Cairo University

Faculty of Computer Science and Information Technology

Data Mining for Medical Informatics

Thesis Submitted to Department of Computer Science in Partial Fulfilment of the Requirements for Obtaining the Degree of

Doctor of Philosophy in Computer Science

Submitted by

Mostafa Salama Abdelhady Mohamed

M.S. in Computer Science
Lecturer assistance
British University in Egypt

Supervised by

Professor Aly A. Fahmy
Department of Computer Science
Faculty of Computers & Information
Cairo University

Professor Aboul Ella Hassanien
Department of Information Technology
Faculty of Computers & Information
Cairo University

October 2011, Cairo

Approval Sheet

Data Mining for Medical Informatics

Submitted by

Mostafa Salama Abdelhady Mohamed

This Thesis Submitted to Department of Computer Science, Faculty of Computer Science and Information Technology, Cairo University, has been approved by:

Name	**Signature**
1. Prof. Dr. Ismail Abdel Ghafar Ismail	……..
2. Prof. Dr. Amir Ateya	……..
3. Prof. Dr. Aly Aly Fahmy	……..
4. Prof. Dr. Aboul Ella Hassanien	……..

December 2011, Cairo

List of Publications

Journal Papers:

1. **Mostafa A. Salama**, O.S. Soliman, I. Maglogiannisa, A.E. Hassanien, Aly A. Fahmy, "*Frequent pattern-based classification model without data presumptions*", Computers and artificial intelligence, 2011. [Submitted]
2. **Mostafa A. Salama**, A.E. Hassanien, Aly A. Fahmy, "*Binarization and validation in formal concept analaysis*", International Journal of Machine Learning and Cybernetics, 2011. [Submitted]
3. **Mostafa A. Salama**, A.E. Hassanien, Aly A. Fahmy, "*Fuzzification of Euclidean space in machine learning techniques*", International Journal of Approximate Reasoning, 2011. [Submitted]
4. **Mostafa A.Salama**, Kenneth Revett, Aboul Ella Hassanien, Aly A. Fahmy, "*An investigation on mapping classifiers onto data sets*", Journal of Intelligent Information Systems, 2012. [Submitted]

Peer Reviewed Book Chapters:

5. **Mostafa A. Salama,** O.S. Soliman, I. Maglogiannisa, A.E. Hassanien and Aly A. Fahmy, "*Rough set-based identification of heart valve diseases using heart sounds*", Intelligent Systems Reference Library, ISRL series, 2011. [In press]

Peer Reviewed International Conference:

6. **Mostafa A. Salama**, Aboul Ella Hassanien, Aly A. fahmy, Jan Platos and Vaclav Snasel, "*Fuzzification of Euclidian Space in Fuzzy C-mean and Support Vector Machine Techniques*", The 3rd International Conference on Intelligent Human Computer Interaction (IHCI2011), Pragu, published by Springer as part of their Advances in Soft Computing series, Aug. 29-31, 2011.
7. **Mostafa A. Salama**, Aboul Ella Hassanien and Aly A. Fahmy, "*Feature Evaluation Based Fuzzy C-Mean Classification*", The IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Taibai, Taiwan, jun 30, pp. 2534-2539, 2011.

8. **Mostafa A. Salama**, Kenneth Revett, Aboul Ella Hassanien and Aly A. Fahmy, "*Interval-based attribute evaluation algorithm*", The 6th IEEE International Symposium Advances in Artificial Intelligence and Applications, Szczecin, Poland, Sep 18-21, pp. 153-156, 2011.
9. **Mostafa A. Salama,** Aboul Ella Hassanien, Aly A. Fahmy, Tai-hoon Kim, "*Heart Sound Feature Reduction Approach for Improving the Heart Valve Diseases Identification*", The 2nd International Conference on Signal Processing, Image Processing and Pattern Recognition (SIP 2011), Dec. 8-10, 2011, International Convention Center Jeju, Jeju Island, Korea, CCIS/LNCS Springer series, vol. 260 (Indexed by SCOPUS), 2011.
10. **Mostafa A. Salama,** Aboul Ella Hassanien, Jan Platos, Aly A. Fahmy and Vaclav Snasel, "*Rough Sets-based Identification of Heart Valve Diseases using Heart Sounds*", The 3rd International Conference on Intelligent Human Computer Interaction (IHCI2011), Prague, published by Springer as part of their Advances in Soft Computing series, Aug. 29-31, 2011.
11. **Mostafa A. Salama**, Aboul Ella Hassanien and Aly A. Fahmy, "*Uni-Class Pattern-based Classification Model*", The 10th IEEE International Conference on Intelligent Systems Design and Applications (ISDA2010), Cairo, Egypt, pp. 1293-1297, Dec. 2010.
12. **Mostafa A. Salama**, Aboul Ella Hassanien and Aly A. Fahmy, "*Pattern-based Subspace Classification Model*", The second World Congress on Nature and Biologically Inspired Computing (NaBIC2010), Kitakyushu, Japan, pp. 357-362, Dec. 2010.
13. **Mostafa A. Salama**, Aboul Ella Hassanien and Aly A. Fahmy, "*Reducing the Influence of Normalization on Data Classification*", The 6th International Conference on Next Generation Web Services Practices (NWeSP 2010), Gwalior, India, pp. 609-703, Nov. 2010.
14. **Mostafa A. Salama**, Aboul Ella Hassanien and Aly A. Fahmy, "*Deep Belief Network for clustering and classification of a continuous data*", The IEEE International Symposium on Signal Processing and Information Technology, Luxor (SSPT 2010), Egypt, pp. 473-477, 2010.

Papers outside the medical data study:

15. **Mostafa A. Salama**, Heba F. Eid, Rabie A. Ramadan, Ashraf Darwish and Aboul Ella Hassanien, "*Hybrid Intelligent Intrusion Detection Scheme*", Advances in Intelligent and Soft Computing, vol. 96, pp. 295-302, 2011.
16. Heba F. Eid, **Mostafa A. Salama**, Aboul Ella Hassanien, Tai-Hoon Kim, "*Bi-Layer Behavioral-based Feature Selection Approach for Network Intrusion Classification*", The international Conference on Security Technology (SecTech 2011), Dec. 8-10, Jeju Island, Korea, 2011.

Abstract

Knowledge discovery in database (KDD) describes the process of automatically searching large volumes of data for patterns that can be considered knowledge about the data. Data mining is considered as the analysis step of the Knowledge Discovery in Databases process where it composed of the preprocessing, machine learning and visualization of the input data. Data preprocessing handles the input raw data to be more easily and effectively processed by machine learning and visualization techniques, such as data cleaning, transformation and feature reduction. Machine learning is a scientific discipline concerned with the design and development of algorithms that have an artificial learning capabilities such as supervised and unsupervised learning techniques. Finally the visualization of data to enable the understanding of hidden patterns and trends like Formal Concept Analysis. The data mining of real-life data like medical data is a key challenge in knowledge discovery applications. The diversity of the characteristics of the input data is hard to be handled by a single data mining technique like classification or clustering. Most of the data mining techniques have presumptions on the input characterizations like assuming that data is in a discrete form, in a normally distributed form or assuming the independence between attributes. Visualization data mining techniques like Formal Concept Analysis technique assumes that this input data is in a binary form. These characterizations presumptions may not exist in the most of the real-life data sets like medical data sets. The medical data sets could have many characteristics like it may contains continues features, and multi-variate features and it may suffers from the high dimensionality where these characterizes could violate assumptions of data mining techniques. The absence of these assumptions in the real-life data could negatively affects the results of the data mining techniques. Consequently, preprocessing techniques like

feature reduction, discretization and normalization algorithms should be applied to the input in order to be available to the data mining techniques. These preprocessing techniques itself also may have presumptions on the input and it could cause a distortion input data structure.

The work presented in this thesis investigates the nature of real-life data, mainly in the medical field, and the problems in handling such nature by the conventional data mining techniques. Accordingly, a set of alternative techniques are proposed in this thesis to handle the medical data in the three stages of data mining process.

In the first stage which is preprocessing, a proposed technique named as interval-based feature evaluation technique that depends on a hypothesis that the decrease of the overlapped interval of values for every class label leads to increase the importance of such attribute. Such technique handles the difficulty of dealing with continuous data attributes without the need of applying discretization of the input and it is proved by comparing the results of the proposed technique to other attribute evaluation and selection techniques. Also in the preprocessing stage, the negative effect of normalization algorithm before applying the conventional PCA has been investigated and how the avoidance of such algorithm enhances the resulted classification accuracy. Finally in the preprocessing stage, an experimental analysis introduces the ability of rough set methodology to successfully classify data without the need of applying feature reduction technique. It shows that the overall classification accuracy offered by the employed rough set approach is high compared with other machine learning techniques including Support Vector Machine, Hidden Naive Bayesian network, Bayesian network and other techniques.

In the machine learning stage, frequent pattern-based classification technique is proposed, it depends on the detection of variation of attributes among objects of the same class. The preprocessing of the data like standardization, normalization, discretization or feature reduction is not required in this technique which enhances the performance in time and keeps the original data without being distorted. Another contribution has been

proposed in the machine learning stage including the support vector machine and fuzzy c-mean clustering techniques, this contribution is about the enhancement of the Euclidean space calculations through applying the fuzzy logic in such calculations. This enhancement has used chimerge feature evaluation techniques in applying fuzzification on the level of features. A comparison is applied on these enhanced techniques to the other classical data mining techniques and the results shows that classical models suffers from low classification accuracy due to the dependence of un-existed presumption.

Finally, in the visualization stage, a proposed technique is presented to visualize the continuous data using Formal Concept Analysis that is better than the complications resulted from the scaling algorithms.

To my parents, my wife and teachers who gave their love, support, and time freely.

Acknowledgements

At first, at last and all the time, thanks to ALLAH the God of the world, for every thing in my life. Nothing in my life could be done without his permission, and no success could be gained without his mercy. Thanks to Prof. Aboul Ellah Hassanien for his very much support and encouragement to accomplish this thesis in a professional and valuable way, he sparked my interest in machine learning and classification techniques in the first place. Thanks to Prof. Aly Fahmy for his very important guidance and leading to me and to my group all the time during our research. Thanks to the sincere support, help and guidance of Eng. Heba Eid and Dr. Nashwa el-Bendary. My parents have set the cornerstone of this work. They have encouraged me all the time and who are always motivating and supporting me through my academic career. I also want to thank my wife for all of her support.

Contents

List of Figures

List of Tables

List Of Abbreviations

BN	Bayesian network classifier
DT	Decision tree classifier
EM	Expectation maximization algorithm
FCA	Formal concept analysis
FCM	Fuzzy c-mean clustering
FLCM	Fuzzy latent class model
HCV	Hepatitis C virus
HIV test	Human Immunodeficiency Virus
IB	Interval-based feature evaluation and selection technique
IB1	Instance-based classifier
IG	Information gain
KDD	Knowledge discovery in database
KS test	Kolmogorov-Smirnov test
LVQ	Supervised learning vector quantization
MLP	Multilayer perceptron
NB	Naive bayes classifier
NN	Neural network classifier
PBC	Pattern-based classification technique
PCA	Principle component analysis
SOM	Unsupervised self organizing maps
SVM	Support vector machine classifier
SVMB	SVM-based feature selection technique
TB	Tuberculosis bacterium disease

Chapter 1

Introduction

This chapter presents an introduction to the importance of handling medical data characteristics in knowledge discovery. Data in the medical field represents a real-life data that is considered as a great challenge in different data mining techniques due to the diversity of characteristics it concludes. If any of these characteristics are not handle probably, this will reflects on the correctness of the extracted knowledge. Some techniques in data mining may contain limitations that in some cases may be considered as conflicts with the existing characteristics in real life data set. A summery is presented about the problems that are facing the knowledge discovery in the medical field and the proposed solutions. Finally an overview about the organization of the thesis is shown at the end of the chapter.

1.1 A Background

Knowledge discovery in databases (KDD) is the process of extracting hidden patterns of useful and predictive information and patterns in bodies of medical data sets for use in decision support and estimation. Medicine requires KDD in Diagnosis, therapy and Prognosis. KDD is required in diagnosis to recognize and classify patterns in multivariate patient attributes, in therapy to select the most effective and suitable method to a patient from available treatment methods and finally in prognosis to predict future outcomes based on previous experience and present conditions. KDD could produce efficient screening techniques to reduce demand on costly health care resources, help physicians cope with the information overload and offers a better insight into medical

survey data. Data mining is considered as the main step in KDD that consists of three stages which are the data preprocessing, machine leaning and visualization techniques. Data mining offers methodological and technical solutions to deal with the analysis of medical data and construction of prediction models (1). Even though there are many data mining techniques to analyze the medical databases, most of them are still experimental and in the hands of computer scientists (2). Many attempts have been made in the last decades to design systems to enhance the performance and outcome of data mining techniques. Hybrid systems for classification is an example of these techniques that is applied by combining different and individual classification techniques (3). These hybridization techniques have been proposed to gather the strong points in different techniques. Recently, a number of approaches adopted a semi-supervised model for classification (4). Another approach for the movement from clustering to classification model appears like in supervised learning vector quantization (LVQ), which is based on a standard and unsupervised self organizing maps (SOM) (5).

