instances. More formally, each observation is described as $(X; C)$, where X is an attribute vector and C is a class label. A classifier is a function mapping every attribute vector (input data) to the corresponding class labels (output data). In other words, it is a function that assigns a class label to the observations.

A training set or data set in classification can be considered as an information system (IS) composed of attributes and observations such that $IS = (U, A, V_a, f_a)$, where $A = (a_1, a_2, ..., a_n, C)$, is a finite set of attributes including both conditional and decision attributes, $U = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$ is a finite set of

Instances	a_1	a_2	a_3	...	a_n	C
$\mathbf{x}_1$	x_{11}	x_{12}	x_{13}	...	x_{1n}	c_1
$\mathbf{x}_2$	x_{21}	x_{22}	x_{23}	...	x_{2n}	c_2
$\mathbf{x}_3$	x_{31}	x_{32}	x_{33}	...	x_{3n}	c_3
...	...	...	...	....	...	...
$\mathbf{x}_m$	x_{m1}	x_{m2}	x_{m3}	...	x_{mn}	c_q

Table 2.1: Training Set

observations (universe), V_a is the domain of attributes and $f_a : U \longrightarrow V_a$ is an information function such that $f(X, a) \in V_a$ for each $a \in A, X \in U$. The attributes $a_i, i = 1, 2, ..., m$ are called conditional attributes and C is called a decision attribute or class label [29]. We will denote the finite number of different classes by $c_i, i = 1, 2, ..., q$; that is $f(x, c) \in V_c$, where $V_c = \{c_1, c_2, ..., c_q\}$. Table 2.1 is another way for representing such an IS with one decision attribute.

It is seen that $a_1, a_2, ..., a_n$ are attributes, C is a decision attribute or class label and $\mathbf{x}_i$'s are observations. For example, consider observation $\mathbf{x}_1 = (x_{11}, x_{12}, , x_{1n})$, which is an n-dimensional attribute vector corresponding to the attributes $a_1, a_2, ..., a_m$, respectively.

The above definition of a classification problem can be found in many books which deal with the classification problem. However, in this study, we assume that the classification problem can be defined as follows.

Definition. Classification task is composed of a training set $D = (U, A, f, V_c)$, where U is a finite set of observations, A is a finite set of attributes, f is an information function and V_c is a set of class labels. Assume that each class can be represented by a class center which is unknown. We denote those class centers by $V = \{v_1, v_2, ..., v_n\}$. In other words, it is assumed that the space X is partitoned into q well-defined subsets. Under this assumption each class can be identified by a Voronoi diagram [6] as follows:

$$Vor(v_i) = \left\{x \in \mathbb{R}^d | d(v_i, x) \leq d(v_j, x), \forall v_j \in V\right\} \tag{2.1}$$

Many algorithms have been proposed for classification. Some well-known

classification algorithms are given below:

- Fisher's Linear Discriminant[7]: This is one of the oldest classification used and implemented in computer packages. The aim of this approach is to find a projection of high dimensional data onto a low dimensional space such that the distance within the classes is minimized. To do so, variance and covariance matrices are used. The problem emerges when the covariance matrix is singular. This happens, for example, when we have the total number of observations smaller than the total number of attributes.

- Rule Based Method [8]: This classifier is a function of attributes based on rules. It is very practical in a small data set, though not very convenient for large data sets. Another problem is how to define rules and make generalization of those rules to obtain meaningful classifications.

- K-Nearest Neighbor [9]: In this method, a new object is classified with respect to the closest k-objects in the training set. Even though it seems very basic and very applicable, there are some shortcomings. For instance, the distribution of observations may be sophisticated and may not be comprehensible. In addition, it may be very slow compared to other methods.

- Neural Networks [10]: Networks have the ability to get information from complicated data sets. They were very popular in the past but now they are not the one applied because of the complexity. Therefore, one cannot figure out an efficient algorithm for classification. Those algorithms are learned by examples which is changing all the time; thus, such algorithms cause too much waste of time and misclassifications.

- Decisicion Tree [11]: Decision tree is a powerful, widely used classification by using a tree structure. This approach is based on partioning the sample space, that is, each space is divided into subspaces and those subspaces are also divided into other subspaces and so forth. This approach is similar to rule based method, major difference being in the manner of drawing a tree structure. Furthermore, representing rules by a decision tree makes it more attractive than neural networks. It becomes very simple to interpret and

classify objects. Nonetheless, the main concern of this approach is that it gets much more complicated while the decision tree is growing.

- Support Vector Machines: This method is an attractive type of classification that is used mostly nowadays. This classifier is based on decision planes which can be a linear plane or a curve. These decision planes which have decision boundaries split a set of examples which have different class memberships. (For more details see [12].)

- Bayesian Networks [13]: This method is also called belief networks. Bayesian approach depending on probability distributions was firstly used by Fisher. (For details see [14].) Naive Bayes Classification is a special case of Bayesian Networks under some independence assumptions. This method will be discussed in details in the following sections.

Chapter 3

Bayesian Classification

Throughout this chapter Bayes' theorem is overviewed and Bayesian Classification is represented in details. Recent studies are given as well. Bayesian classification is a statistical classification method based on Bayes Theorem. Duda and Hart [15] studied this method in pattern recognition for the first time. Following this, Clark and Niblett pointed out the practical importance of Bayes Classification [17]. Then, in 1992, Langley et al. [18] improved this method.

3.1 Bayes' Theorem and Bayesian Classification

Bayes' theorem which took its name from an English man was first proposed in 18th century by Thomas Bayes, who studied probability and decision theories. There exist two distinct interpratations which are the Bayesian interpratation and the frequentist interpretation of Bayes theorem, in probability theory. While in the Bayesian interpretation, probability measures a degree of belief, in the frequentist interpretation, probability measures a proportion of outcomes. This theorem is based on a simple idea, which is, using prior knowledge, in order to obtain posterior knowledge about an observation. In other words, according to the Bayesian interpretation, the posterior probability can be derived from an appropriate prior probability and a "likelihood function" derived from a probability

model for the data to be observed. Let $\mathbf{x}$ be an observation. Then the posterior probability is defined as:

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C)p(C)}{p(\mathbf{x})} \tag{3.1}$$

where
C stands for a class label
$p(C)$ is posterior probability of C before $\mathbf{x}$ is observed
$p(C|\mathbf{x})$ is posterior probability of C given $\mathbf{x}$, that is, after $\mathbf{x}$ is observed
$p(\mathbf{x}|C)$ is likelihood, the probability of observing $\mathbf{x}$ given C
$p(\mathbf{x})$ is marginal likelihood of observed $\mathbf{x}$, assuming that $p(x) > 0$.

Bayesian classification is a classification method that has many different variants depending on structure and assumptions of attributes, such as independence. Many different types of this classification approach are proposed, such as naive bayes, semi-naive bayes, selective naive bayes and others. A specific kind of this approach is known as the Naive Bayesian Classifier (NBC) which has totally independent attributes. Unless the attributes are totally independent, semi-naive bayes type of classification is used [19].

3.2 Naive Bayesian Classification(NBC)

Consider a training set of examples (tuples) D together with their related class labels. Each tuple, X, n-dimensional attribute vector, is represented as $X = (x_1, ..., x_n)$, showing m values of n attributes $a_1, a_2, ..., a_n$, respectively. Thus, each attribute represents a feature vector. Assume that there are q classes; $c_1, c_2, ..., c_q$. Given a new tuple or example, the classifier will estimate the class, by using the maximum posterior probability given X. Namely, $p(c_i|X)$ will be maximized. This probability will be evaluated with Bayes theorem

$$p(C_i|X) = \frac{p(X|C_i)p(C_i)}{p(X)}. \tag{3.2}$$

Since $p(X)$ is the same for all classes, just the maximization of $p(X|C_i)p(C_i)$ is required. However, it is known that X is an n-dimensional attribute vector which

may be computationally unattractive to compute. To simplify the computation of naive Bayesian classification, in which attribute values are independent of each other, under the given class, the following formula can be used

$$p(X|C_i) = \prod_{k=1}^{n} p(\mathbf{x}_k|C_i), \quad i = 1, 2, ..., n. \tag{3.3}$$

Therefore, $p(C_i) \prod_{k=1}^{n} p(\mathbf{x}_k|C_i)$ will be maximized in order to classify a new instance. More formally,

$$\text{NBC}(C^*) = \arg\max_{C_i} \left\{ p(C_i) \prod_{k=1}^{n} p(\mathbf{x}_k|C_i) \right\} \tag{3.4}$$

where C^* is a new class to be determined, X_k's are the attributes for the given X instance, C_i represents each possible class and $p(\mathbf{x}_k|C_i)$ represents the posterior probability of instance x given class i. In 2009, a naive Bayesian possibilistic network classifier which is very similiar to NBC was introduced by Haouari et al. [21] to cope with imperfect data and uncertainty that may occur in data sets. The difference is that NBC depends on probability theory while this approach depends on possibility theory.

It is obvious to see that the aim of Bayesian classification is to find the best class for a new observation by using probability, namely, new observation is assigned to the best class with the highest probability. Due to the fact that the data sets are spesific data sets, the resulting probabilities may not reflect the real scores and therefore it may result in low performance in classification. It is not a realistic classification although it can generate a confidence value with respect to its choice. Clark and Niblett, in 1989 showed that despite all these negations, naive bayes classification has the highest performance among other types. Recent researches declared the good performance of naive Bayesian classification, and in 1997, Friedman [20] analyzed this classification method illustratively. Nevertheless, owing to the dependency/indpendency assumptions, it is shown that there exist a classification error [22]. It is mentioned that NBC is based on the conditional independence assumption. Hence, NBC is extended through the independency conditions such as selective naive Bayes classification [23], semi-nave Bayes

classification [19], tree augmented naive Bayes classification [20], K-dependence Bayesian classification [24], etc. Indeed, dependency between attributes may reflect the accuracy of classification. For instance, the aim of selective naive Bayes classification is to eliminate redundant attributes in order to make efficient classification. Moreover, in order to get high accuracy in classification weighted naive Bayes classification based on frequency of attributes or correlation coefficient is proposed and experiments have proved effectiveness of this method. Weighted Naive Bayes Classifier (WNBC) was proposed for first time as:

$$\mathrm{WNBC}(C^*) = \arg\max_{C_i} \left\{ p(C_i) \prod_{k=1}^{n} p(X_j|C_i)^{w_j} \right\} \tag{3.5}$$

where w_j is the weight of j-th attribute [25]. Thus, when weight of an attribute is great then that means its impact is great. This method is constructed for data set containing linguistic attribute values. Another important key in this model is how weights of each attributes are determined. There have been many methods used for determining weights. In the method proposed in the next chapter we use weights in a different way instead of the powers of conditional probability. WNBC softens independency assumptions by taking weight of each attribute into consideration. In 2006, Yager extended naive Bayes classifier in a different manner by using ordered weighted averaging (OWA) operators. He showed that the extended naive Bayes classifier is a special case of OWA operators:

$$\mathrm{ENBC}(C^*) = \arg\max_{C_i} \left\{ p(C_i) \sum_{j=1}^{n} w_j \prod_{j=1}^{k} p(X_j|C_i) \right\} \tag{3.6}$$

where w_j is the weight of the j-th attribute [26]. In this study, we use only weights without dealing with the independency assumption. We expect that weights are sufficient for assigning or identifying essential attributes for classification since those weights take active roles when it will have a great impact, in which case it is close to 1 and a small impact in which case it is close to 0.

3.3 Fuzzy Bayes Classification

In this section, recent studies about fuzzy bayes classification and its importance are explained. It is observed that there has been considerable research on fuzzy Bayesian classification. Nevertheless we note that mainly rule-based methods or some classification methods, using expert knowledge, are applied in order to make classification. An important part of our goal is to construct fuzzy membership functions without consulting an expert. For this reason, membership functions are constructed by using FCM clustering in which a learned Mahalanobis distance is used instead of other known types of distances. Thus, FCM clustering is also described briefly in this section. However, distance metric learning will be explained later in the proposed method, in chapter 5.

Even though NBC operates under the strong naive assumptions of independency, problems may emerge when imperfect and imprecise data are encountered because Bayesian approaches basically use probability theory. Another important problem which is observed is that, in NBC attributes or variables may have discrete domains whereas in real life most of the variables are continuous. Therefore, using such type of classification, one cannot handle all kinds of uncertainties in data classification. To overcome such problems several approaches have been suggested in the literature. The main approach and the one that is mostly used is Fuzzy Set theory, which was proposed by Zadeh in 1965 [27].

Foundation of fuzzy logic is emerged against the binary system of Aristotle logic and it tries to identify at which ratio (degree) events take place by assigning membership degrees to events. Fuzziness occurs when information is not clear. The word fuzzy which was first proposed by Zadeh in 1962 means vagueness or uncertainty. In fuzzy logic systems, fuzzy sets, which have been introduced by Zadeh in 1965, are used for analysis as an extension of classical sets. Fuzzy set theory has been developed to simplify the complexity of real life problems where human judgment is at the forefront and to obtain more effective results. Fuzzy set theory, in conjunction with helping a decision-maker to make the best decision under known constraints, makes it possible to produce models with new alternatives, taking the human factor into consideration. On the other hand,

when uncertainty in the models can be expressed by linguistic variables, it also allows those variables to be used in a mathematical expression [16].

In 2002, Störr proposed a fuzzy naive Bayes classifier (FNBC) combining the Naive Bayes Classifier and fuzzy theory without loss of information

$$FNBC(c^*) = \arg\max_{c_i \in C} \left\{ \sum_{x_1 \in X_1} p(x_1|c_i)\mu_{x_1} ... \sum_{x_n \in X_n} p(x_n|c_i)\mu_{x_n} \right\} \tag{3.7}$$

where $\mu_{x_i} \in [0, 1]$ is a membership function of $x_i \in X_i$ [28]. The general idea here is to obtain posterior information by using the likelihood function (prior information). Then the classification procedure is applied. However, in both approaches probability theory is used. Here, we note that the membership functions of fuzzy numbers already contain information which is obtained from probabilities. Thus, in this study a new classifier that includes only membership functions instead of probabilities is proposed. It is expected that the new classifier will perform at least as good as classical classifiers.

In the same year, Tang et. al. studied the classification problem based on clustering [29]. In their study, instances are classified in accordance with fuzzy naive Bayesian classifier based on fuzzy clustering. It is noted that in NBC all attributes are assumed to be nominal meaning that they have finite number of records but in large data sets attributes may take continuous records which may result in complexity. To cope with this complexity, the domain of continuous attributes is partitioned. They proposed a new fuzzy Bayesian classifier after applying unsupervised FCM clustering algorithm.

In this thesis, a new Fuzzy Bayesian Classifer constructed only with conditional membership functions, is proposed. To construct the conditional membership functions, an approach similiar to the approach in [29] is used. However, there exist important differences between our approach and the approach used in [29]. The first important difference is that in the FCM algorithm the Euclidean distance is used whereas we use learned Mahalanobis distance in order to obtain cluster centers. The second important difference is that they define posterior knowledge in terms of probability, i.e., membership functions are turned into

probability. However, after evaluating maximum membership degree we again use memberhsip functions in order to obtain posterior knowledge.

Chapter 4

Proposed Method

Over the past few years, research in the literature has shown that distance metrics and their use in various algorithms such as clustering or classification is important for high performance especially in visual pattern recognition, face verifications, biological data classifications and others. Choice of an appropriate distance function to be used in the learning algorithm is important in order to obtain valid results and high performance [1], [3], [31], [5]. Two of the most commonly used metric distances are the Euclidean and Mahalanobis distances which are special cases of Bregman divergences. It is known that k-Means Clustering algorithms will work only if the distance function is a Bregman divergence [4]. As a consequence the Bregman divergences make it possible to apply the k-Means clustering algorithm with different dissimilarity measures. This means that the researcher can try different potentially suitable measures before deciding of the final choice of the measure to be used.

Although Euclidean distance is used in many applications because of its simplicity, it may not reflect the real distance in some applications. A generalization of the Euclidean distance is the Mahalanobis distance. Instead of using predefined distance functions on the basis of some prior information, a different approach is to learn distance functions from available information. Therefore, in this study, one of our aims is to learn an appropriate Mahalanobis distance from the data set. It is expected that this will be important in order to achieve high performance

algorithms to be used in clustering. Xiang et. al [1] used must-link and cannot-link information for the purpose of learning a Mahalanobis distance. Motivated by this approach we investigate how to obtain such side information by using rough set theory. The application of rough sets is expected to give better results because the main goal of rough set theory is to draw inferences from data objectively without referring to experts or without using subjective prior information [2].

In the literature, the problem of learning distance functions is defined as a pairwise constraint problem where pairs of data points are specified as similar or dissimilar. Such pairwise constraints are used to learn a distance metric. Xiang et. al used pairwise constraints in the form of must-link and cannot-links to learn a Mahalanobis distance metric. A must-link identifies pairs of data points that must be in the same class. However, a cannot-link identifies pairs of data points that must be in different classes. This motivated us to consider these links in terms of indiscernibility relations, indirectly related to rough set theory. In other words, those links will be constructed based on similarities or dissimilarities, which are given implicitly in the data set by the features or attributes of the examples.

The basic idea is to identify instances that are considered to be indistinguishable as must-links. Similarly, sets of distances that are considered to be definitely distinguishable are identified as cannot-links. After identifying these links, the same steps as in the article of Xieng et. al [1] are applied to learn a Mahalanobis distance that will be used in Fuzzy C-Means (FCM) clustering algorithm in order to find cluster centers which will be used in the proposed classification algorithm.

In this thesis, we use cluster centers in order to obtain fuzzy membership functions and use these fuzzy membership functions in a new FBC method. It is observed that there has been considerable research on fuzzy Bayes classification. Nevertheless we note that mainly rule-based methods [34], [35] or methods that use expert information are applied in classification problems. An important part of our goal here is to construct fuzzy membership functions without consulting an expert since membership functions or prior probabilities are generally designed

with the help of human expert. Due to the fact that knowledge of human experts is not comprehensive which means that they are usually specialized on a specific field of study, prior information may actually not be reliable for using in posterior information. In other words, human experts may not be objective just because of their limited knowledge. For that matter, we will be focusing on the solution of a problem that can arise from experts in order to make accurate classification. Namely, new examples will be classified such that there will be no need for consulting an expert. As a result accuracy in classification is expected to be better. Another important aspect of learning algorithms is the generalization ability. By using this ability one can generalize accuracy of classification for testing set. In chapter 5, generalization performance of two data sets, Iris and Seed, are given [39].