One of the main problems in data mining that is considered as an important area of research that is targeted in this thesis is the relation between the data characteristics of real-life data sets and the data mining techniques. Medical field is considered as a resource of the real-life data sets that needs a lot of research and analysis. The variety in the characteristics of medical data is considered as a hard problem for data mining techniques that depend on presumptions on the input data. As shown in this thesis, the experimental analysis shows a deterioration in the results of data mining techniques if these assumptions are violated. This chapter introduces the general characteristics of data sets, then introduces the medical data issues including a brief idea about its relations to different characteristics of medical data. Then the thesis motivation, problem statement and a brief summery about the techniques proposed in this thesis to solve the maintained problem.

1.1.1 General data characteristics

The characteristics of any data can describe either the unique attributes in the input data set or a collection of attributes. The characteristics that describe *each attribute* are as follows:

- The values in the attribute are either continuous, discrete or binary values. The binary attribute could be considered as a discrete attribute except that only two discrete values are available. The difference between discrete and continuous attributes is that in continuous attributes the number of objects that have a certain value of the attribute is not significant. To deal with continues attributes, data mining techniques uses a range of values to detect the number of the objects in this range rather than using specific values. But usually the determination of the range of values depends on expert decision that could not be accurate all the time.

- The distribution of values in each attribute could be in any form, one of the most known distribution is the normal, gaussian distribution. In a normal distributed attribute, most of the values are close to a certain value called the mean value such that it appears as a peak in the distribution curve. Other distributions that may be considered are uniform distribution, binomial for distribution of more than one peak or gamma distribution where no secondary peak, only a single cure, in the distribution.

- Whether the attribute may contain missing values or not. Many of the real-world applications suffer a common drawback which missing or unknown data (incomplete feature vector). For example, in medical diagnosis, some tests cannot be done because either the hospital lacks the necessary medical equipment, or some medical tests may not be appropriate for certain patients. There exist many techniques to manage data with missing items, but no one is absolutely better than the others, as Allison says, "the only really good solution to the missing data problem is not to have any" (6).

The characteristics that describe a *group of features* are mainly two points which are more related to the classification of objects. Usually these characteristics are determined or handled through the corresponding class of each object in the input data:

- The dependence of some features on each other has a very important role in data mining techniques. This dependence means that a feature could be meaningless by its own unless in the presence of another feature or a group of features, in other words, a group of features could act as a single descriptive feature. The

dependence could appear in any form, like the trend of variation of one feature could be related the trend of another.

- The curse of dimensionality is one of the important problem in the data mining techniques. The high number of features (dimensions) could include redundant or irrelevant features where they have a noisy effect on the data mining technique. Two main benefits are gained if these features are ignored, the first is to enhance the computational performance and the second is to avoid a deterioration in the accuracy of classification, and finally is to study which feature are really related to the classification problem.

1.1.2 Medical data issues

Medical data applications include the prediction of the effectiveness of surgical procedures, medical tests and medications, and discovery of relationships among clinical and pathological data. There are various resources of medical data where this is an important factor in determining its nature. The ability to understand the nature of data characteristics in the medical field is an important factor to achieve medical data application of a high quality. This section shows the resources of medical data, the characteristics of these data and the effect of these characteristics in the accuracy resulted from different data mining techniques.

1.1.2.1 Sources of medical data

Medical data maintained for a patient can be summarized in the following categories:

Diagnoses Information about a patient's condition, disease, date, site of disease, etc. used to classify the client's status Previous Medical History - Information about previous medical conditions

Test Data Various tests performed for a patient including skin tests, chest x-rays, blood tests, HIV tests, bacteriology testing, and physical examination. i.e. HIV antibody tests are the most appropriate test for routine diagnosis of HIV, the virus that causes AIDS, among adults.

Medications Drugs prescribed for a patient to include start time, stop time, stop reason, dosages, etc.

Contacts Information about people with whom the client has been associated and who may have been exposed to TB through that association. i.e. TB is a disease caused by a bacterium called Mycobacterium tuberculosis.

Hospitalizations Information about previous or current hospital care the patient has received

Referrals Information about the person or organization that sent the patient to be examined

1.1.2.2 Characteristics of medical data

The capabilities required in data mining techniques to be handled if it is applied on a real-life medical data are as follows:

- Medical data could be in a discrete or a continuous form, where this is considered as one of the main problems in data mining in general. Several techniques do not handle the case when the data features are in a continuous form.
- Medical data distribution could be in any form, not only in the form of normal (gaussian) distribution. In other words, normal distribution is not a common or normal distribution (7). Deviations of empirical distributions from the normal distribution can be described in several ways. Comparing values of the mean and the median is one of them. If the mean and median do not coincide, the distribution is skewed. In the skewed distribution, the mean is pulled toward the skewed side more than the median. Variables that demonstrate marked skewness or kurtosis may be transformed or normalized to a better approximate normality (8). Normalizing the distribution made the scores of the new distribution evenly spaced.
- More realistically, a data model and its physical implementation must represent the relationships with other characteristics of the object. This relation indicates that the input data set is a multivariate data set. For example, an attribute that represents the weight of 150 kilograms is possible for a person in the adult age, but it is impossible for a newborn baby.

- Medical decisions often involve missing and error values in data. The records may have missing values for several reasons: limited number of tests required for diagnoses, logical exclusion of not applicable data (e.g., data specific to female gender is omitted from a record of a male patient) (9).

- Some data sets may contain irrelevant or redundant features, these data could increase the number of features (dimensions) that could lead to what is known as curse of dimensionality problem (10).

1.1.2.3 The effect of medical data characteristics on data mining technique

The effect of the medical data characteristics on data mining technique is an important area of study. Figure 1.1 shows different data sets, used in this thesis, and the corresponding classification accuracy resulted from different classification techniques. The results show that the machine learning performance is dependent on the input data set. The reason is that each input data set has different characteristics, where these characteristics may not fit to the classifier applied. This fact declares the importance of taking the characteristics of the input data sets into consideration in order to reach high classification accuracy. Also it refers to the fact that there is no machine learning technique is always perfect for all real life data sets, otherwise one of these machine learning techniques would show the highest classification accuracy for all existing data sets. The needed is to propose techniques that deal with these characteristics with avoiding of limitations that could cause deterioration in the classification accuracy.

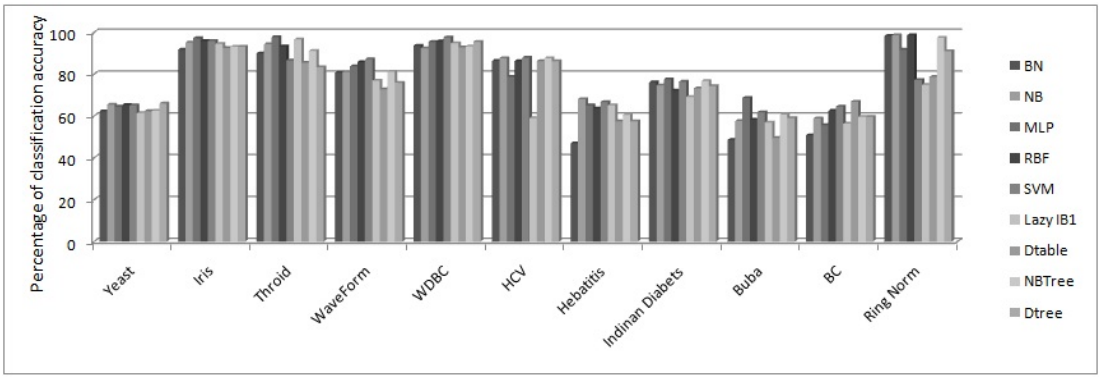


Figure 1.1: Classification Accuracy - Applying different classification techniques on different data sets.

A summery about some problems that are discussed in this thesis is list in the table 1.1. This table is another proof about the importance of handling and putting into consideration the characteristics of the input data.

Table 1.1: Some data mining techniques and the corresponding problems

Data mining technique	Challenge description
Decision tree	Perform better on discrete data sets
Bayes belief	Univariate attributes assumption
Principal component analysis	Normal distribution assumption
Support vector machine (SVM)	User defined parameters
Neural network	Difficulty in rule extraction
Formal concept analysis	Representation of binary data only
Some data mining techniques	Difficulty in handling missing data
Fuzzy c-mean methods and SVM	Sensitivity to outliers
Frequent pattern based clustering	User defined threshold according to the input data
General problem	Curse of dimensionality

1.2 Thesis Motivation

Data mining has attracted growing research attention for computing applications in medical informatics. The implications of data mining methodology and applications are manifested in the areas of health informatics, patient care and monitoring systems, assisting technology to knowledge extraction and automatic identification of unknown classes. Various algorithms associated with data mining have significantly helped to understand medical data more clearly, by distinguishing pathological data from normal data, for supporting decision-making as well as visualization and identification of hidden complex relationships between diagnostic features of different patient groups. In medical practice the collection of patients data is often expensive, time consuming and harmful for patients (11). Therefore, it is required to have the data mining methodology that is able to reliably diagnose with small amount of data about patients. In order to achieve this requirement, feature selection techniques should be applied to achieve the least amount of cost and time. However, the process of determining the right feature set is time consuming and may be a combinatorial problem. In addition, the transparency of diagnostic knowledge in the medical field is an important requirement to present a well explanation of knowledge to the physician. It should be easy for physician to be able to analyze and understand the generated knowledge. This requirement may be not available in techniques like neural networks and support vector machine techniques (12, 13). In medical diagnosis, the description of patients lacks certain data very often. The machine learning techniques should consider missing and error values in different attributes. A small percentage of error should be taken into consideration such that it doesn't lead to a major decrease in the classification accuracy.

1.3 Problem statement

In the knowledge discovery field, a problem that is targeted in this thesis is the presumptions of data mining techniques about the input data sets. In the case of real-life medical data sets, the presumptions of data mining techniques are not always true or applicable. The presumptions are mainly about the input data characteristics and its variety among different data sets. The characteristics under investigation are the data distribution, discreetness, correlation between attributes, and relevance to target class labels. For example, the decision tree classification technique assumes that the input

data is in a discrete form. In order to handle such presumptions, a preprocessing of data is always added as the first stage in data mining before applying the machine learning and visualization stages. The problem of the existing preprocessing techniques applied on data is that it may cause deformation in the input data to the next stages. In addition, these preprocessing techniques themselves may also have presumptions about the input data. These problems cause the decrease in the accuracy and correctness of the output from the machine learning and visualization techniques.

The preprocessing stage makes use of the data cleaning, data transformation and feature reduction techniques. In the case of data transformation techniques, data discretization and data normalization are applied. In data mining, it is often necessary to transform a continuous attribute into a categorical attribute using discretization process (14). Feature reduction techniques have presumptions on the input data which affects also on the correctness of the resulted data. An example of feature reduction techniques is the principle component analysis (PCA) which assumes that all attributes in data is normally distributed, where this is not common or normal in real life cases (7). Another example is that many feature selection techniques, as chisquare (χ^2) and information gain techniques, are shown to work effectively on discrete data or more strictly, on binary data (15). Data discretization and data normalization techniques could cause a distortion in the internal structure of the input data (16, 17), this on the return could cause a decrease in the correctness of the processing that is applied later in the data mining process.

In the machine learning stage, some techniques suffer from the presumption of discreteness of data. Logic-based learning machines like decision trees tend to perform better when dealing with discrete/categorical features (18). Bayesian model has a common assumptions which indicate that all variables should be discrete and normally distributed (19). Another problem appears in univariate models like Bayes belief models that assumes features are independent (20) where it will be computationally intractable unless an independence assumption (often not true) among features is imposed (21). On the other hand, some the machine learning techniques are independent on user defined thresholds like frequent pattern-based clustering technique. An inappropriate user defined threshold value may result in too many or too few patterns, with no coverage guarantees (22). Another indirect example which is considered as a big limitation in the support vector machine technique, is the dependance on the expert users in the

choice of the kernel and the selection of the corresponding parameters.
Finally in prominent visualization techniques like formal concept analysis, the expected input data is to be in a binary form (23). This limitation of using continuous attributes in formal concept analysis technique is solved by using a scaling algorithms, but this technique is highly computational and causes the generation of a complicated and an unreadable lattice(24).

1.4 Scope of work and proposed solutions

Proposed solutions have been presented to avoid most of the problems discussed in the previous section; these solutions are involved in each of the stages of data mining. In the preprocessing stage, an interval based feature evaluation and selection technique (25) is proposed to handle data of a continuous or a discrete form. The technique avoids the use of an extra preprocessing and avoids the distortion of the input data. Also an experimental study has been applied to prove the negative effect of applying the normalization algorithm before the PCA feature extraction technique, and discusses the use an sternutative method of PCA that depends on the use of correlation matrix to avoid to the use of a normalization technique (26). In the machine learning stage, classification model based on pattern-based clustering technique (27, 28) have been proposed. This technique, which is named as the frequent pattern-based classification technique, depends on the variation of the attribute values among different attributes in the same object. The advantage of this technique that it does not have presumptions that could be unavailable in the medical data sets. Another contribution in the machine learning stage is the fuzzification of Euclidean space calculations implemented in machine learning techniques like fuzzy c-mean clustering and support vector machine (29). The fuzzification is applied through the use of ranks evaluated by the chimerge feature selection technique. The reason of such enhancement is that the data features even after applying a feature selection technique do not have the same degree of relevance to the classification problem. In the visualization stage, to avoid the limitation on continuous input when applying formal concept analysis, a binarization and validation techniques are proposed. This leads to the generation of a more understandable and simple lattice rather than the use of scaling algorithm.