Most of the classification problems are based on probability theory. One of the problems when using probability theory is that in a typical classification problem we only have one data set to be used. If the quality of the sample is bad the results may also be inaccurate. One possible alternative approach is to use fuzzy set theory. In particular, the proposed classifier basically uses membership functions obtained directly from the data set. Here, we note that the membership functions of fuzzy numbers already contain information which may also be obtained from probabilities. The proposed classifier is defined as:

$$FBC(x_0) = \arg\max_{1\leq j\leq t} \{\mu(\, c_j |\, x_0)\}\,, \tag{4.1}$$

where

$$\mu(\, c_j |\, x_0) = \frac{\mu_l(c_j)}{\max\{\mu_1(x_0), ..., \mu_l(x_0), ..., \mu_k(x_0)\}}$$

where

$l = \arg\max_{1\leq i\leq k} \{\mu_i(x_0)\}$, and $\max_i \{\mu_i(x_0)\} > 0$.

In this formula, $\mu_i(x_0)$ denotes the membership degree of a new example x_0 belonging to cluster i such that $\mu_i(x_0) = \sum_a w_a^* \mu_{i,a}(f(x_0, a))$ where $\mu_{i,a}$ denotes

the component of the membership function for the ith cluster corresponding to attribute $a \in C$, and w_a^* is the weight of attribute a such that $\sum_a w_a^* = 1$. The weights w_a are determined using formula 5.1, given in chapter 5. However, in this step, clusters are taken into consideration instead of classes. Let l denote the index of the cluster to which x_0 belongs, that is

$$l = \underset{1 \leq i \leq k}{\operatorname{argmax}} \{\mu_i(x_0)\}$$

Using the fuzzy membership function of the chosen cluster the memberships of the class centers (c_1, c_2, ..., c_q) are evaluated in order to determine the class. Hence,

$$c_0 = \underset{1 \leq j \leq q}{\operatorname{argmax}} \{\mu_l(c_j)\}.$$

will be the class assigned to the new example x_0.

In order to explain our approach explicitly, let us give an example. Table 4.1 is a sample data set containing 20 examples with 3 attributes one of which is the class attribute. This sample data set is chosen from a well-known data set, the Fisher Iris data set. In Figure 4.1, one can see that optimal fuzzy partition is done by applying unsupervised FCM clustering. The chosen data set is seperated into 3 clusters. It is also shown which example belongs to which class in Figure 4.1, by using class label information.

Instances	a_1	a_2	C_k
$\mathbf{x}_1$	$5,4$	$1,7$	1
$\mathbf{x}_2$	5	$1,4$	1
$\mathbf{x}_3$	$4,5$	$1,3$	1
$\mathbf{x}_4$	$4,7$	$1,6$	1
$\mathbf{x}_5$	$4,7$	$1,3$	1
$\mathbf{x}_6$	$5,4$	$1,5$	1
$\mathbf{x}_7$	$5,8$	$1,2$	1
$\mathbf{x}_8$	$5,1$	$1,5$	1
$\mathbf{x}_9$	$5,2$	$1,5$	1
$\mathbf{x}_{11}$	$5,9$	$4,2$	2
$\mathbf{x}_{12}$	6	4	2
$\mathbf{x}_{13}$	$6,2$	$4,5$	2
$\mathbf{x}_{14}$	5	$3,5$	2
$\mathbf{x}_{15}$	$6,7$	5	2
$\mathbf{x}_{16}$	$6,1$	$4,7$	2
$\mathbf{x}_{17}$	$5,9$	$4,8$	2
$\mathbf{x}_{18}$	$6,1$	$4,6$	2
$\mathbf{x}_{19}$	$6,2$	$4,3$	2
$\mathbf{x}_{20}$	$5,7$	$4,5$	2

Table 4.1: A subset of the Fisher Iris data set

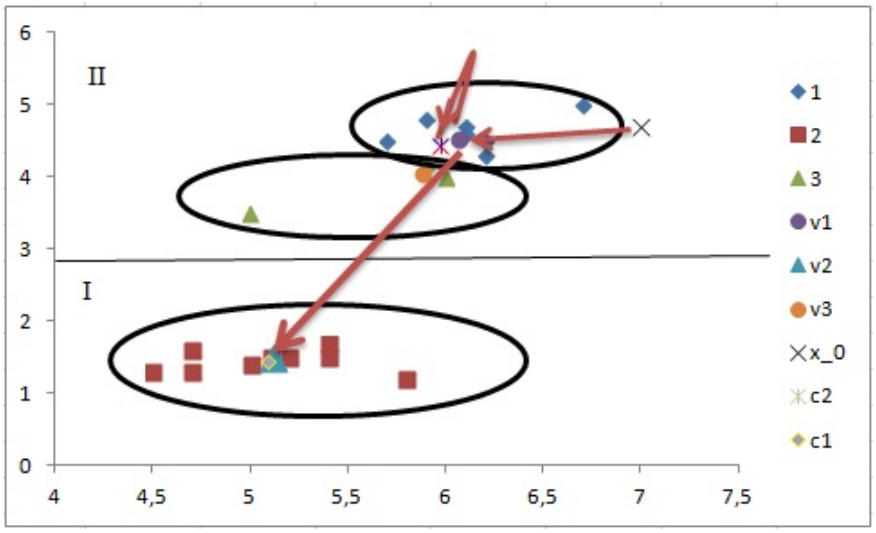

Figure 4.1: Clusters of the chosen Iris sample data set

The proposed method can be explained by analogy to gravitational forces of planets acting on nearby objects. Let x denote a new example to be classified. We first try to find the planet from which the new object is attracted the most by its cluster center; that is, the center of clusters are used as gravitational centers of planets for assigning a new object. Then, we check the distance of this planet in order to determine its class. Namely, the classification is completed by checking the location of the chosen gravitational center with respect to the class centers (See Figure 4.1). It is seen that, in the given example, a new example has a maximum membership degree to the first cluster. Accordingly, the membership degrees of each class center are evaluated and it is observed that the second class center has maximum value of belongingness, which means that the new example is assigned to the second class. (See Figure 4.2).

Bayes theory is a kind of probability theory providing a mathematical framework for making inference with probabilities and Bayes theorem is a statement in conditional probabilities such that prior probabilities are mapped into posterior probabilities by using class label information or outcome of classification events. Generally, prior probabilities are obtained by frequencies of attributes or knowledge of an expert which may not yield high accuracy in classification algorithms. However, prior probabilities may be generated objectively without consulting to

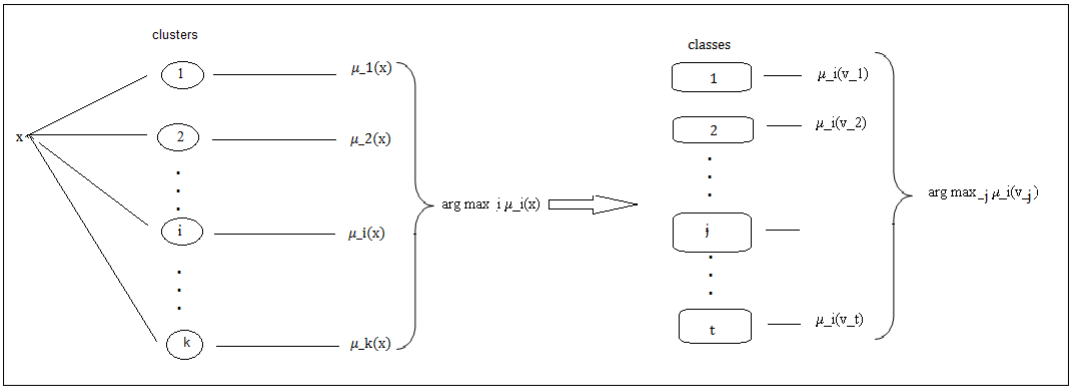

Figure 4.2: The process for classification of a new example

an expert. In this thesis, it is shown that this prior knowledge can be derived from the data set. Then, posterior information is obtained based on this prior knowledge. To sum up, the logic behind Bayes theorem is to get posterior knowledge by using prior knowledge. In the proposed method one can see that Bayesian logic is used implicitly such that in order to classify a new example class label information is used. To be more precise, in the proposed method partition of samples into clusters are achieved as in the Naive Bayes Classifier (see Figure 4.3). Then, the basic principle as in Bayes theorem is applied in order to find conditional membership functions. More formally, conditional probabilities and conditional membership functions are defined as

$$p(C_j|\mathbf{x}) = \frac{p(\mathbf{x}|C_j)p(C_j)}{\sum_i p(\mathbf{x}|C_i)p(C_i)}$$

and

$$\mu(C_j|\mathbf{x}) = \frac{\mu_l(C_j)}{\max_i \mu_i(\mathbf{x})}, l = \arg\max_{1\leq i\leq k} \{\mu_i(x)\}$$

respectively.

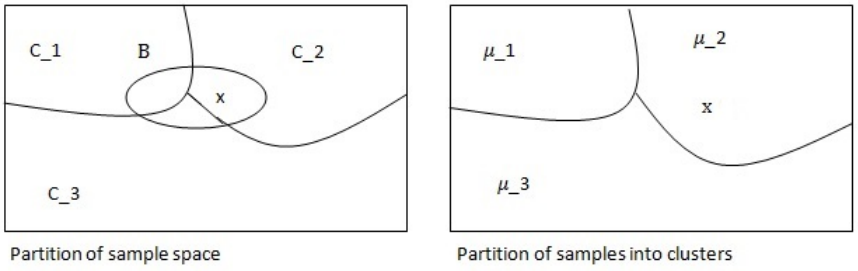

Figure 4.3: The partition of the sample space within probabilities and fuzzy approaches

The main steps of the proposed classifier are as follows (see Figure 4.2):

A. Obtaining Clusters

1. Determine the sets for must-links and cannot-links using similarities.
2. Use Xiang et al.'s (2008) algorithm to find optimum matrix W^* for $A = W^*(W^*)^T$ (see Table 4.2) to be used as a Mahalanobis distance

$$d_A(x, y) = \sqrt{(x - y)^T A(x - y)}$$

3. Apply FCM algorithm with $d_A(x, y)$ to find optimal k clusters.
4. In a similar way with [29], construct membership functions $\mu_1(x), \mu_2(x), ..., \mu_k(x)$ based on cluster centers for each cluster.

B. Applying proposed Fuzzy Bayes Classifier with constructed membership functions.

1. Find cluster with highest membership for a new example x_0.
2. By evaluating memberships of class centers determine the class.

It is seen that the new proposed classifier consists of two main parts. In the first part, unsupervised fuzzy clustering is applied and optimal fuzzy partition is found. In the design of unsupervised fuzzy clustering, a positively definite matrix A is learned from data by using pairwise constraints; must-links and cannot-links. Second, FCM algorithm is applied in order to find cluster centers $v_i, i = 1, 2, ..., k$.

Third, membership functions of each attribute for each cluster are constructed by using cluster centers that are found (achieved) in the third step of the first part [29]. In the last step of the first part, weights of each attributes are determined. We note here that the first part is tested and run by R programming [40]. In the second part, a new FBC is applied to new examples, using fuzzy membership functions obtained in the first part.

In the following, we will explain the first part in more detail. One of the main steps in the new FBC is to construct fuzzy membership functions. In order to achieve this, must-links and cannot-links have to be determined. A must-link identifies pairs of data points that must be in different classes. For various other distance learning algorithms one may refer to [1]. Motivated by this study, such links were considered in terms of indiscernibility relations which is also related to rough set theory. We note here that rough set theory is used to draw inferences objectively from data without referring to an expert.

4.1 Obtaining Clusters

4.1.1 Step 1. Obtaining must-links and cannot-links sets

In the first step, we focus on pairwise constraints in the form of must-links and cannot-links to be used in the algorithm in order to learn a Mahalanobis distance. The basic idea in this step is to minimize dissimilarity values between point pairs (examples) which are determined as must-links and maximize dissimilarity value between those point pairs which should be cannot-links. If applied appropriately, this step will also help to overcome problems such as poor quality of data set and inconsistencies in the data set. In the following, we give some basic definitions used for constructing those must-link and cannot-link sets.

Since in clustering and classification problems, objects/examples are described by attributes we will use the notion of a *data table* to describe a general data set. A data table is also referred to as an information system and consists of a 4-tuple

$\langle U, A, V, f\rangle$, where U is a finite set of objects and $A = \{a_1, a_2, \ldots, a_m\}$ is a finite set of attributes. The domain of an attribute $a \in A$ is denoted by V_a and $V = \bigcup_{a\in A} V_a$. The function f is a total function such that $f(x, a) \in V_a$ for each $a \in A$, $x \in U$, and it is called an information function. If the set of attributes A is divided into *condition* attributes ($C \neq \phi$) and *decision* attributes ($D \neq \phi$), then the data table is called a *decision table* [3].

An important step in the proposed method is the construction of the must-link and cannot-link sets. This will be achieved by using similarities between examples of the data set. The following definitions will be used to obtain the sets of must-links and cannot-links.

Definition. [2] The similarity value between two examples $x_i, x_j \in U$ with respect to attribute a_q is defined as

$$\mathrm{sim}_{a_q}(x_i, x_j) = 1 - \frac{|f(x_i, a_q) - f(x_j, a_q)|}{\max(a_q) - \min(a_q)}, \tag{4.2}$$

if a_q is a numerical attribute and as

$$\mathrm{sim}_{a_q}(x_i, x_j) = \begin{cases} \frac{1}{\mathrm{card}\, a_q}, & \text{if} f(x_i, a_q) \neq f(x_j, a_q) \\ 1, & \text{otherwise} \end{cases}$$

if a_q is a nominal attribute.

We note that, in general, distance functions are used in order to compute similarity. However, in our method, the similarity is based on differences between attribute values of attribute a_q, that is, attribute values are directly used in order to compute similarity.

Definition. [2] The similarity value between two examples $x_i, x_j \in U$ with respect to an attribute set $B \subseteq A$ is defined as

$$\mathrm{sim}_B(x_i, x_j) = \sum_{a\in B} w_a \mathrm{sim}_a(x_i, x_j) \tag{4.3}$$

where the weights w_a correspond to the attribute $a \in B$.

Remark. Note that, in order to determine the weights the mutual entropy values are generally used. However, in our method we will use a different approach. This approach is almost a new approach that is not used before in the literature. The weights are determined by using class labels as follows.

Suppose that there are t classes in the data set. For $a \in C$, let

$$A_i(a) = \{x \in U \,|\min\left(C_i(a)\right) \leq f(x,a) \leq \max\left(C_i(a)\right)\}$$

where $C_i(a)$ is the set of values for attribute a belonging to class i, $1 \leq i \leq t$. Denoting by

$$B_j(a) = A_j(a) \backslash \bigcup_{\substack{i=1 \\ (i \neq j)}}^{t} A_i(a),$$

the weights w_a are defined as

$$w_a = \frac{\sum_{i=1}^{t} s\left(B_i(a)\right)}{s(U)} \tag{4.4}$$

The number of examples, here, is counted such that the examples having common values are eliminated in this counting. However, weights are normalized in order to see the impact of that attribute with respect to each class. Therefore;

$$w_a^* = \frac{w_a}{\sum_a w_a} \tag{4.5}$$

For instance, for the same sample data set given before in Table 4.1, the weights of each attribute are $w_1^* = 11/30$ and $w_2^* = 19/30$. See figure 4.3.

Let's consider another example. We have chosen another sample data set consisting of 4 attributes with 2 classes from Fisher Iris data set and by the same logic given above weights of each attribute are calculated as $0.15, 0.19, 0.33, 0.33$ (See Fig. 4.4).

Definition. [2] The indiscernibility relation IR_B with respect to an attribute set

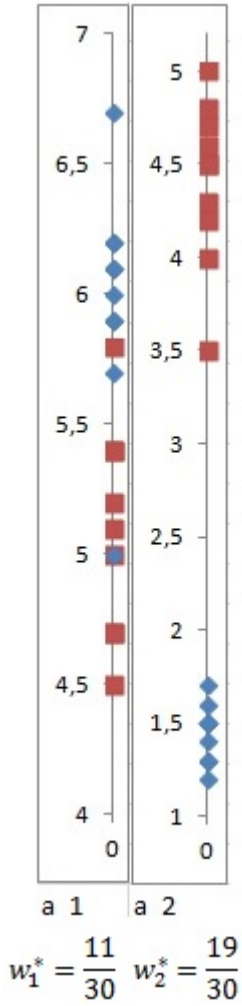

Figure 4.4: Weights of two attributes for data set in Table 4.1

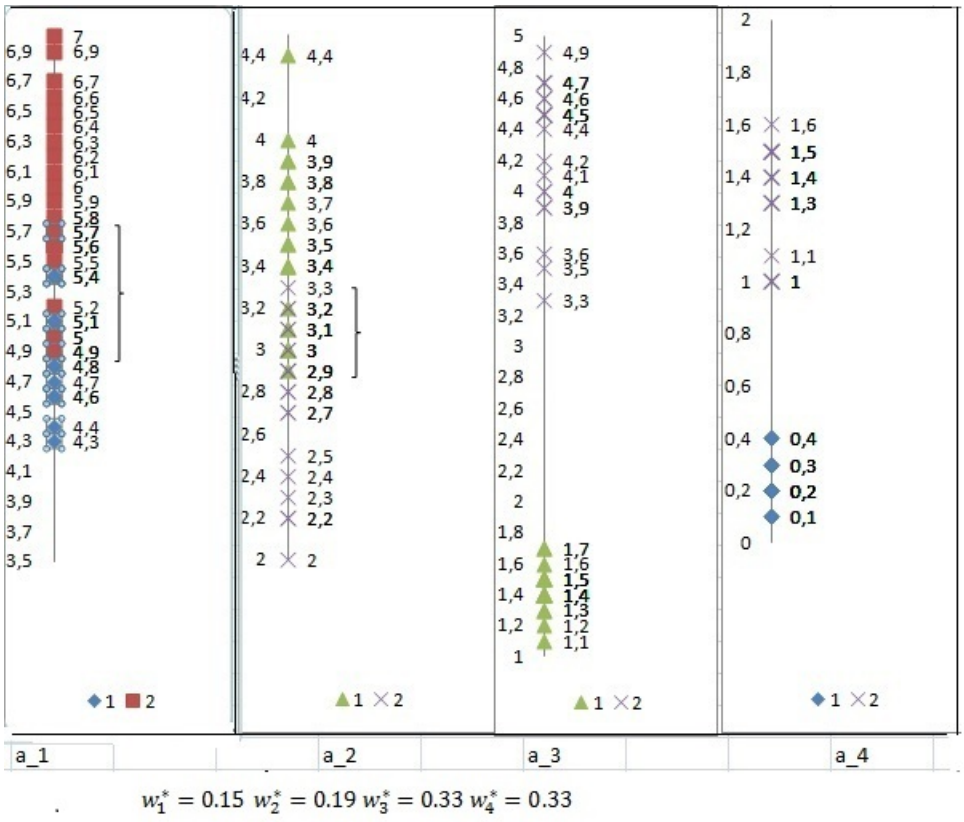

Figure 4.5: Weights of 4 attributes of Iris data set

$B \subseteq U$ is defined as

$$IR_B = \{(x_i, x_j) \in U \times U : \text{sim}_B(x_i, x_j) \geq th_B\}$$

where th_B is the threshold value for the similarity relation.