1.5 Thesis Organization

The thesis consists of seven chapters, as seen in figure 1.2 that shows an outlined flow chart of the thesis, including the introduction. The introduction explores the characteristics of the input data and how the current classifiers have limitations on such characteristics, where these limitations could affects the accuracy and correctness of many learning algorithms. Chapter 2, shows the main steps of data mining, and shows the role of each stage in data mining in separate.

chapter 3 introduces the different data preprocessing techniques and the effect of high dimensionality and feature relevance on the learning algorithm. Second, it reviews a variety of different dimensionality reduction and feature selection through feature weighting. Third, it declares the negative effect of normalization and discretization algorithms on the results of learning techniques. Finally a feature evaluation technique is proposed to handle the continuous data problem.

Chapter 4 reviews a variety of different machine learning techniques and explores their behavior towards the characteristics of the input data. A proposed pattern-based subspace classification technique is provided to overcome the limitations on different characteristics exists in different conventional machine learning techniques. A modifications is applied on the Euclidean space calculations that is considered as a core method in support vector machine and fuzzy c-mean clustering techniques.

Chapter 5 presents an approach to represent continuous data using a visualization known as formal concept analysis technique. The approach uses chimerge technique in the binarization algorithm rather than using the usual scaling algorithms.

Chapter 6 shows the experimental work performed and the characteristics of the data used in test the proposed approaches or techniques. This chapter includes the empirical evaluation of the proposed techniques where the results for each technique are discussed. This chapter includes a comparison between the different approaches and the proposed techniques.

Finally, chapter 7 summarizes the results obtained from the empirical investigation presented in the chapter of experimental work. The impact of these results is discussed with respect to the proposed techniques. Finally, a number of the issues that have been raised by this work are discussed, and presented as directions for further study.

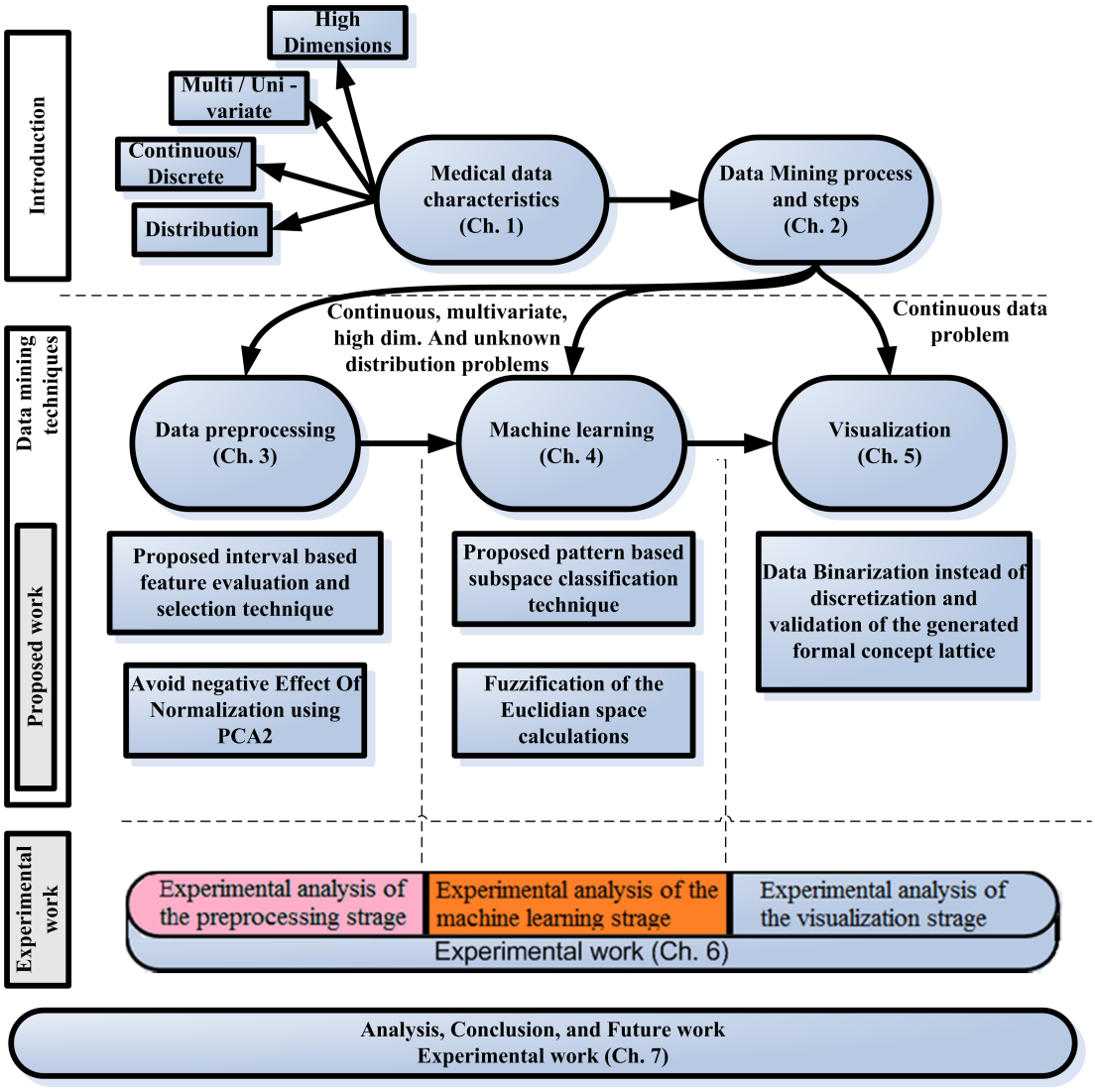


Figure 1.2: Thesis Organization - Outlined flow chart of the thesis

Chapter 2

Data Mining: Architecture and background

Data mining is the first step before knowledge interpretation in knowledge discovery. Data mining can be viewed as three main stages which are data preprocessing, machine learning and visualization. In order to grantee the success of discovering knowledge from medical data, all of these three stages should be investigated. The capabilities of each data mining techniques in dealing with the characterizes of the medical data and the possibility of finding solutions for there difficulty is are a very important area of research. This chapter shows these three stages to be discussed briefly in the next chapters.

2.1 Introduction

Data mining is considered as a part of knowledge discovery that is applied before the knowledge interpretation part. Data mining techniques predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining techniques can answer business questions that are traditionally too time-consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The first conclusion that appears here that Data mining is not knowledge discovery, the results from data mining techniques are fed to another method named as the knowledge interpretation method to extract the knowledge. And since that any technique that is responsible

for extracting hidden patterns from data is considered as a data mining technique then second conclusion is that Formal concept analysis is considered as data mining technique (30). Formal Concept Analysis as visualization tool is used both for its abilities in data-mining and information representation. The data mining can be described into three main stages which are data preprocessing, machine leaning and visualization techniques. And in general, the stages of knowledge discovery can be described in the following figure, figure 2.1:

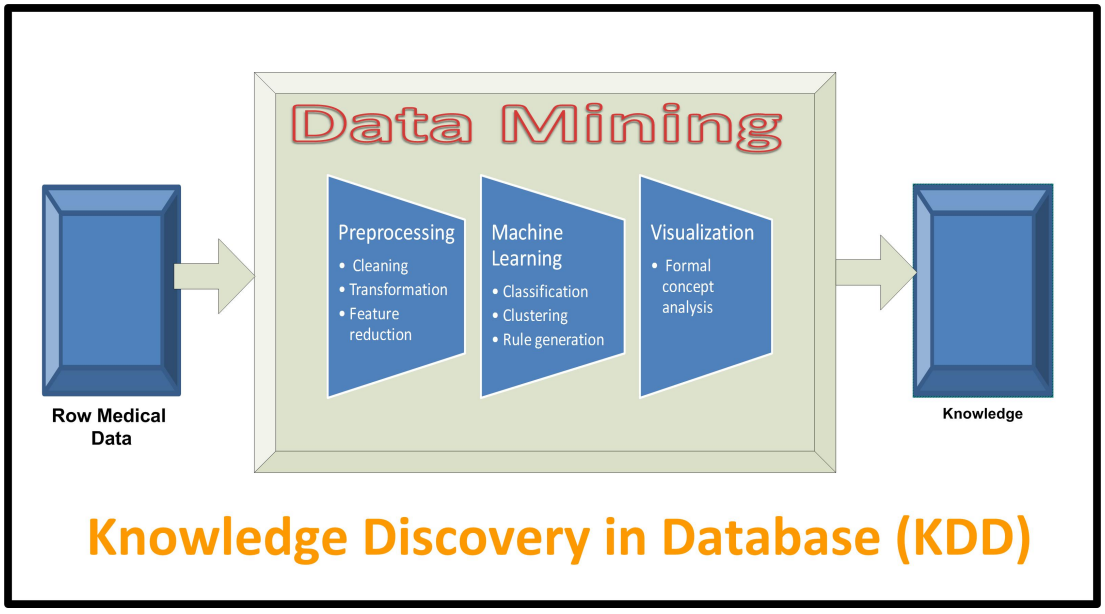


Figure 2.1: Knoweldge discovery - Stages of knowledge discovery

2.2 Data mining stages

The different stages of data mining are not fully separated, as in many techniques they can work together. As an example, support vector machine could be used in feature selection. Also in rough set, a part of the rough set classification technique is the extraction of reducts. This section investigates each stage separately but explores the hydride techniques when needed.

2.2.1 Data preprocessing stage

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice where it prepare data for machine learning and for visualization. Data preprocessing transforms the data into a format that will be more easily and effectively processed. For example, in a neural network, the data should be in a discrete form. There are a number of different tools and methods used for preprocessing, including:

- Data cleaning
- Data transformation
- Data reduction

However still some of the data preprocessing tools have problems, like presumptions on the input data, deformation in the original structure of data and increasing the complexity of the data mining.

2.2.2 Machine learning stage

Machine learning systems are a well-established paradigm with current systems having many of the characteristics of biological computers and being capable of performing a variety of tasks that are difficult or impossible to do using conventional techniques. Recent trends aim to integration of different components to take advantage of complementary features and to develop synergistic systems leading to hybrid architectures such as rough-fuzzy systems, evolutionary-neural networks, or rough-neural approaches for problem solving. The combination or integration of more distinct intelligent methodologies can be done in any form, either by a modular integration of two or more intelligent methodologies, which maintains the identity of each methodology, or by integrating one methodology into another, or by transforming the knowledge representation in one methodology into another form of representation, characteristic to another methodology. However these techniques fails to avoid the negative points in each of the combined techniques, where the presumptions on the input data still used.

2.2.3 Visualization stage using formal concept analysis

Formal concept analysis (FCA) is a mathematical theory of data analysis using formal contexts and concept lattices for automatically deriving an ontology from a collection of objects and their properties (31). In this thesis, the concept lattices are considered as the visual representation of the data. Before finding formal concepts in a many-valued context, this context has to be turned into a formal context (one-valued): many-valued attributes are discretized. This procedure is called discretization in data analysis, and termed also conceptual scaling in FCA. The problem of scaling method that is discussed in this thesis is that it increases the number of attributes, and so increases the complexity of the generated lattice.

2.3 Chapter conclusion

It is stated that the three core stages of data mining are the preprocessing, machine learning and visualization. Each of these stages sufferers from a group of problems that are needed to be handled. These problems are all about data as discussed in the introduction. The details of the each of these data mining stages will be discussed in depth in the next three chapters. Also propose solutions to handle the problems of each of the data mining stages will be shown in each chapter.

Chapter 3

Preprocessing stage and proposed techniques

The data plays an important role in the selection of the data mining technique that are required to perform the classification of the input data. The selection should not depends only on how much the classification technique is famous or trustable. The data should be described according to its characteristics as discussed previously. Different researches have studied the characteristics and the quality of data in order to enhance the performance of the applied classifiers. The results of these researches proposes different preprocessing techniques that became an important part in data mining. This chapter discusses these characteristics and shows the different dimensions of data quality. Then shows different preprocessing techniques. And finally shows two different proposed approaches that leads to a better classification accuracy results.

3.1 Introduction

Recent researches that are related to data mining have considered the data according to its quality. They considered the data to be of high quality if it can fit for their intended uses in operations, decision making and planning (32). In other words, the data is said to be of a high quality, if it can correctly represent the real-world construct to which it refers. In order to measure the data quality, it should be focused on a number of aspects that are meaningful and relevant for the classification problem without spending too many resources. There are six common relative aspects to the data quality definition

named as the dimensions of the data quality. These dimensions can be listed as follows:

- Completeness: Is all necessary data present, no lacking attributes or certain attributes of interest and no existence of missing data.
- Validity: Are all data values within domains specified by the problem
- Accuracy: The data reflect the real-world objects verifiable source, this means that there are no noise in data. The data is said to be noisy when it contains error or outlier values that deviate from the expected
- Consistency: Data consistent between systems, such that no discrepancies between different database that could cause the existence of duplicate records.
- Integrity: Are the relation between entities and attributes consistent?, Also some attributes that represents a given concept may have different names in different databases causing inconsistencies.
- Timeliness: Is the data available at the time needed?

The quality of the input data can be measured using several methods and parameters. These measures could determine the nature of the input data and clarify its quality and applicability to be used in data mining techniques. An example of these measures are the normality, sparsity and fuzziness measures of the input data.

- Normality test algorithm is applied on each column on each data set like the Kolmogorov-Smirnov test (K-S test) to compare the values in each column to a standard normal distribution.
- Input data or base data is said to be sparse if it is not densely populated. As the number of dimensions increase, data will typically become sparser (less dense).Sparsity has also played a central role in the success of many machine learning algorithms and techniques such as matrix factorization (33). The comparison in (34) shows that Gini Index is the best measure where it satisfies all the criteria maintained in this thesis
- Crisp items belong exclusively to one category, whereas fuzzy items belongin varying degreeto multiple categories. This relaxation in the assumption about the

nature of qualitative data makes fuzzy latent class model (FLCM) more widely applicable. The study in (35) proposes a moment-based measure of overall data fuzziness that is bounded by 0 (completely crisp) and 1 (completely fuzzy).It is based on the cross-product moments of the g_{ik} distribution which capture the extent/degree to which an item i is a member of category k, this grade is subject to the constraints in equation 3.1

$$\sum_{k=1}^{K} g_{ik} = 1, 0 \leq g_{ik} \leq 1 \tag{3.1}$$

3.2 Data preprocessing

In order to enhance the quality of the input data sets after preprocessing, several stages should be followed in order to reach a high accuracy when applying different machine learning techniques. These stages are named as the preprocessing of the input data sets so it could be applicable for processing in machine learning and visualization techniques. As discussed previously there are five main characteristics of the input data sets that are recently discussed in many researches. These characteristics should be handled and taken into consideration before applying a classification technique. Each classification technique has a special requirement (assumptions) in the input data set that should satisfied in the input data sets. Data preprocessing is required to do such job. The substages of preprocessing can be applied through three main stages, data cleaning, transformation and reduction are shown in figure 3.1.

3.2.1 Data cleaning

The first stage in the data preprocessing is the data cleaning. Data cleaning algorithms work to "clean" the data by three main steps:

Handling the missing values Most of the approaches in the literature can be grouped into four different types depending on how both problems are solved:

- Deletion of incomplete cases and classifier design using only the complete data portion.

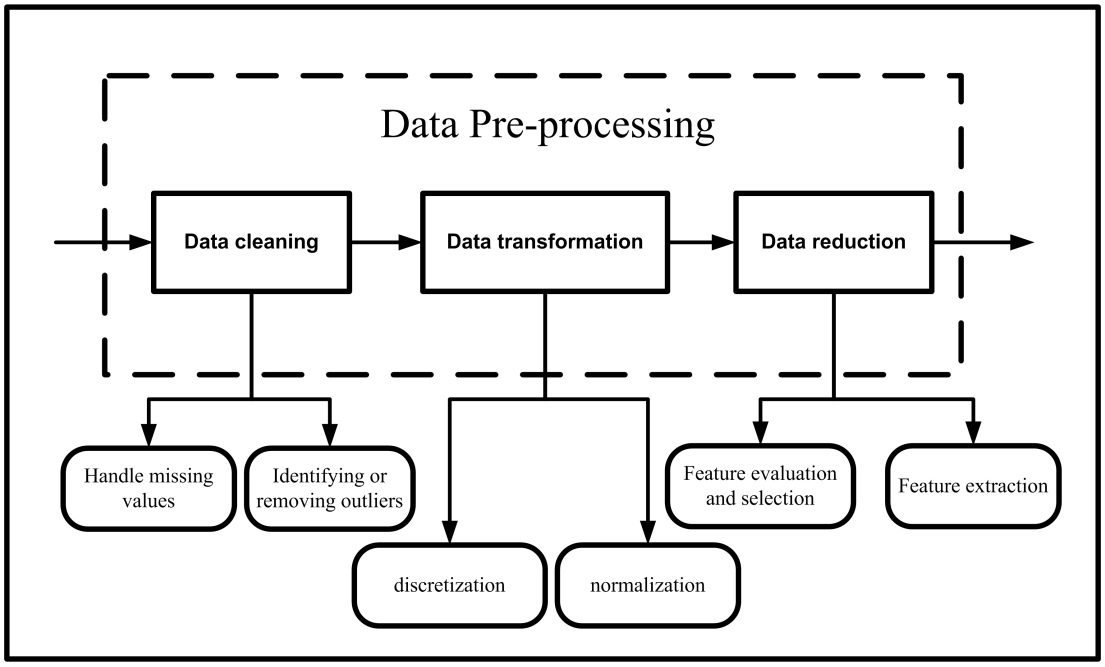


Figure 3.1: Preprocessing stage - Three substages of preprocessing

- Imputation or estimation of missing data and learning of the classification problem using the edited set, i.e., complete data portion and incomplete patterns with imputed values.
- Use of model-based procedures, where the data distribution is modeled by means of some procedures, e.g., by expectation maximization (EM) algorithm.
- Use of machine learning procedures, where missing values are incorporated to the classifier.

Identifying or removing outliers Outliers represent very different points from the rest of the data. It often contains useful information on abnormal behavior of the system described by the data (36). The basic approaches used currently in data mining systems for solving the outlier detection problem can be summarized as follows (37):

- Statistical techniques: Are based on the construction of a probabilistic data model. If the object does not suit the probabilistic model, it is considered

to be an outlier. Probabilistic models are constructed with the use of standard probability distributions and their combinations. The construction of this model based on empirical data and finding the needed parameters are considered as complicated computational tasks.

- Distance-based techniques: Are based on the calculation of distances between objects of the database and have a clear geometric interpretation like k-Nearest Neighbors. These techniques can find the local outliers for each class, but it has two main problems. The first problem is that it has a quadratic complexity. The second problem is its difficulty in dealing with the fact that the majority of modern information systems contain heterogeneous data of complex structure.
- Kernel Function techniques: Deals with heterogeneous structured data problem. The kernel functions are defined for discrete structured objects. The problem in this approach is that the decision function which describes the outlier factor is discrete. In other words, it is impossible to estimate how much one object is "worse" than another. Another problem is the case when dealing with large data sets.
- Fuzzy approaches: Combines techniques of fuzzy set theory to other techniques, for example the kernel function techniques.

Resolving other problems Resolving inconsistencies and removing redundancies that could be resulted from the data integration.

3.2.2 Data transformation

There are two main types of transformation which are Data discretization and data normalization.

Data discretization Data could be in a discrete or continuous form, where this is considered as one of the main problems in data mining in general. Several techniques do not handle the case when the data features are in a continuous form. The algorithms that suffers from such characteristic are:

- Logic-based learning machines like decision trees tend to perform better when dealing with discrete/categorical features (18).

- Statistical-based learning machine techniques like naive bayes classifiers and bayesian networks.
- rough set classification techniques
- visualization techniques like Formal Concept Analysis that requires data to be in a binary form.

It is often necessary to transform a continuous attribute into a categorical attributes using discretization process (14).

Data normalization Data could be in any form, not only in the form of normal (gaussian) distribution. In other words, normal distribution is not a common or normal distribution (7). Variables that demonstrate marked skewness or kurtosis may be transformed or normalized to a better approximate normality (8). In some data mining techniques like PCA and Bayesian techniques, each observed variable should be normally distributed (21). However, Data sets of many real-life cases like medical data-sets do not follow such assumption and it could be considered as a limitation for many data-sets. Again normalization process could lead to a distortion in the actual structure of the data that could lead to a decrease in the classification accuracy. There are two main types of data normalization, which are either:

- Min-max normalization, where it performs a linear transformation on the original data. Such that if max_A, min_A are the maximum, minimum values of attribute A respectively. The required is to normalize data to range of the interval $[new_{min_A}, new_{max_A}]$. And the $\grave{v}$ is a value after applying normalization and v is the original data value.

$$.\grave{v} = \frac{v - min_A}{max_A} * (new_{max_A} - new_{min_A}) + new_{min_A} \tag{3.2}$$

- z-score (z-mean) normalization, where the values of an attribute A are normalized based on the mean and standard deviation of A which are $\bar{A}$ and σ_A respectively.

$$\grave{v} = \frac{v - \bar{A}}{\sigma_A} \tag{3.3}$$

3.2.3 Data reduction

Data sets may contain irrelevant or redundant features that are considered as misleading to the classification technique applied. These features could lead to what is known as curse of dimensionality problem (10). The application of feature selection techniques greatly reduces the computational cost and increases the classification accuracy of classifying high dimensional data. The selection of the most features important features to the classification problem could be based on two phases, feature evaluation or ranking and feature selection. There are a lot of possible combinations between each feature search and each feature evaluation algorithms(38). Feature evaluation techniques like chimerge, information gain and gain ration evaluate the importance of each feature in the input data set. This does not mean that all (or any) of the features in the data set have high individual importance. Chimerge feature evaluation technique is the most applicable technique in dealing with continuous data, and accordingly is most suitable to be used in the case of medical data sets like the data extracted from heard sound. Feature evaluation technique involves the evaluation of each feature according to the target class labels, while feature selection techniques perform the evaluation of subset of feature explicitly via a predictive model, classifier, built from just those features. The reason of using feature selection via feature evaluation techniques is that there is no direct way of evaluating the correctness of the order of the individual features defined by the feature evaluation techniques. Feature selection is grouped in two ways according to the attribute evaluation measure: depending on the type (filter or wrapper techniques) or on the way that features are evaluates (individual or subset evaluation). The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. Feature subset selection algorithms can be divided into three categories: exponential, randomized, and sequential (39). In exponential search algorithms such as exhaustive search and branch and bound (B&B), the number of subsets grows exponentially with the dimensionality of the search space. Sequential algorithms have reduced computational complexity, Sequential forward and backward selection, Plus-l Minus-r selection, and sequential floating selection are examples of such methods. Randomized search methods. Genetic algorithms and simulated annealing fall into this category. Sequen-

tial algorithms tend to become trapped in local minima due to the so-called nesting effect. Randomized algorithm try to avoid the problem of local minima by adding randomness to the search. Several algorithms have been proposed previously like in (40), where the wrapper selection over Feature Ranking has been implemented. There are two phases in these algorithms, firstly, the features are ranked according to some evaluation measure. In second place, the list of attributes is used once, crossing the ranking from the beginning to the last ranked feature. This behavior of the variation of the classification accuracy according to the number of selected attributes is shown in figure 3.2. It is noticed that the classification accuracy increases as the number of selected attributes increases until a certain number of attributes where a specific peak of accuracy, then the accuracy decreases, also it must be stated that the chart may contain many local extrema besides the global maximum value. Another way of feature

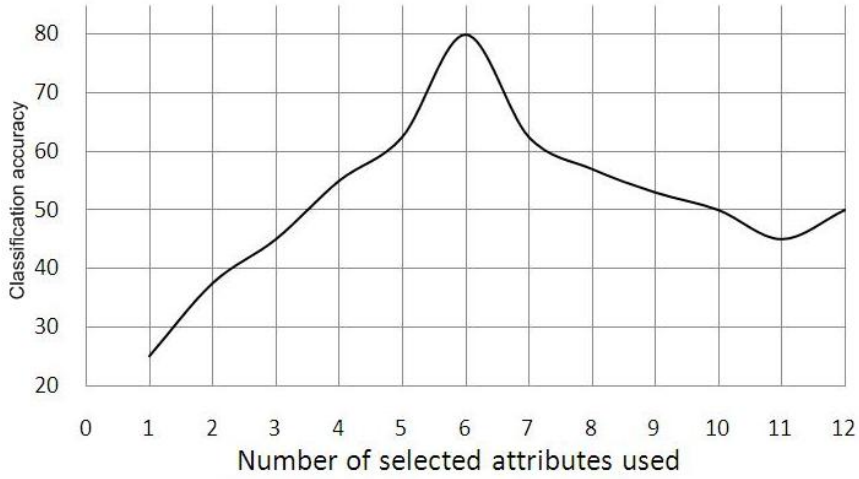


Figure 3.2: Classification behaviour - Apply the classifier on data sets that composes the most important feature gradually until all feature are used

reduction instead of the removal of irrelevant and redundant features is the extraction of less number of features out of the original feature set. An example of such techniques is the principal component analysis technique (PCA). But these methods may have an effect on the input data set due to its dependence on normalizaiton algorithms as discussed previously.

The next subsections discuse two examples of feature evaluation techniques which are the chimerge and information gain feature evaluation and ranking techniques, and one

example of feature extraction technique which is the conventional PCA. These techniques are used in the proposed and the experimental work in the thesis.

3.2.3.1 Chimerge feature evaluation

One of the most popular feature selection techniques is the chi-Square χ^2 algorithm, which measures the lack of independence between each attribute A and the target class c. Chimerge or Chi2-Square is a χ^2-based discretization algorithm (15, 41). It uses the χ^2 statistic to discretize numeric attributes repeatedly until some inconsistencies are found in the data that achieves attribute selection via discretization. The χ^2 value is calculated for each continuous attribute as follows: Initially, each distinct value of a numeric attribute A is considered to be one interval. The values,intervals, of attribute A are sorted and the χ^2 is applied for every pair of adjacent intervals as follows:

$$\chi^2 = \sum_{i:1..2} \sum_{j:1..k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{3.4}$$

Where:
A_{ij} is the number of values in the ith interval and jth class,
R_{ij} is the number of values in the jth class $= \sum_{j:1..k} A_{ij}$,
C_{ij} is the number of values in the ith interval $= \sum_{i:1..2} A_{ij}$,
N is the total number of values $= \sum_{i:1..2} R_{ij}$,
and E_{ij} is the expected frequency of $A_{ij} = \frac{R_{ij} * C_{ij}}{N}$

Adjacent intervals with the least χ^2 values are merged together, because low χ^2 values for a pair indicates similar class distributions. This merging process proceeds recursively until all χ^2 values of all pairs exceeds a parameter *signlevel* (initially 0.5). Then repeat the previous steps with a decreasing *signlevel* until an inconsistency rate is exceeded, where two patterns are the same but classified into different categories.