Using these definitions, elements similar to an element $x_i \in U$ can be defined as:

$$S_B(x_i) = \{x_j \in U : \text{sim}_B(x_i, x_j) \in IR_B\}$$

Based on these definitions, for a given data set, a must-link set is defined as:

$$S = \{(x_i, x_j) | \text{sim}_B(x_i, x_j) \geq th_B\},$$

and a cannot-link set is defined as:

$$D = \{(x_i, x_j) | \text{sim}_B(x_i, x_j) \leq th_B\},$$

where th_B is the threshold value for the similarity relation.
As can be seen from these definitions, S is a set which contains points which are considered to be definitely in a same class whereas D is a set which contains points which are considered definitely to be in different classes. To illustrate this idea, we used the sample data set before, obtained from the Fisher Iris data set. In Figure 4.5 one can see points inside circles indicating some of the elements belonging to S whereas points connected with lines indicate some elements belonging to D.

4.1.2 Step 2. Learning Optimal Mahalanobis Distance

In the second step the algorithm given in Xiang et al.'s study [1] is used in order to find an optimum matrix W^* for $A = W^*(W^*)^T$ to be used as a Mahalanobis distance. For this purpose Xiang et al. [1] introduced a transformation such that $y = W^T x$ where $W \in \mathbb{R}^{n \times d}$, with $d \leq n$. Based on this transformation, the sum

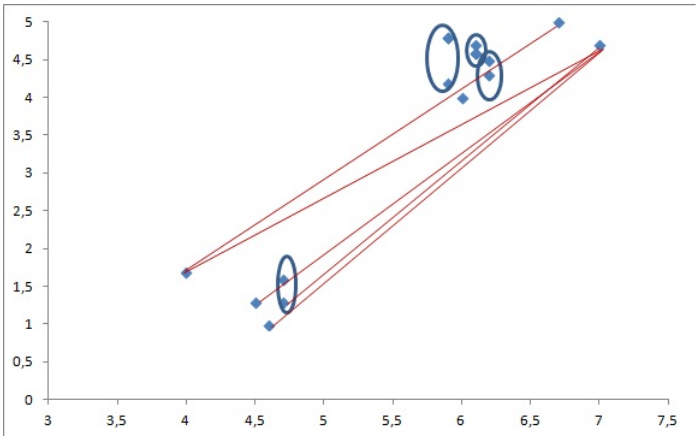

Figure 4.6: Some similarity-dissimilarity pairs

of the squared distances of the point pairs in S is defined as:

$$d_w = \sum_{(x_i,x_j)\in S} (W^T x_i - W^T x_j)^T (W^T x_i - W^T x_j) = tr(W^T S_w W),$$

where tr is the trace operator and S_w, covariance matrix of the point pairs in S is calculated as:

$$S_w = \sum_{(x_i,x_j)\in S} (x_i - x_j)(x_i - x_j)^T$$

Similarly, for the point pairs in D, we have

$$d_b = tr(W^T S_b W),$$

where S_b is the covariance matrix of the point pairs in D is calculated as:

$$S_b = \sum_{(x_i,x_j)\in D} (x_i - x_j)(x_i - x_j)^T.$$

Since d_w and d_b represent the sum of squared distances between point pairs in must-links and cannot links, respectively the optimal matrix W^* can be calculated as

$$W^* = \underset{\{W^T W = I\}}{\arg\max} \frac{tr(W^T S_b W)}{tr(W^T S_w W)},$$

where I is identity matrix and the constraint $W^TW = I$ is given in order not to have degenerate solutions. The important point here is that W cannot be a square matrix when $d < n$. In that case A is defined as follows:

$$A = \begin{cases} (W^*(W^*)^T, & \text{if } d < n \\ I, & \text{if } d = n \end{cases}$$

In order to find optimal W^* $tr(W^T S_w W)$ is used.

Theorem 4.1 *[1] Suppose that $W \in \mathbb{R}^{n\times d}$, $W^TW = I$, and $r(\leq n)$ is the rank of matrix S_w. If $d > n - r$, then $tr(W^T S_w W) > 0$. If $d \leq n - r$, then $tr(W^T S_w W)$ may be zero.*

Case1: $d > n - r$ Assume that λ^* is the optimal solution of the equation

$$\lambda^* = \max_{\{W^TW=I\}} \frac{tr(W^T S_b W)}{tr(W^T S_w W)}$$

Then

$$\max_{W^TW=I} tr(W^T(S_b - \lambda^* S_w)W) = 0$$

From above equation a new function which is a function of λ can be defined such as:

$$\eta(\lambda) = \max_{W^TW=I} tr(W^T(S_b - \lambda^* S_w)W)$$

Thus, the aim is to find a λ such that $\eta(\lambda) = 0$ holds.

In that case, $tr(W^T S_w W) > 0$, and then not only $\eta(\lambda) < 0$ implies that $\lambda > \lambda^*$ but also $\eta(\lambda) > 0$ implies that $\lambda < \lambda^*$. This shows that one can find λ with an iteration method.

In order to find the optimal value for λ^*, lower and upper bounds are determined by using the following theorem:

Theorem 4.2 *[1] Let r be the rank of S_w. If $d > n - r$ then*

$$\frac{tr(S_b)}{tr(S_w)} \leq \lambda^* \leq \frac{\sum_{i=1}^{d} \alpha_i}{\sum_{i=1}^{d} \beta_i}$$

where $\alpha_1, ..., \alpha_d$ are the first d largest eigenvalues of S_b, and $\beta_1, , \beta_d$ are the first d smallest eigenvalues of S_w.

The optimal matrix W^* is finally calculated by using the eigenvalue decomposition of $S_b - \lambda^* S_w$. For this reason, the null space of $S_b + S_w$ can first be eliminated by using the following theorem:

Theorem 4.3 *[1] W^* can be found in the orthogonal complement space of the null space of $S_b + S_w$.*

When the dimensionality (n) is greater than the number of data points (N), then the rank of $S_b + S_w$ will be smaller than N. In this case, there is no need for eigenvalue decomposition of $n \times n$ dimensionality. When $n < N$, then symmetrical indicator matrices are defined in order to facilitate computational complexities. The following notation is used as in [1].

Let X be the data matrix containing N points such that $X = [x_1, x_2, ., x_N] \in \mathbb{R}^{n \times N}$. A similar indicator matrix $L_s \in \mathbb{R}^{N \times N}$ can be introduced as follows:

$$L_s(i,j) = L_s(j,i) = \begin{cases} 1, & (x_i, x_j) \in S \\ 0, & (x_i, x_j) \notin S \end{cases}$$

where S is a must-link set.

Moreover, a symmetrical indicator matrix $L_d \in \mathbb{R}^{N \times N}$ can be introduced as follows:

$$L_d(i,j) = L_d(j,i) = \begin{cases} 1, & (x_i, x_j) \in D \\ 0, & (x_i, x_j) \notin D \end{cases}$$

where D is cannot-link set.

Now, suppose that $L_w = diag(sum(L_s)) - L_s$ and $L_b = diag(sum(L_d)) - L_d$, where sum is an N-dimensional vector which records the sum of each row of the matrix. Therefore, we can justify that $S_w = \frac{1}{2} X L_w X^T$ and $S_b = \frac{1}{2} X L_b X^T$ and those imply that $S_w + S_b = X \left(\frac{1}{2} L_w + \frac{1}{2} L_b \right) X^T$

Preprocess

P1: Eliminate the null space of $S_w + S_b$ and get a linear transformation $y = W_1^T x$ where W_1 only consists of the eigenvectors corresponding to the non-zero eigenvalues of $S_w + S_b$.

P2: Reconstruct the matrices $S_w = W_1^T S_w W_1$ and $S_b = W_1^T S_b W_1$.

Algorithm

Input $S_w, S_b \in \mathbb{R}^{n \times n}$, the lower dimensionality d, and an error ϵ.

A1: Compute the rank r of the matrix S_w

A2: If $d \leq n - r$ go to step 7

Case 1: $d > n - r$

A3: $\lambda_1 \leftarrow \frac{tr(S_b)}{tr(S_w)}$, $\lambda_2 \leftarrow \frac{\sum_{i=1}^{d} \alpha_i}{\sum_{i=1}^{d} \beta_i}$, $\lambda \leftarrow (\lambda_1 + \lambda_2)/2$.

A4: Find optimal λ value

While $\lambda_2 - \lambda_1 > \epsilon$, do

Compute $\eta(\lambda)$.

If $\eta(\lambda) > 0$ then $\lambda_1 = \lambda$; otherwise $\lambda_2 = \lambda$.

Then $\lambda = (\lambda_1 + \lambda_2)/2$.

End while.

A5: $W^* = [\mu_1, ..., \mu_d]$, where $\mu_1, ..., \mu_d$ are the d eigenvectors of $S_b - \lambda S_w$.

A6: $A = W^*(W^*)^T$; STOP

Case 2: $d \leq n - r$

A7: $W^* = Z.[v_1, ..., v_d]$, where $v_i, i = 1, 2, ..., d$ are d eigenvectors corresponding to the d largest eigenvalues of $Z^T S_b Z$ and $Z = [z_1, z_2, ..., z_{n-r}]$ are the eigenvectors corresponding to $n - r$ zero eigenvalues of S_w.

A8: $A = W_1 W^*(W^*)^T (W_1)^T$.

Table 4.2: Xiang et al.'s Algorithm for Learning Matrix A

Let $L = X^T X(\frac{1}{2}L_w + \frac{1}{2}L_b) \in \mathbb{R}^{N \times N}$. Non-zero eigenvalues of L and corresponding eigenvectors can be calculated if $N < n$.