3.2.3.2 Mutual information attribute evaluation

The mutual information MI (also called cross entropy or information gain) is a widely used information theoretic measure for the stochastic dependency of discrete random variables (42, 43). The mutual information $I(A;C)$ between values of attribute A and the set of classes C can be considered as a measure of the amount of knowledge on C

provided by A (or conversely on the amount of knowledge on A provided by C). In other words $I(A;C)$ measures the interdependence between A and C where it can be computed as follows:

$$I(A;C) = H(C) - H(C|A) \tag{3.5}$$

The entropy $H(C)$ measures the degree of uncertainty entailed by the set of classes C, and can be computed as

$$H(C) = -\sum_{c \in C} p(c) \log p(c) \tag{3.6}$$

where $p(c)$ is the probability density function (PDF) of c. The conditional entropy $H(C|A)$ measures the degree of uncertainty entailed by the set of classes C given the set of attribute values A, and can be computed as

$$H(C|A) = \sum_{c \in C} \int_{a \in A} p(a,c) \log \frac{p(a|c)p(c)}{p(a)} da \tag{3.7}$$

The integration in the expression above signifies that the attribute space is continuous.

3.2.3.3 Conventional principle component analysis (PCA1)

Conventional PCA (PCA1) (44) is an unsupervised feature extraction technique. It extracts a subset of new features from the original set of normally distributed features by means of some functional mapping, keeping as much information in the data as possible. The approach can be summarized as shown in the following steps:

- Check the normality of the each variable in the input data set
- If (not normally distributed)
- Normalization algorithm is applied
- Suppose N vectors of M dimensions $x : x_1, x_2, .., x_M$
- for each dimensions calculate:

$$\bar{x} = \sum_{i=0}^{M} x_i \tag{3.8}$$

- Subtract the mean $\varphi_i = x_i - \bar{x}$

- Form the NxM matrix $A = [\varphi_1 \varphi_2 \varphi_3 .. \varphi_M]$
- Compute the covariance $N \times N$ matrix C using:

$$C = \sum_{n=1}^{M} \varphi_n \varphi_n^T = AA^T \tag{3.9}$$

- Compute the Eigen values of C such that $\rho_1 > \rho_2 > .. > \rho_N$
- Compute the Eigen vectors $C : u_1, u_2, , u_M$
- Since C is symmetric, any vector $x(x - \bar{x})$ can be written as a linear combination of Eigen vectors.
- Dimensionality reduction ($N \rightharpoonup K$) keeps only terms principle components corresponding to K largest Eigen values where $K << N$.
- The necessary cumulative percentage of variance explained by the principal axes should be consulted in order to set a threshold, which defines the number of the principle components k to be selected.

Accordingly, PCA is found to have the following properties:

1. It maximizes the variance of the extracted features;
2. The extracted features are uncorrelated;
3. It finds the best linear approximation in the mean-square sense;
4. It maximizes the information contained in the extracted features.

Also it have several assumption, one of these assumptions is the normality of data. So normalization step is an important preprocessing step in case that the variables are not normally distributed.

3.3 Preprocessing negative effect and proposed techniques

According to the maintained problems in preprocessing stage, two enhancements have been proposed in order to increase the classification accuracy next to this stage. The first enhancement introduces the importance of avoiding using the normalization algorithm before applying PCA, by using the correlation matrix in PCA. The second enhancement shows the importance of avoiding using the discretization algorithm before applying feature selection techniques, by proposing a novel supervised feature selection and evaluation technique.

3.3.1 The effect of applying normalization

Not all the real-life data have normal distribution, and applying the normalization algorithm could affect the structure of the input data set. Also it affects the outcome of multivariate analysis and calibration used in data mining. The study in (26) shows the negative effect of applying normalization algorithm on the input data set. It proves that using standard deviation in PCA technique, Correlation PCA technique (PCA2), without applying the normalization step has a better performance than using normalization before applying PCA1.

According to (45), Pearson correlation can be used as a ranking criterion of the dependency between features and classes (discrimination power of each feature). Pearson's correlation is used to find the correlation between two continuous variables, it is computed as follows:

$$R(i,j) = \frac{cov(x_i, x_j)}{\sqrt{var(x_i) * var(x_j)}} \tag{3.10}$$

Where R(i, j) represents the correlation between x_i and x_j variables, $cov(x_i, x_j)$ is covariance between these two variables and $var(x_i)$ is the variance of the each variable. On the other hand, According to (46), feature selection of q features out of n feature could be applied using PCA. This technique performs the selection by trying all combinations of q feature to calculate the covariance matrix, then choose the combination the mostly maximizing the covariance matrix. This means that the calculation of the covariance matrix is effecting on the discrimination between features. PCA2 technique depends on Pearson's correlation in the calculation of φ_i. It uses square root of the

variance (standard deviation σ_i) in the calculation of φ_i that is defined as follows:

$$\varphi_i = \frac{(x_i - \bar{x})}{\sigma_i}. \tag{3.11}$$

The resulted matrix is called correlation matrix instead of covariance matrix which indicates the strength and direction of linear relationship between features. The PCA that uses the correlation matrix instead of the covariance matrix is named PCA2, such that PCA2 depends on Pearson's correlation in the calculation of φ_i. Steps of comparing the two preprocessing types of PCA1 and PCA2 are as follows:

- Normality test algorithm is applied on each column on each data set like the Kolmogorov-Smirnov test (K-S test) to compare the values in each column to a standard normal distribution.
- The preprocessing technique applied before the classification could be any of the following three preprocessing techniques, where these techniques will be compared in this study:
  - Normalization of all features in the data set followed by PCA1
  - PCA1
  - PCA2
- The classification will be repeated for four times, as there are three models of preprocessing and the case when no preprocessing is applied.
- The classification technique that will be used is Multilayer perceptron (MLP) with 10-fold cross validation, having 10 folds means 90% full data is used for training (and 10% for testing) in each fold test.
- In case of using PCA in the second, third and the fourth model, feed-forward feature selection technique will be used to select the number features extracted from the PCA to be used in the MLP Classifier, These extracted features are defined as Principle Components. The test will be repeated for several numbers of the Principle Components starting from one and ending with original number of features. The selected number of Principle Components is the number that leads to the highest classification performance.

PCA2 shows the best classification performance, the results of this thesis will be discussed in the chapter of the experimental study 6.

3.3.2 The effect of applying discretization and a proposed Interval-based feature evaluation technique

Most of the feature selection techniques depends on an assumption that the input has a discrete data. Discretization algorithm is applied to satisfy such condition, where it could not preserve the original structure of the input. A proposed technique named interval-based attribute selection techniques is used to resolve the need of discretization preprocessing stage (25). The interval-based attribute selection algorithm depends on a hypothesis that as the intersection between attribute value ranges of different class labels decreases as the importance of this attribute increases. The reason of this hypothesis is that if an attribute has a certain continuous range of values appears only in the case of a certain class label, then this attribute can help as an indication to this class label. And as the length of this kind of ranges increases as the importance of this attribute increases. Figure 3.3 shows an attribute that contains an interval for each class. The hashed areas show the ranges of values that are not overlapped between

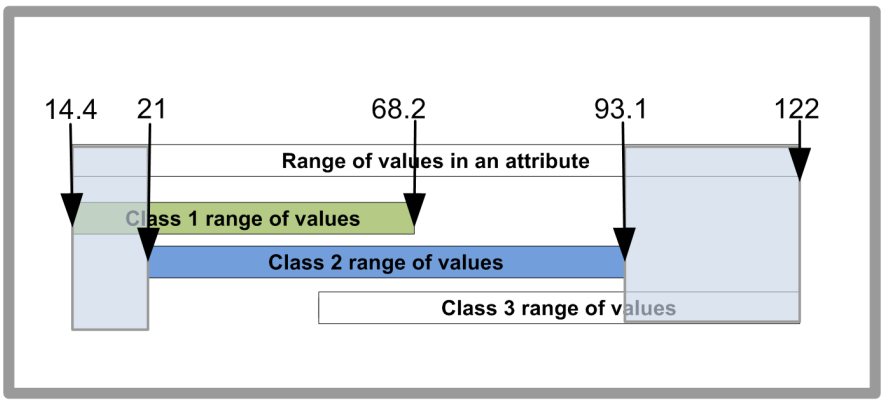


Figure 3.3: Intervals - Non overlapping Intervals between different class labels

multiple class, only a single class label is assigned to this label. In order to evaluate the importance of such attribute, the number of values in ranges that falls in a single class will be calculated for every class and summed. i.e. as shown in figure 3.3, the number of values that falls in the hashed areas are counted. Then the resulted value, after refinement of this count as shown in the equation 3.12, will be considered as the

attribute rank among other attributes.

$$\mu_a = \frac{1}{n} * \sum_{c \in C} \frac{n_{ci}}{n_c} \tag{3.12}$$

μ_a represents the rank of attribute a, n is the number of objects in the data set, n_c is the number of values where the corresponding objects are of class label c, and n_{ci} is the number of values in a rang that is completely falls in class c where this range is not overlapped with other class labels. For a two class data set, algorithm [1] can be used to calculate the interval-based ranking value which is μ_a. The algorithm detects, for every class, the range of values for an attribute that are not in the class and hence counts the number of objects in that range.

The removal of misleading values in Algorithm [2] is an optional step as it depends

Algorithm 1 Calculate Interval-based rank μ_a of attribute a

μ_a: Attribute a's rank, initial value is 0
AttributeLength: Number of objects
x_a and n_a: max and min values of attribute a
for Each Class label c **do**
  Remove misleading values.
  Determine the interval that represent the range of values of the attribute in that class label.
  IntervalLength: Number of objects in class c
  x_{ac} and n_{ac}: The max and min values of this interval.
  //Calculate the number of values outside the interval range.
  μ_c: Initial value is 0
  for Each value v in attribute a **do**
    if $v < n_{ac}$ **or** $v > x_{ac}$ **then**
      $\mu_c = \mu_c + 1$.
    end if
  end for
  $\mu_c = \mu_c$ / *IntervalLength*
  μ_a=$\mu_a + \mu_c$
end for
$\mu_a = \mu_a$ / *AttributeLength*

on the collection methodologies whether it is accurate or not. This step decrease the

sensitivity to outliers by removing the values that are most far away from the average of the attribute values. This step should remove only a small percentage of the values in the attribute in order not to affect the accuracy of the results.

Algorithm 2 Remove percentage x of misleading

Input: $Inteval$ values of an attribute a for objects lies in class c
Output: avg average of the values of an attribute a in a class c
for $x * IntervalLength$ values **do**
  Remove the value of max difference from the average avg.
end for

3.3.3 The independence of rough set classifier on feature reduction

rough set theory (47) provides the tools that could successfully produce high classification accuracy and generate an interpretable rules. The main goal of the rough set analysis is the induction of approximations of concepts, as it is based on the premise that lowering the degree of precision in the data makes the data pattern more visible. In order to applying classification, first the input data will be discretized using a rough set and boolean reasoning discretization method (48), then rules are generated, and finally classification is applied based on the generated rules. Also the reducts concept in rough set theory allows to keep only the attributes that are not redundant and their removal could not worsen the classification, where this could a privilege for rough set over decision trees (49). Rough set put into account the relation among attributes. (50, 51)

On applying rough set discretization and classification on medical data sets, the need of applying feature selection is not required due to the reducts concept, the rough set produces the highest classification accuracy. When the reducts were found, the job of creating definite rules for the value of the decision feature of the information system was practically done. The rules generated that are used in classification are also useful to detect some of the knowledge and facts that exist in the input data set. The model in figure 3.4 involves the extraction of reducts and rule generation. In this model, the discretization based on rough set and boolean reasoning (RSBR) (48), and the rule generation from the extracted set of reducts is applied by (GDT RS) (52). After