Case 2: $d \leq n - r$ If W is the null space of S_w, then $tr(W^T S_w W) = 0$ and $(\lambda)^*$ is finite. Therefore, $tr(W^T S_w W)$ is maximized after $y = Z^T x$ transformation:

$$V^* = \arg\max_{V^T V} (V^T (Z^T S_b Z) V),$$

where $Z \in \mathbb{R}^{n \times n-r}$ is a matrix whose column vectors represent the eigenvectors corresponding to $n - r$ zero eigenvalues of S_W. $W^* = ZV^*$ is obtained after V^* is evaluated.

In Table 4.3, there are 19, 3-dimensional instances with 2 classes. By applying the algorithm given in Table 4.2 a transformation is obtained for the case $d = 2$. It is observed that the instances within the same class are gathered together. Therefore, the examples in different classes are split up very well. (See Figure 4.6).

Instances	a_1	a_2	a_3	C_k
$\mathbf{x}_1$	4	3,9	1, 7	1
$\mathbf{x}_2$	4, 6	3,6	1	1
$\mathbf{x}_3$	4, 5	2,3	1, 3	1
$\mathbf{x}_4$	4, 7	3,2	1, 6	1
$\mathbf{x}_5$	4, 7	3,2	1, 3	1
$\mathbf{x}_6$	5, 4	3,9	1, 5	1
$\mathbf{x}_7$	5, 8	4	1, 2	1
$\mathbf{x}_8$	5, 1	3,7	1, 5	1
$\mathbf{x}_9$	5, 2	3,5	1, 5	1
$\mathbf{x}_{10}$	7	3,2	4, 7	1
$\mathbf{x}_{11}$	5, 9	3	4, 2	2
$\mathbf{x}_{12}$	6	2,2	4	2
$\mathbf{x}_{13}$	6, 2	2,2	4, 5	2
$\mathbf{x}_{14}$	5	2	3, 5	2
$\mathbf{x}_{15}$	6, 7	3	5	2
$\mathbf{x}_{16}$	6, 1	2,9	4, 7	2
$\mathbf{x}_{17}$	5, 9	3,2	4, 8	2
$\mathbf{x}_{18}$	6, 1	3	4, 6	2
$\mathbf{x}_{19}$	6, 2	2,9	4, 3	2

Table 4.3: Another subset from Fisher Iris data set

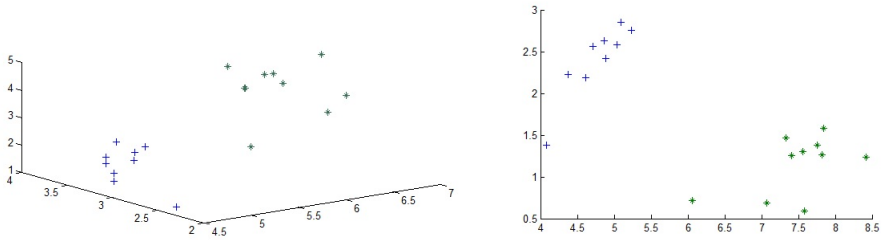

Figure 4.7: Transformation

So far constructing must-links, cannot-links and learning a positively definite matrix to be used in Mahalanobis were explained. Next, the FCM algorithm will be outlined in order to obtain the optimal cluster number and their corresponding cluster centers.

4.1.3 Optimal Fuzzy C-Means Clustering

4.1.3.1 Fuzzy C-Means Clustering

Cluster analysis has been a major research tool since the 1960's. Since then it became a well-known method that divides a training set into several subsets(clusters) which have similar objects. Untill now, many resarchers have proposed different types of clustering methods such as fuzzy clustering, conducted with respect to similarity/dissimilarity between cluster centers and data points. Zadeh, in 1965, approached similarity/dissimilarity by a function (membership function) because some objects may not belong to one cluster only. We know that membership functions take values between zero and one. Therefore, a similarity value close to one means that there exists a big difference in similarity between the sample and cluster [30]. However, it is not that easy to calculate this difference. To solve such problems, many different algorithms are proposed and coded. Since the inputs are vectors, to compute the difference, we will use norms. Note that, in this study we will use Mahalanobis and Euclidean norms.

Let X be a set of observations such that

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1q} \\ x_{21} & x_{22} & ... & x_{2q} \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ x_{p1} & x_{p2} & ... & x_{pq} \end{bmatrix}$$

and let each row be an example, $X_k, k = 1, 2, ..., p$. Suppose we will divide the data set into c clusters and let u_{ik} be the membership degree of object k in the i-th fuzzy cluster. with $0_{ik} \leq 1; \forall i, k$ and $\sum_{i=1}^{c} u_{ik} = 1, \forall k$. First of all, we have to find out optimal c and optimal level of fuzziness which is represented as m. To find out one has to use iteration methods. The most common one used is FCM that is stated as [30]

$$J(u, v) = \sum_{k=1}^{nd} \sum_{i=1}^{c} (u_{ik})^m \left\| x_k - v_i \right\|$$

where
$0 < u_{ik} \leq 1, \forall i, k$
$\sum_{i=1}^{c} u_{ik} = 1, \forall k.$

4.1.3.2 FCM Algorithm

The main objective of the FCM algorithm is to minimize the following function [33]:

$$J(u, v) = \sum_{i=1}^{n} \sum_{j=1}^{k} (u_{ij})^m \| x_i - v_j \|_A^2$$

where
$1 \leq j \leq k$, and $m \geq 1$ is the weight of fuzzy membership values,
$\|x\|_A = \sqrt{x^T A x}$ is an inner product norm,
u_{ij} is the membership degree of the i^{th} instance j^{th} cluster,

1. Initialize $U = [u_{ij}]$ matrix, $U^{(0)}$ randomly such that

$$\sum_{i=1}^{c} u_{ik}^{(0)} = 1, 1 \leq k \leq n.$$

2. Calculate cluster centers:

$$v_i^{(l)} = \frac{\sum_{k=1} n u_{ik}^{(l-1)\alpha} X_k}{\sum_{k=1} n u_{ik}^{(l-1)\alpha}}, 1 \leq i \leq c$$

3. Update U^k, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}}}$$

4. If $\left\| U^{(k+1)} - U^{(k)} \right\| \leq \epsilon$ then STOP; otherwise return to step 2.

Table 4.4: FCM Clustering Algorithm

v_j is j^th cluster center,
n is the number of data points,
k is number of cluster centers and x_i is i^{th} data.

It is important to note at this point that in this algorithm a learned Mahalanobis distance is used in the inner product. Therefore, it is expected that the error in clustering will be minimized. Table 4.4, shows the main steps for the FCM clustering algorithm [42].

It can be observed that the learned Mahalanobis distance improves the performance of the clustering algorithm. For instance, when applied to Fisher's Iris data set consisting of 150 instances, the learned Mahalanobis distance performs better than the Euclidean distance. Obtaining prior knowledge by constructing must-links and cannot-links may be an effective way for learning a positive definite matrix A from data set instead of consulting an expert, especially when it will not be possible to consult an expert. This means that a distance metric with good quality should identify irrelevant attributes and crucial attributes.

4.1.4 Constructing Membership Functions

After the FCM clustering algorithm is applied, cluster centers are determined. Then those centers are sorted in an ascending order such that $v_{1j} < v_{2j} < ... < v_{kj}$ where $v_{1j}, v_{2j}, ..., v_{kj}$ are the first components of each cluster center, namely cluster centers are are ordered with respect to their attributes [29]. In the study published in [29], one can see that they used only three different types of fuzzy membership functions which are given below:

$$\mu_{1,a_j}(x_j) = \begin{cases} 1, & \text{if } x_j \leq v_{1j} \\ \frac{v_{2j}-x_j}{v_{2j}-v_{1j}}, & \text{if } v_{1j} \leq x_j \leq v_{2j} \\ 0, & \text{if } x_j > v_{2j} \end{cases}$$

$$\mu_{t,a_j}(x_j) = \begin{cases} 0, & \text{if } x_j \leq v_{(t-1)j} \\ \frac{x_j-v_{(t-1)j}}{v_{kj}-v_{(k-1)j}}, & \text{if } v_{(t-1)j} \leq x_j \leq v_{tj} \\ \frac{v_{(t+1)j}-x_j}{v_{(t+1)j}-v_{tj}}, & \text{if } v_{tj} < x_j \leq v_{(t+1)j} \\ 0, & \text{if } x_j > v_{(t+1)j} \end{cases}$$

$$\mu_{k,a_j}(x_j) = \begin{cases} 0, & \text{if } x_j \leq v_{(k-1)j} \\ \frac{x_j-v_{(k-1)j}}{v_{kj}-v_{(k-1)j}}, & \text{if } v_{(k-1)j} \leq x_j \leq v_{kj} \\ 1, & \text{if } x_j > v_{kj} \end{cases}$$

where $1 < t < k$, and μ_{k,a_j} is the set of fuzzy partition of domain of j^{th} attribute.

As mentioned before that Naive Bayes Classifier is the most common classifier used in practice whereas it has strong independency assumptions. Although all attributes in the NBC are assumed to be nominal or discrete which have finite number of values (records) the variables may take continues values in large data sets. One approach to handle continues values is to make discretization, namely, crisp partitioning the domain of each attribute but this may lead to loss of information. Thus, in order to overcome such handicaps, fuzzy partitioning is done instead of discretization [29]. Moreover, fuzzy membership functions are constructed without consulting an expert or without using subjective prior knowledge. It is expected that this approach will increase the accuracy of clustering

or classification. By using the constructed membership functions, a conditional membership function

$$\mu(c_j|x)$$

is proposed, which is used in the proposed classifier. This expression means that given an example $\mathbf{x}$ to be classified, the membership function of each class is evaluated with respect to the cluster chosen for $\mathbf{x}$. The logic behind this idea is to apply Bayes theorem for classification. Thus, posterior information is obtained by this expression.