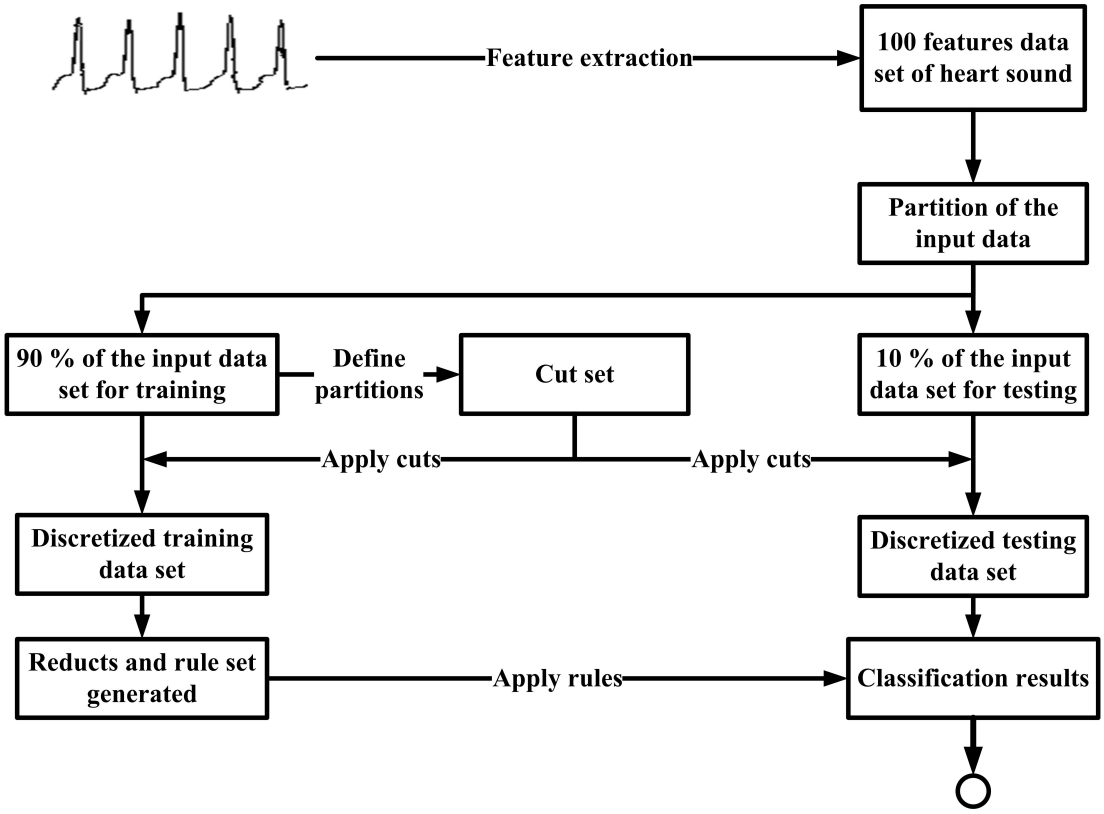


Figure 3.4: Rough set model - Model of extracting reducts and rule generation

rule generation, chimerge feature selection is applied on the training data set and the selected features are compared to the extracted reducts. It is required to show that the rules generate from the rough set are dependents on the attributes selected by features like chimerge and information gain methods. This could declare the reason that the resulted classification accuracy in rough set is not affected by the absence of feature selection techniques, which again declare the negative effect of discretization algorithms on the internal structure of the input data sets. The comparison between the output from classifiers and rough sets are shown in the experimental work in chapter 6.

3.4 Chapter conclusion

This chapter introduces several preprocessing techniques that are required to solve several problems. These problems arises from the requirement of machine learning techniques from the input data sets. Some of these preprocessing techniques could be avoided due to its negative effect on the classification accuracy as it causes a deterioration in the internal structure of the input data. The example presented here of such techniques is the PCA1 which required the data distribution to be normal (Gaussian). If the normalization process is avoided by using PCA2 that depends on the usage of standard deviation in its calculations, an increase in the classification accuracy appears. Other preprocessing feature selection techniques like chisquare and mutual information work better if the input data is in a discrete form which is available in real-life data sets like medical data. A proposed Interval-based feature selection technique that contains no presumptions on the input data is presented in order to reaches a better classification accuracy. This technique depends on a hypothesis that as overlap between values of different classes decreases as the importance of the corresponding feature increases. The results and the experimental analysis will be discussed in the experimental work chapter (6).

Chapter 4

Machine learning stage and proposed techniques

The selection of a specific machine learning technique is a critical problem that is recently under investigation in different researches. Different studies have applied to solve the problems of each technique according to the nature of the input data set. Solutions are either includes enhancements that have been applied to the technique itself or propose hybrid models to overcome the drawbacks of the technique. This chapter discusses the machine learning techniques briefly the drawbacks of each model from the point of view of data's nature. Then it discusses latest upgrades and enhancements applied to each technique.

4.1 Introduction

The learning techniques can be categorized into six main categories:

- Perceptron-based techniques like single and multi layered perceptron, Artificial neural network is another name for multi-layered perceptron.
- Logic-based learning machines like decision trees and rule-based classifiers
- Statistical-based learning machine techniques like naive bayes classifiers and bayesian networks.
- Instance-based learning machine techniques like K-Nearest neighbor approach.

- Set-based machine learning techniques like rough set and fuzzy set techniques.
- Kernel-based machine learning like the support vector machine.

Every machine learning technique has its own strong and weak points. Generally, SVMs and neural networks tend to perform much better when dealing with multidimensions and continuous features (53). On the other hand, logic-based systems like decision trees tend to perform better when dealing with discrete/categorical features. For neural network models and SVMs, a large sample size is required in order to achieve its maximum prediction accuracy whereas NB may need a relatively small data set. The main judge of which machine learning technique to select depends on the nature of the input data set, as selecting inappropriate algorithm may lead to either high processing cost or low classification accuracy. This chapter discusses the characteristics that defines the nature of the input data set.
The comparison among different classification techniques can be defined according to following criteria, where the most of these criteria are problem-dependent.

- Accuracy in general
- Speed of learning with respect to number of attributes and the number of instances
- Speed of classification
- Tolerance to missing values
- Tolerance to irrelevant attributes
- Tolerance to redundant attributes
- Tolerance to highly interdependent attributes (e.g. parity problems)
- Dealing with discrete/binary/continuous attributes
- Tolerance to noise values, or outliers
- Dealing with danger of overfitting
- Attempts for incremental learning
- Explanation ability/transparency of knowledge / classifications

- Model parameter handling

Another way of categorizing supervised learning whether they are either lazy or eager learning. In the case of Lazy learning, it simply stores training data (or only minor processing) and waits until it is given a test tuple. An example of lazy learning is the instance-based learning. On the other hand, eager learning techniques get a set of training set then construct a classification model before receiving new (e.g., test) data to classify. Example of eager techniques are decision trees, SVM and neural network.
The main goal of this chapter is to:

- Demonstrate the well known classification techniques and the strong and weak points in each one technique,
- Show the most recent researches applied to solve the problem either through a modification in the technique or through a hybrid model.

In several domains of interest such as in medicine, the training data often have characteristics that are not handled directly in conventional classification techniques. Many attempts have been made in the last decades to design hybrid systems for pattern classification by combining the merits of individual techniques (18). Also some recent approaches have adopted a semi-supervised model for classification. These approaches first apply unsupervised flat clustering algorithms, like the k-mean clustering, to cluster all instances in the training and testing data sets, and then use the resulting clustering solution to add additional instances to the training set (54). Both hybrid and semi-supervised models depends on classification algorithms and statistical algorithms that requires assumptions that may not exist in real-life and medicine data sets (55).
The selection of a classification technique is still a try and error manner problem. As a simplification for the above techniques the 4.1 is describing the most known techniques and the available hybridization techniques. The rest of this chapter is organized as follows: Section 4.2 demonstrate the machine learning techniques like perceptron, logic, statistical, instance-based and SVM machine learning techniques respectively. And also discuses the approaches introduced for enhancing such techniques. Section 4.3 introduces the proposed techniques to solve the problems appear due to the nature of medical data sets. While section 4.4 discusses a conclusion about these techniques.

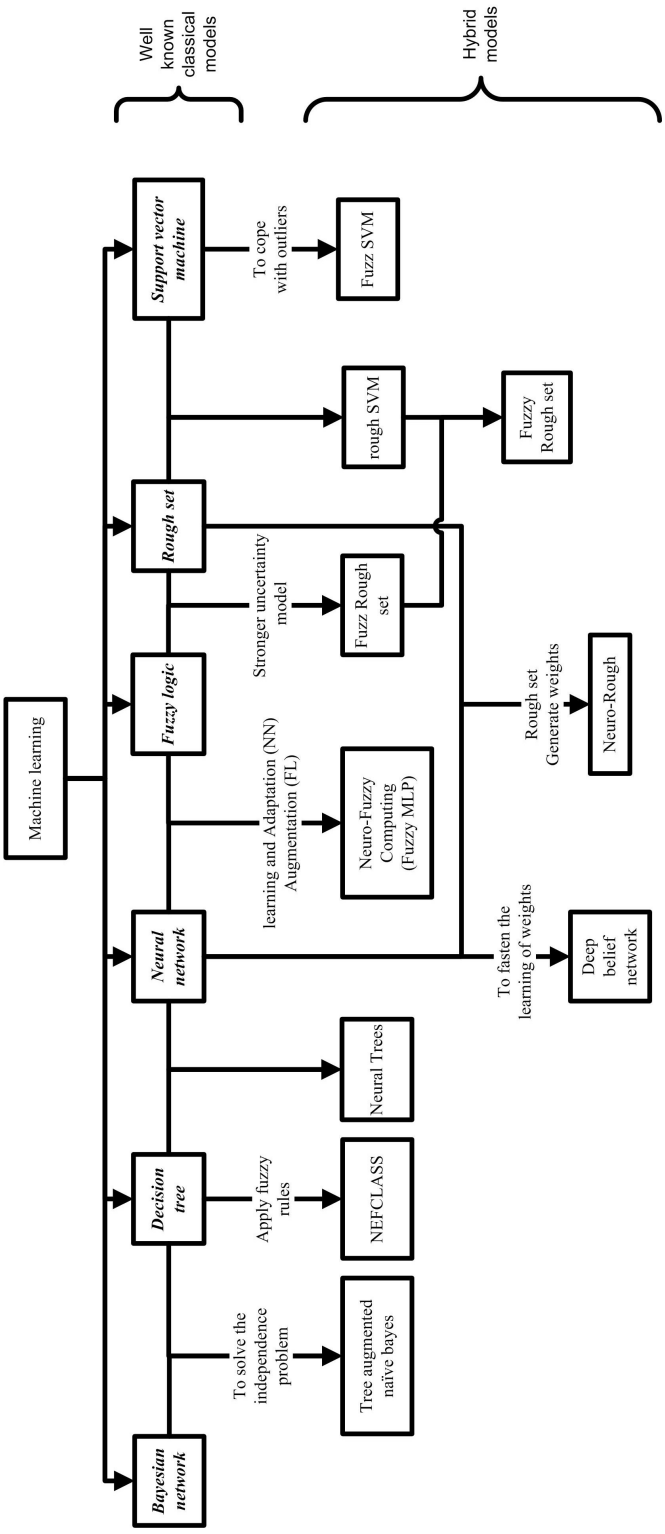


Figure 4.1: classification Techniques - classification Techniques

4.2 Conventional classification techniques and its applications

4.2.1 Perceptron-based machine learning techniques

4.2.1.1 Artificial neural network

Artificial Neural Networks (ANN) - or simply neural networks (NN) - are a learning algorithm and a non-linear statistical data modeling technique. A neural network is composed of interconnected groups of artificial neurons (a mathematical function) as shown in figure 4.2. It applies the processing information to model complex relationships between inputs and outputs, to find patterns in data or capture the statistical structure. Where the nonlinear generalization of the McCullogh-Pitts neuron:

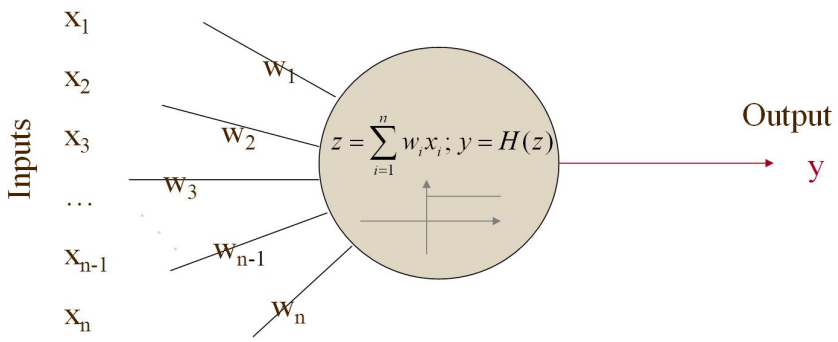


Figure 4.2: Artificial neurons - The McCullogh-Pitts model

$$y = f(x, w) \tag{4.1}$$

Where w represents the weights of the connectors and x represents the input while y represents the output. The function f could be Sigmoidal or Gaussian as in equations 4.2 and 4.3 respectively.

$$y = \frac{1}{e^{-w^T x - a}} \tag{4.2}$$

$$y = e^{-\frac{\|x-w\|^2}{2a^2}} \tag{4.3}$$

An ANN is composed of a set of neurons (perceptrons), connected by synapses, where each neuron's task is a basic yes/no decision. The most common types of neural networks consist of 3 layers: input - hidden - output. In the input layer, raw information

is fed into the network. Then the input is processed in the hidden layer, depending on activities of input units and weights on connections. The behavior of the output is dependant on hidden layer activities and weights. The hidden/processing layer can be extended to multiple layers, without feedback connections like an feed-forward networks, or with feedback connections like in recurrent networks. In the feed-forward networks, signals move one way only - from input to output - with no feedback. So, they are straight forward networks that associate inputs with outputs. In the feedback networks, signals can travel noth ways by using loops in the network. So, they are very powerful and can get more complicated. They are dynamic cause their states keep changing continuously until they reach the best point (stable state), via a relaxation process (56).
Types of Artificial Neural Networks: There are many types of neural networks, where each one is a computational simulation of a biological neural network model. Each type has different strengths particular to their applications, and they have different abilities related to their structure, dynamics, and learning techniques.

- Feedforward NN: a simple neural network with an input layer, zero or more hidden layers, and an output layer. It's the most common type, and can be constructed from different types of units.

- Radial Basis Function (RBF) NN: a feedforward network with an input layer, a hidden layer, and an output layer. The hidden layer is based on a radial basis function (usually the Gaussian), and it can employ more complex activation function that a typical feedforward. RBF networks are usually used for pattern recgnition applications.

- Kohonen Self-Organizing Network (Self-Organizing Maps): an unsupervised NN that contains two layers (input and output). Instead of taking the output of individual neurons, the neuron with the highest output os considered the winner and takes all strategy in the output layer.

- Learning Vector Quantization (LVQ): related to SOM (winner-take-all) learning strategy. LVQ systems can be applied to multi-class classification problems in a natural way.

- Recurrent NN: It's a neural network with bi-directional data flow for feedback, so it may contain loops.
- Modular NN: it's a bigger neural network, where the building blocks are modules, each module is a neural network in itself. The architecture of a single module is simpler, and the subnetworks are smaller than a monolithic network. Each module is connected to other modules rather than neurons, where they are independent to some level, so they can work in parallel.
- Physical NN: A network that simulates artificial synapses, represented in electrically adjustable resistance material.
- Others like Neuro-fuzzy networks, Holographics associative memroy, Instantaneously trained networks, Cascading NN.

4.2.1.2 Deep belief network

Neural networks suffer from the problem of weights initialization, also the number of forth and back propagation for wight initialization could increase the time of training. DBN should have a better a performance than the traditional neural network due the initialization of the connecting weights rather than just using random weights in NN (57). Deep Belief Network (DBN) is a deep architecture that consists of a stack of Restricted Boltzmann Machines (RBM). RBMs were originally developed using binary stochastic units for both the visible and hidden layers. A technique is proposed in (57) that depends on Deep Belief Network (DBN) in the clustering and classification of continuous input data without using the back propagation in the DBN architecture. This technique solve the problem of RVM, where RBMs were originally developed using binary stochastic units for both the visible and hidden layers.

Restricted Boltzmann Machine (RBM) is an energy-based undirected generative model that uses a layer of hidden variables to model a distribution over visible variables (58, 59). The undirected model for the interactions between the hidden and visible variables is used to ensure that the contribution of the likelihood term to the posterior over the hidden variables is approximately factorial which greatly facilitates

inference (60). Energy-based model means that the probability distribution over the variables of interest is defined through an energy function. It is composed from a set of observable variables $V = \{v_i\}$ and a set of hidden variables $H = \{h_j\}$, i node in the visible layer, j node in the hidden layer. It is restricted in the sense that there are no visible-visible or hidden-hidden connections. The steps of the RBM learning algorithm can be declared as follows:

1. Due to the conditional independence (no connection) between nodes in the same layer (Property in RBM), the conditional distributions are given in equations {4.4, 4.5, 4.6} and equations {4.7, 4.8, 4.9}.

$$P(H|V) = \prod_j p(h_j|v) \tag{4.4}$$

$$p(h_j = 1|v) = f(a_i + \sum_i w_{ij} v_i) \tag{4.5}$$

$$p(h_j = 0|v) = 1 - p(h_j = 1|v) \tag{4.6}$$

   And

$$P(H|V) = \prod_i p(v_i|h) \tag{4.7}$$

$$p(v_i = 1|h) = f(b_j + \sum_j w_{ij} h_j) \tag{4.8}$$

$$p(v_i = 0|h) = 1 - p(v_i = 1|h) \tag{4.9}$$

   Where f is a sigmoid function (σ) which takes the form $\sigma(z) = 1/1 + e^{-z}$ for binary data vector.

2. The likelihood distribution between hidden and visible units is defined as:

$$P(v, h) = \frac{e^{-E(v,h)}}{\Sigma_i e^{-E(v_i,h)}} \tag{4.10}$$

   Where $E(x, h)$=$-\bar{h}wv - \bar{b}v - \bar{c}h$,
   And $\bar{h}$,$\bar{b}$, $\bar{c}$ are the transposes of matrices h, b and c.

3. The average of the log likelihood with respect to the parameters is given by

$$\begin{aligned} \Delta w_{ij} &= \varepsilon^*(\delta \log p(v)/\delta w_{ij}) \\ &= \varepsilon(< x_i h_j >_{data} - < v_i h_j >_{model}) \end{aligned} \tag{4.11}$$

$$\Delta v_i = \varepsilon(< v_i^2 >_{data} - < v_i^2 >_{model}) \tag{4.12}$$

$$\Delta h_i = \varepsilon(< h_i^2 >_{data} - < h_i^2 >_{model}) \tag{4.13}$$

4. The term $<>_{model}$ takes exponential time to compute exactly so the Contrastive Divergence (CD) approximation to the gradient is used instead (61). Contrastive divergence is an algorithm that depends on the approximation that is to run the sampler for a single Gibbs iteration, instead until the chain converges. In this case the term $<>_1$ will be used such that it represents the expectation with respect to the distribution of samples from running the Gibbs sampler initialized at the data for one full step, the new update rule will be.

$$\Delta w_{ij} = \varepsilon(< v_i h_j >_{data} - < v_i h_j > 1) \tag{4.14}$$

$$\Delta v_i = \varepsilon(< v_i^2 >_{data} - < v_i^2 > 1) \tag{4.15}$$

$$\Delta h_i = \varepsilon(< h_i^2 >_{data} - < h_i^2 > 1) \tag{4.16}$$

The Harmonium RBM is an RBM with Gaussian continuous hidden nodes (61). Where f is normal distribution function which takes the following form:

$$P(h_j = h|x) = N(c_j + w_j.x, 1) \tag{4.17}$$

Harmonium RBM is used for a discrete output in the last layer of a deep belief network in classification.

Deep belief network architecture as shown in figure 4.3 is a deep architecture

that consists of a stack of Restricted Boltzmann Machines (RBM). RBM is an energy-based undirected generative model that uses a layer of hidden variables to model a distribution over visible variables (62). The undirected model for the interactions between the hidden and visible variables is used to ensure that the contribution of the likelihood term to the posterior over the hidden variables is approximately factorial which greatly facilitates inference (63). Energy-based model means that the probability distribution over the variables of interest is defined through an energy function. The key idea behind training a deep belief network by training a sequence of RBMs is

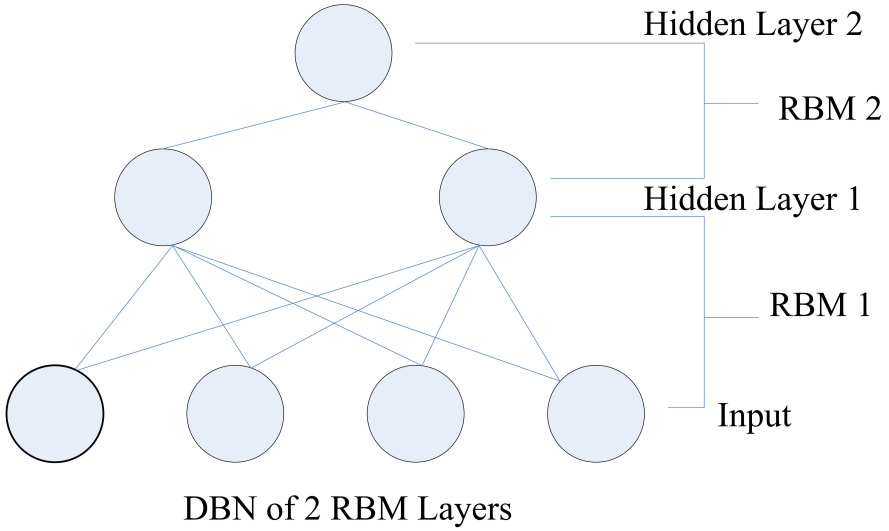


Figure 4.3: Deep belief network - Stack of RBMs

that the model parameters θ, learned by an RBM define both $p(v \mid h, \theta)$ and the prior distribution over hidden vectors, $p(h \mid \theta)$, so the probability of generating a visible vector, v, can be written as:

$$p(v) = \Sigma_h p(h \mid \theta).p(v \mid h, \theta) \tag{4.18}$$

After learning θ, $p(v \mid h, \theta)$ is kept while $p(h \mid \theta)$ can be replaced by a better model that is learned by treating the hidden activity vectors $H = h$ as the training data (visible layer) for another RBM. This replacement improves a variation lower bound on the probability of the training data under the composite model. The study in (64) proves the following three rules:

1. Once the number of hidden units in the top level crosses a threshold; the performance essentially flattens at around certain accuracy.

2. The performance tends to decrease as the number of layers increases.

3. The performance increases as we train each RBM for an increasing number of iterations.

In case of not using class labels and back-propagation in the DBN Architecture (unsupervised training) (65), DBN could be used as a feature extraction technique for dimensionality reduction. On the other hand, when associating class labels with feature vectors, DBN is used for classification. There are two general types of DBN classifier architectures which are the Back-Propagation DBN (BP-DBN) and the Associate Memory DBN (AM-DBN) (66). For both architectures, when the number of possible classes is very large and the distribution of frequencies for different classes is far from uniform, it may sometimes be advantageous to use a different encoding for the class targets than the standard one-of-K softmax encoding.

4.2.2 Logical-based machine learning techniques

4.2.2.1 Decision trees

A decision tree is a hierarchical model for supervised learning implementing the divide and conquer strategy whereby the local region is identified in a sequence of recursive splits in a smaller number of steps, which applied in many real-world applications as powerful solution to classification problem. Therefore, at the beginning, let us briefly summarize the basic principles of classification. in general classification is the process of assigning objects to one of several predefined class. example include detecting spam email message based the message sender and header.

How a decision tree works? Decision tree is an efficient non parametric method. In nonparametric method estimation, all we assume is that the similar input has similar outputs. This is a reasonable assumption therefore, our algorithm is composed of finding the similar past instance from the training set and interpolate from them to find the right output. A decision tree is a hierarchical structure that consists of nodes and directed edges. The tree has three types of nodes:

- A root node: Is the node that has no incoming edges and zero or more outgoing edges.
- Internal node: each of which has exactly one incoming edge and two or more outgoing edges.
- Leaf or terminal nodes: each of which has exactly one incoming edge and no outgoing edges.

Decision trees are either univariate or multivariate trees. In a nonlinear multivariate tree; a multilayer perceptron at each node divides the input space arbitrarily where this increases the complexity and the risk of over-fitting. In a linear multivariate tree, the problem of finding the optimal split at a node is NP-hard problem. This problem is investigatable and solved in the classification and regression trees (CART) algorithm (67).

Hybrid models applied on Decision trees : In order to solve the problem of high-dimensionality, parallel algorithms for constructing classification decision trees are desirable for dealing with large data sets in reasonable amount of time. (68) show how to parallelize C4.5 algorithm in three ways: (i) feature-based, (ii) node-based and (iii) data-based manner. Hybridization techniques have been applied between Decision trees and neural networks into two ways, first is to construct a decision tree and then convert the tree into a neural network, the other way is using neural networks as building blocks in decision trees (56). Also the neuron-fuzzy rule-based algorithm, namely NEFCLASS, is used in the construction of fuzzy decision trees and the association rule-based techniques. This algorithm gives better rules (means appropriate for each specific problem) and keep the level of interpretability and accuracy in the decision making tasks (69) (70).

4.2.3 Statistical-based machine learning techniques

4.2.3.1 Bayes Model

The optimum Bayesian classifier (in the sense that it minimizes the total misclassification error cost) is obtained by assigning to the example $x = (x_1, \ldots, x_n)$ the class with the highest posterior probability, i.e.

$$\gamma(x) = \arg\max_{c} p(c|x_1, ..., x_n) \tag{4.19}$$

Where according to naive Bayes model, in order to calculate this posterior probability, we have:

$$p(c|x) \propto p(c,x)\Pi_{i=1}{}^{n}p(x_i,c) \tag{4.20}$$

The estimation of the prior probability of the class, p(c), as well as the conditional probabilities p(xi—c), is performed based on the database of selected individuals in each generation (71). This Bayesian model has always the same structure: all variables X1 . . .Xn are considered to be conditionally independent given the value of the class value C. Figure 4.4 shows the structure that would be obtained in a problem with four variables. Tree augmented naive Bayes (Friedman et al., 1997) is another Bayesian

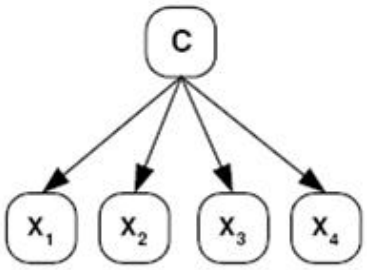


Figure 4.4: Naive Bayes model - Graphical structure of the naive Bayes model

network classifier in which the dependencies between variables other than C are also taken into account. These models represent the relationships between the variables X_1, . . ., X_n conditional on the class variable C by using a tree structure. Tree augmented naive Bayes (72) classifier put the dependencies between variables other than C are into account. It is built in a two-phase procedure.