To sum up, unsupervised fuzzy C-Means clustering is applied with a learned Mahalanobis distance and then using those cluster centers that are obtained by the FCM algorithm, a new FBC is proposed. In this method one can see that the same principles as in Bayes theorem are applied with conditional membership functions, given the class label information. Here, we take the membership function of i^{th} cluster which gives the maximum degree for X and then compute the membership degree of each cluster center. Finally, a new example is classified according to the maximum membership degree of each class.

In the proposed method, classification is made according to maximum global preference. What if one faces the situation in which membership degrees of several classes with respect to the chosen cluster membership function has the same maximum value. In 2009, L. Peng et al., stated that distance and data mass are two important concepts that should be considered for classification. They studied data gravitation based classification using Newton's gravitation law. In fact, the logic behind their classification method is very similar to our method. However, our method is based on fuzzy clustering algorithm and pairwise constraints which are based on similarities. On the other hand, in their study, similarity between data is defined as distance and data mass. It is stated that distance is the first concept that should be considered for classification but when the distances are equal then the concept of 'data mass' is considered in order to determine the class of a new example. The same approach may be used when two or more classes have the same maximum degree. In this case the data mass or density will help to determine the class of a new example.

Chapter 5

Application

In this section, we have applied the algorithm given in Table 4.2 on two different data sets and compared them for different cases of d with different must-link and cannot-link sets. The applications to data clustering are analyzed. Moreover, the proposed method is performed on these data sets and the results are given. In section 5.1, all examples in the Fisher Iris data set are used for both training set and testing set. In section 5.2, accuracies in different classification methods are given.

The first data set is the Fisher Iris data set which is the best known data set in the literature. It contains 3 classes with 150 instances (50 in each one of three classes). Information of each attributes are given in Figure 5.1. In the training set 120 instances are included and the remaining 30 examples are used for testing the proposed Fuzzy Bayes classification. The second data set which we have worked with is the Seed data set contains 3 classes with 210 instances. In this experiment, 168 instances are used for training set and 42 instances are used for testing set. These two databases can be obtained from UCI Machine Learning Repository.

	Min	Median	Mean	Max	Min	Median	Mean	Max	Min	Median	Mean	Max
a_1	4	5	5,006	5,8	4,9	5,9	5,936	7	4,9	6,5	6,588	7,9
a_2	2,3	3,4	3,418	4,4	2	2,8	2,77	3,4	2,2	3	2,974	3,8
a_3	1	1,5	1,464	1,9	3	4,35	4,26	5,10	4,5	5,55	5,552	6,9
a_4	0,1	0,2	0,244	0,6	1	1,3	1,326	1,8	1,4	2	2,026	2,5

Table 5.1: Descriptive Statistics of Iris Data Set

5.1 Training Performance for Fisher Iris Data Set

The proposed classifier is firstly applied on the Fisher Iris data set with different values of d used in learning a Mahalanobis distance. In addition, it is also applied on the same data set with the Euclidean distance for comparison. Fisher's Iris data set consists of 150 instances with four attributes and having three classes. In the following, descriptive statistics of all attributes for each classes are summarized. When we apply FCM clustering algorithm with learned Mahalanobis distance we get the following matrix that shows cluster centers for constructed optimal number of clusters. As a first experiment we have chosen the cluster number as 3. In this experiment there are a couple of cases for d. Although we obtain minimum value for objective function used in FCM when $d = 1$ we have applied the classification procedure also for $d = 2$ and for $d = 3$ to ensure that classification accuracy is better when $d = 1$. It is also important to note that minimization of the objective function is not only sufficient for the proposed classification. Since classification is based on cluster centers, it is very crucial to get a meaningful matrix that shows cluster centers. Therefore, when $d = 2, m = 2.3$ and $k = 3$ the following matrix is obtained from the algorithm given in Table 4.4

$$v = \begin{bmatrix} 6.499376 & 2.963389 & 5.359669 & 1.9399501 \\ 6.074130 & 2.845930 & 4.502108 & 1.4529828 \\ 5.043911 & 3.398109 & 1.566182 & 0.2811095 \end{bmatrix}$$

Using this matrix, the membership functions of each attributes with respect to each cluster are constructed, as shown in the Figures 5.1, 5.2, 5.3, and 5.4:

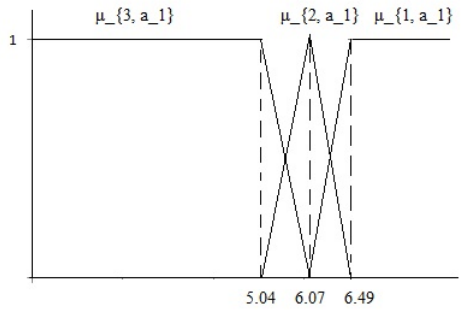

Figure 5.1:

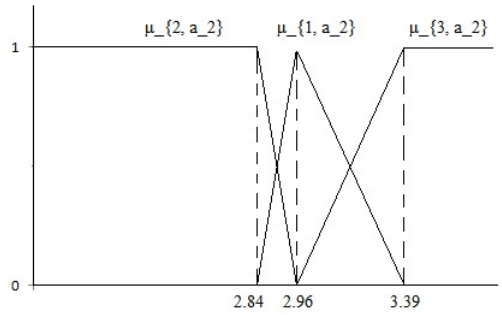

Figure 5.2:

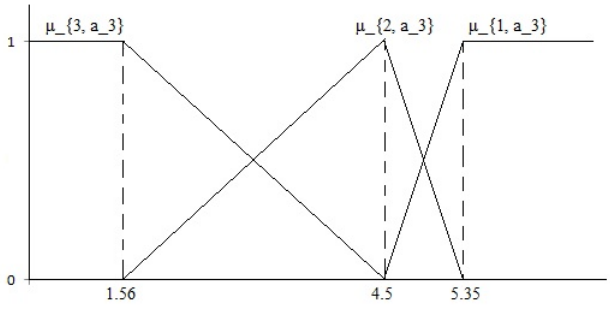

Figure 5.3:

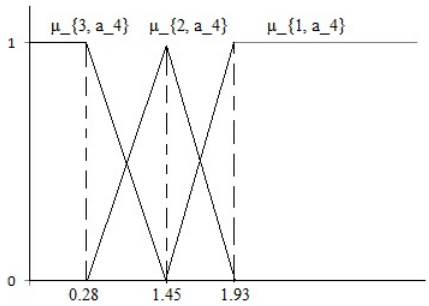

Figure 5.4:

Figures 5.1, 5.2, 5.3, and 5.4 show the membership functions of each attribute with respect to each cluster. In addition to those membership functions, weights of attributes, which are shown in figure 5.5 are calculated as $w_1 = \frac{27}{290}, w_2 = \frac{10}{290}, w_3 = \frac{141}{290}$ and $w_4 = \frac{112}{290}$.

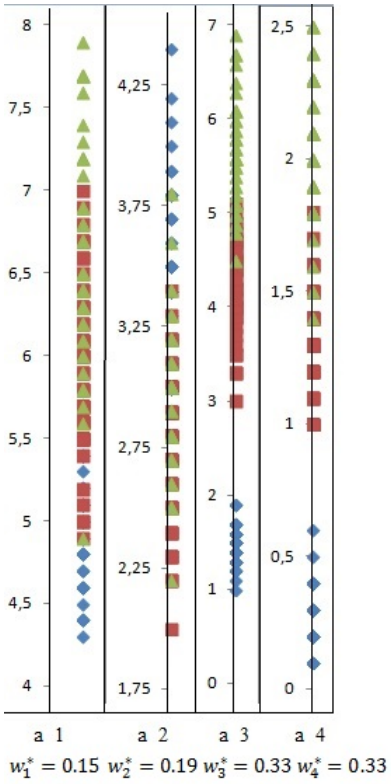

Figure 5.5: Weights

Note that the accuracy in clustering of the proposed method is better than for the Euclidean distance. Besides, the performance in classification is the same for both NBC and the proposed method (for $d = 1 \& d = 3$).

An interesting point in Table 5.2 is that increasing the number of pairs in the must-link and cannot-link sets does not provide the expected improvement in the performance of the algorithm.

Method	Dimensionality(d)	Accuracy in clustering	Accuracy in classification
Learned Mahalanobis when (must-link set S=A=20)	1	0,960	**0,960**
	2	**0,986**	0,953
	3	0,900	0,953
Learned Mahalanobis when (must-link set S=A=121)	1	0,880	0,946
	2	0,933	0,953
	3	0,933	**0,960**
Euclidean distance		0,980	**0,960**

Table 5.2:

Database	#Data	#Tr	#Ts	#Attributes	#Classes
Iris	150	120	30	4	3
Seed	210	168	42	7	3

Table 5.3:

5.2 Generalization Performance

We have applied the proposed approach to two real datasets obtained from repository of Machine Learning dataset, namely, Fisher Iris dataset and Seeds database which consist of only numerical attributes. A brief description of the data sets is given in Table 5.3, where #Data denotes the number of examples in the data set, #Tr denotes the number of training instances, #Ts denotes the number of testing instances.

The proposed classifier is compared with the NBC classifier. In addition, to see the effect of distance learning the same method is applied to the Euclidean distance. Accuracy rate of classification is computed as:

$$\frac{\text{number of correctly classified instances}}{\text{number of classified instances}} 100.$$

The results in Table 5.4 show that for the Iris testing data set, the proposed method with learned Mahalanobis distance outperforms the same method with Euclidean distance. However, in that case performances of NBC and our classifier seem to be the same. When we look at the Seeds data set, we see that generalization performance of the proposed method is better compared to NBC.

Data set	Accuracy rate (%)	
Iris	d=1	43.33
	d=2	**100**
	d=3	93.3
	Euclidean	80
	NBC	**100**
Seeds	d=1	90.48
	d=2	**92.86**
	d=3	90.48
	d=4	90.48
	d=5	90.48
	d=6	**92.86**
	Euclidean	66.66
	NBC	**90.48**

Table 5.4:

Chapter 6

Conclusion and Further Studies

In this study a new classification method which is called Fuzzy Bayesian Classifier is proposed. The proposed method is applied to Fisher's Iris data and Seed data with the Euclidean and learned Mahalanobis distances. These data sets are chosen since classes for these data sets are known. The FCM clustering algorithm is applied in order to achieve an optimal fuzzy partition. Based on this partition fuzzy membership functions for each attribute is constructed, which are then used in classification. Since in the proposed FBC there are several parameters to be considered, such as number of clusters and the reduction parameter d, several cases were examined. The results show that changing distance from Euclidean distance to Mahalanobis distance increases the classification success rate. It is also seen that, for generalization, the effect of distance becomes more important. See Table 5.4. As a consequence, the results for the considered data sets show that the new FBC is an effective and efficient method for classification. The performance of the proposed FBC needs to be investigated further with respect to different parameters such as dimension size and number of classes. We note here that a well designed simulation study will be needed in order to analyze the performance of the proposed method. A further direction for research is to extend our implementation for both linguistic and numerical variables.

BIBLIOGRAPHY

[1] Xiang, S., Nie, F. and Zhang, C. (2008). "Learning Mahalanobis distance metric for data clustering and classification", *Pattern Recognition*, 41, 3600-3612.

[2] Chen, C.B., and Wang, L.Y. (2006). "Rough set-based clustering with refinement using Shannon's entropy theory", *Computers and Mathematics with Applications*, 56, 1563-1576.

[3] Greco, S., Matarazzo, B., and Slowinski R. (2001). "Rough sets theory for multicriteria analysis", *European Journal of Operational Research*, 129, 1-47.

[4] Banerjee, A., Merugu, S., Dhillon, I.S. and Ghosh, J. (2005). "Clustering with Bregman divergeneces", *Journal of Machine Learning Research*, 6, 1705-1749.

[5] Arslan, G. (2011). "The Use of Bregman Divergences in k-Means Clustering", *Fuzzy Systems Organization*, FUZZYSS'11 Proceedings, 13-16.

[6] Nielsen, F., Boissonnat, J.D., and Nock, R. (2007). "Bregman Diagrams: Properties, Algorithms and Applications", ISSN 0249-6399 ISRN INRIA/RR–6154–FR+ENG, 1-48.

[7] Xu, P., Brock, G.N. and Parrish, R.S. (2009). "Modified linear discriminant analysis approaches for classification of high dimensional microarray data", *Computational Statistic and Data Analysis*, 53, No: 5, 1674-1687.

[8] Giacometti, A., Miyaneh, E.K., Marcel, P. and Soulet, A. (2008). "A Generic Framework for Rule Based Classification", Proceedings of LeGo, 37-54.

[9] Cover, T.M. and Hart, P. (1967). "Nearest Neighbour Pattern Classification", *IEEE Transactions on Information Theory*, 13, 21-27.

[10] Bishop, M. (1996). *Neural Networks for Pattern Recognition*, Oxford University Press, New York.

[11] Quinlan, J.R. (1986). "Induction of Decision Trees", *Machine Learning*, 1, 81-106.

[12] Hsu, C.W., Chang, C.C. and Lin, C.J. (2010). "A Practical Guide to Support Vector Classification", In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1-16.

[13] Friedman, N., Geiger, D., and Goldszmidt, M. (1997). "Bayesian Network Classifiers", *Machine Learning*, 29, 131-163.

[14] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgen Kaufmman, San Francisco, CA, 221-226.

[15] Duda, R.O., Hart, P.E. (1973). *Pattern Classification and Scene Analysis*, Wiley, New York.

[16] Tutuncu, G.Y. (2003). *Rassal Talepli, Tek Donemli Envanter Sistemlerinin Bulank Kume Teorisi Kullanlarak Modellenmesi*, Master Thesis, Baskent University, Ankara, Izmir.

[17] Clark, P., and Niblett, T. (1989). "The CN2 induction algorithm", *Machine Learning*, 3, No: 4, 261-284.

[18] Langley, P., Iba, W., and Thompson, K. (1992). "An Analysis of Bayesian Classifiers", In Proceedings of the national conference on artificial intelligence. JOHN WILEY and SONS LTD, 1-15.

[19] Kononenko, I. (1991). "Semi-Naive Bayesian Classifier", In Proceedings of the 6th European Working Session on Learning, 206-219.

[20] Friedman, N. and Goldszmidt, M. (1996a). "Building Classifiers Using Bayesian Networks", *In Proceedings of the National Conference on Artificial Intelligence* , Menlo Park, CA: AAAI Press, 1277-1284.

[21] Haouari, B., Amor, N.B., Elouedi, Z., and Mellouli, K. (2009). "Naive Possibilistic Network Classifiers", *Fuzzy Sets and Systems*, 160, 3224-3238.

[22] Kuncheva L.I and Hoare, Z.S.J. (2008). Error Dependency Relationships for the naive Bayes Classier with Binary Features", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30, No: 4, 735740.

[23] Langley, P. and Sage, S. (1994). Augmented naive Bayesian classifiers", *In Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence UAI-94*, 399-406.

[24] Sahami, M. (1996). Learning Limited dependence Bayesian classifiers", *In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining KDD'96*, Portland, OR, USA, 335-338.

[25] Harry, Z. and Sheng, S. (2004). Learning Weighted Naive Bayes with Accurate Ranking", *In Fourth IEEE International Conference on Data Mining ICDM'04*, 567-570.

[26] Yager, R.R. (2006). An extension of the naive Bayesian classifier", *Information Sciences*, 176, 577-588.

[27] Zadeh, L.A. (1965). "Fuzzy Set", *Information and Control*, 8, 338-353.

[28] Storr, H.P. (2002). "A compact fuzzy extension of the Naive Bayes Classifier based on Fuzzy Clustering", *IEEE International Conference on Systems, Man and Cybernetics*, 1-6.

[29] Tang, Y., Pan, W., Li, H., and Xu, Y. (2002). "Fuzzy Naive Bayes extension of the Naive Bayes Classifier based on Fuzzy Clustering", *IEEE International Conference on Systems, Man and Cybernetics*, 5, No: 6, 1-6.

[30] Bezdek, J.C., Ehrlich, R., and Full, W. (1984). "FCM: The Fuzzy C-Means Clustering Algorithm", *Computers and Geosciences*, 10, No: 2-3, 191-203.

[31] Khemchandani, R., Jayadeva, and Chandra, S. (2010). "Learning the optimal kernel for Fisher discriminant analysis via second order cone programming" *European J. of Operational Research*, 203, 692-697.

[32] Li, J. and Lu, B.L. (2009). "An adaptive Euclidean distance", *Pattern Recognition*, 42, 349-357.

[33] Pal, N.R., Bezdek, J.C., and Hathaway, R.J. (1996). "Sequential Competitive Learning and the Fuzzy c-Means Clustering Algorithms", *Neural Networks*, 9, No: 5, 349-357.

[34] Wu, T., Chen, S. (1999). "A new Method for Constructing Membership Functions and Fuzzy Rules from Training Examples" *IEEE Transactions on Systems, Man, and Cybernetics*, 29, 25-40.

[35] Hong, T.P. and Lee, C.Y. (1996). "Induction of Fuzzy Rules and membership Functions from Training Examples", *Fuzzy Sets Syst.*, 84, No: 1, 33-47.

[36] Peng, L., Yang, B., Chen, Y., and Abraham, A. (2009). "Data Gravitation based Classification" *Information Sciencies*, 179, 809-819.

[37] Wu, T., and Chen, S. (1999). "A New Method for Constructing Membership Functions and Fuzzy Rules from Training Examples" *IEEE Transactions on Systems, Man, and Cybernetics*, 29, No: 1, 1-47.

[38] Hasuike, T., Katagiri, H., Tsubaki, H., and Tsuda, H. (2012). "Constructing Membership Function Based on Fuzzy Shannon Entropy and Human's Interval Estimation" *Fuzzy Systems, IEEE World Congress on Computational Intelligence.*

[39] Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[40] R Development Core Team. (2008)." R: A Language and Environment for Statistical Computing" R Foundation for Statistical Computing, ISBN 3-900051-07-0, Vienna, Austria, URL: http://www.R-project.org.

[41] Turksen, I.B. (2008). "Fuzzy Functions with LSE" *Applied Soft Computing*, 8, 1178-1188.

[42] Fuzzy C-Means Clustering, "http://home.deib.polimi.it/matteucc/Clustering/tutorial-html/cmeans.html " (access date: 03.04.2013).

VITA

Necla Kayaalp was born in İzmir, Turkey, on December 16, 1987, the daughter of Cindi, and Güler Kayaalp. She completed her primary and high school education in Izmir. She began her B.S degree in 2005 in İzmir University of Economics. After receiving her B.S degree, she continued her academic career at the same university with G. Arslan. In 2010, she began to work as a research assistant in the Department of Mathematics, in İzmir University of Economics.