- Firstly, the dependencies between the different variables X_1, . . ., X_n are learned. This algorithm uses a score based on information theory, and the weight of a branch (X_i,X_j) on a given Bayesian network S is defined by the mutual information measure conditional on the class variable as follows:

$$I(x_i,x_j) = \sum p(c)I(X_i,X_j|C=c) \tag{4.21}$$

  Which means that :

$$I(x_i,x_j) = \sum_{c}\sum_{x_i}\sum_{x_j} p(x_i,x_j,c)log\frac{p(x_i,x_j|c)}{p(x_i|c)p(x_j|c)} \tag{4.22}$$

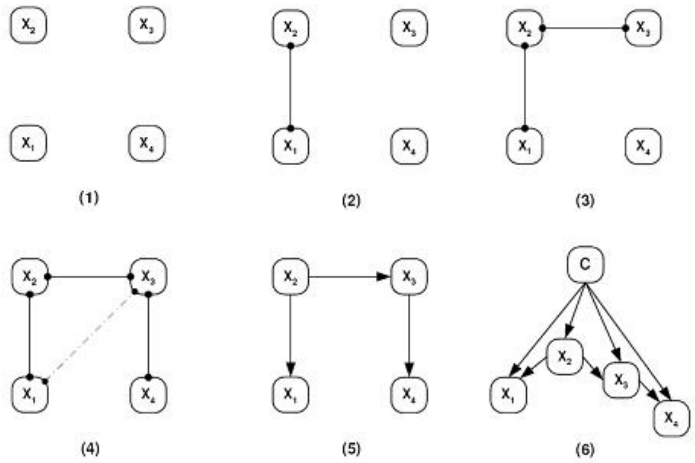


Figure 4.5: Tree augmented naive Bayes steps - Illustration of the steps for building a tree augmented naive Bayes classifier in a problem with four variables. X1,X2,X3,X4 are the predictor variables and C is the variable to be classified

With these conditional mutual information values the algorithm builds a tree structure.

- Secondly, the structure is augmented into the naive Bayes paradigm. Figure 4.4 shows an example of the application of the tree augmented naive Bayes algorithm. This figure assumes that I(X_1,X_2—C) > I(X_2,X_3—C) > I(X_1,X_3—C) > I(X_3,X_4—C) > I(X_2,X_4—C) > I(X_1,X_4—C). In figure 4.5 part (4), the branch (X_1,X_3) is rejected since it would form a loop. Here 4.5 part (6) is the result of the second phase of augmenting the tree structure. Following the tree augmented naive Bayes model, and using the classifier shown in figure 4.5, an individual x = (x_1, x_2, x_3, x_4) will be assigned to the class

$$c^* = \arg\max_c p(c)p(x_1|c,x_2)p(x_2,c)p(x_3|c,x_2)p(x_4|c,x_3) \tag{4.23}$$

The tree augmented naive Bayes algorithm follows a method that is analogous to filter approaches, where only pairwise dependencies are considered.

Particle swarm optimization (PSO)/Bayesian classifier is proposed by Devi in (73), where this classifier concluded that the PSO/Bayesian classifier obtains a promising accuracy. The hybrid algorithm of the combined particle swarm optimization and Bayesian classifier for classification is applied to aid in the prediction of solvation sites

in bio-medical domain. Several evolutionary techniques can optimize the coefficients of the Bayes-derived discriminant function. However, a particle swarm optimization technique using a new way of updating the velocity is employed to prove its effectiveness.

4.2.4 Kernel-based machine learning techniques

4.2.4.1 Support vector machine

Support vector machines outperform conventional classifiers especially when the number of training data is small and there is no overlap between classes. The basic idea behind SVM is to find a hyperplane in the input space, high dimensional feature space of instances, that separates the training data points with as big a margin as possible. SVM searches for the linear optimal separating hyperplane which is a "decision boundary" separating the tuples of one class from another. Most "important" training points are support vectors; they define the hyperplane. If Margin ρ of the separator is the distance between support vectors, the required is to maximize the margin for all point by minimizing $\Phi(w)$. Then the required is to find w and b for any $\alpha_i > 0$ such that 4.24 is minimized:

$$\Phi(w) = w^T w \tag{4.24}$$

Where w and b are calculated as follows:

$$w = \sum \alpha_i y_i x_i \tag{4.25}$$

$$b = y_k - \sum \alpha_i y_i x_i^T x_k \tag{4.26}$$

for all (x_i,y_i),i=1..n and $y_i(w^T x_i + b) \geq 1$

So the linear discriminant, classification, function relies on a dot product between the test point x and the support vectors x_i and it takes the form of equation 4.27 where w is not needed to be calculated explicitly:

$$f(x) = \sum \alpha_i y_i x_i^T x + b \tag{4.27}$$

For any $\alpha_i > 0$ where each non-zero α_i indicates that the corresponding x_i is a support vector.

A kernel function is defined as a function that corresponds to a dot product $x_i^T \cdot x$ appear in equation 4.25 of two feature vectors in some expanded feature space (74):

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j) \tag{4.28}$$

- In the case of Linear kernel,

$$K(x_i, x_j) = x_i^T x_j \tag{4.29}$$

- In the case of Polynomial kernel,

$$K(x_i, x_j) = (1 + x_i^T x_j)^P \tag{4.30}$$

- In the case of Gaussian (Radial-Basis Function (RBF)) kernel, $K(x_i, x_j) = exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$

- In the case of Sigmoid kernel,

$$K(x_i, x_j) = tanh(\beta_0 x_i^T x_j + \beta_1) \tag{4.31}$$

Soft margin SVMs have used to the solve the problem in dealing with noise in the definition of the hyperplane, the solution introduced uses slack variables to relax the constraints used in forming the hard margin SVMs (75). Then due to the lack of discarding corrupted data points by noise, the soft margin SVM is reformulated into fuzzy SVMs by assigning a membership to every training sample (76). Also to handle the problem of the over-fitting due to outliers, Rough sets have been applied to SVMs to develop rough margin based SVM (77). In fuzzy rough sets, a fuzzy similarity relation is employed to characterize the similarity of two objects and a dependency function is also employed to characterize the inconsistency between the conditional features and the decision labels (78), this concept is applied on SVM to improve the hard margin SVMs (79). Tuning SVMs by selecting a specific kernel and parameters is still try-and-see problem.

4.2.5 Instance-based machine learning techniques

4.2.5.1 K-Nearest neighbor technique

One of the most recent emerging techniques is multi-instance learning or multi-label (MIML)-KNN learning technique. This technique solves the problem that some instances like images, documents or genes could contain patches, paragraphs or sections respectively which belong to different class labels. In this case every instance is considered as a bag of sub-instances and every sub-instance has its own class label (80, 81). In MIML, each example is represented by multiple instances and at the same time associated with multiple labels. In KNN, the nearest neighbors are defined in the terms of Euclidean distances between two points, while in in order to define a distance between bags we need to characterize how the distance between two sets of instances could be measured. The Hausdorff distance provides such a metric function between subsets of a metric space. For two sets of points A=a_1,,a_m and B=b_1,,b_n, the Hausdorff distance is defined in eq 4.32:

$$H(A,B) = \max\{h(A,B), h(B,A)\} \tag{4.32}$$

where

$$h(A,B) = \max_{a\in A} \min_{b\in B} \|a-b\| \tag{4.33}$$

The Hausdorff distance is very sensitive to even a single outlying point of A or B. To increase the robustness of this distance with respect to noise, a modification is introduced to Hausdorff distance as follows:

$$h(A,B) = k_{th a\in A} \min_{b\in B} \|a-b\| \tag{4.34}$$

where k_{th} denotes the k-th ranked value.

4.2.6 Set-based machine learning techniques

4.2.6.1 Rough set classification techniques

Rough sets theory proposed by Pawlak is a new intelligent mathematical technique. It is based on the concept of approximation spaces and models of the sets and concepts (82). In rough sets theory, feature values of sample objects are collected in what are known as information tables. Rows of such a table correspond to objects and columns correspond to object features. Let $\mathcal{O}, \mathcal{F}$ denote a set of sample objects and a

set of functions representing object features, respectively. Assume that $B \subseteq \mathcal{F}, x \in \mathcal{O}$. Further, let $[x]_B$ denote:

$$[x]_B = \{y : x \sim_B y\} \tag{4.35}$$

Rough sets theory defines three regions based on the equivalent classes induced by the feature values: lower approximation $\underline{B}X$, upper approximation $\overline{B}X$ and boundary $BND_B(X)$. A lower approximation of a set X contains all equivalence classes $[x]_B$ that are proper subsets of X, and upper approximation $\overline{B}X$ contains all equivalence classes $[x]_B$ that have objects in common with X, while the boundary $BND_B(X)$ is the set $\overline{B}X \setminus \underline{B}X$, *i.e.*, the set of all objects in $\overline{B}X$ that are not contained in $\underline{B}X$. The approximation definition is clearly depicted in figure 4.6.

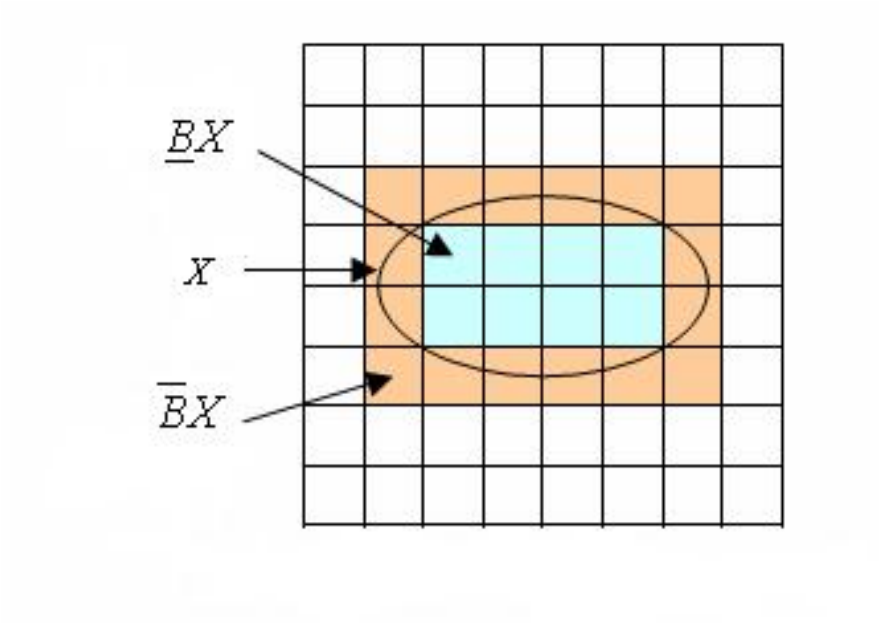


Figure 4.6: Rough boundary region - Rough boundary region

The indiscernibility relation $\sim_B$ is a fundamental principle of rough set theory. Informally, $\sim_B$ is a set of all objects that have matching descriptions. Based on the selection of B, $\sim_B$ is an equivalence relation partitions a set of objects $\mathcal{O}$ into equivalence classes. The set of all classes in a partition is denoted by $O/\sim_B$. The set $O/\sim_B$ is called the quotient set. Affinities between objects of interest in the set $X \subseteq \mathcal{O}$ and classes in a partition can be discovered by identifying those classes that have objects in

common with X. Approximation of the set X begins by determining which elementary sets $[x]_B \in \mathcal{O}/\sim_B$ are subsets of X (83).

The main advantage of rough set theory is that it does not need any preliminary or additional information about data: like probability in statistics or basic probability assignment in Dempster − Shafer theory, a grade of membership or the value of possibility in fuzzy set theory (84). Also rough set theory is very useful, especially in handling imprecise data and extracting relevant patterns from crude data for proper utilization of knowledge (85).

Several software systems based on rough set theory have been implemented in the areas of knowledge acquisition, pattern recognition, medicine, pharmacology, engineering, banking, market analysis, conflict analysis, environment, linguistics, image authentication and gene expression. Many real-life and non-trivial applications of this methodology have also been reported in the literature (86).

4.2.6.2 Fuzzy set theory (Fuzzy c-mean clustering)

A fuzzy set is a set whose elements have degrees of membership. A element of a fuzzy set can be full member (100% membership) or a partial member (between 0% and 100% membership). That is, the membership value assigned to an element is no longer restricted to just two values, but can be 0, 1 or any value in-between. Mathematical function which defines the degree of an element's membership in a fuzzy set is called membership function.

Fuzzification techniques have been applied in most of the machine learning techniques to provide a more human-like behavior and it shows a success in increasing the performance and accuracy of the classification results. Fuzzy logic introduces to the machine learning techniques a framework to deal with quantitatively, mathematically and logically with semantic and ambiguous concepts (87). The membership of data points in a set or class label is not crisp but can be specified as a degree of membership. The machine learning techniques under investigation in this thesis are c-mean clustering, where these techniques are fuzzified to generate fuzzy c-mean clustering, FCM, and fuzzy support vector machine respectively. In instance-based techniques like c-mean clustering technique, the fuzzy logic is used to determine the proximity of a given instance to the training set's instances (88). It allows data instances to belong to two or more clusters

where it is based on minimization of an objective or dissimilarity function (89), (90). The discreteness of each cluster makes the C-Means analytically and algorithmically intractable. Partitioning the data set in a fuzzy manner avoid this problem. The FCM clustering treats each cluster as a fuzzy set and each feature vector is assigned to multiple clusters with some degree of certainty measured by the membership function (10). The FCM optimizes the following objective function

$$J^m = \sum_{j:1..K} \sum_{i:1..N} \mu_{ji}^m * d_{ij} \tag{4.36}$$

Where: K is the number of clusters,
N is the number of objects,
m is the fuzziness controller which reduce the influence of small membership values, m is greater than 1,
d_{ij} represents the Euclidean distance between the object i of value x_i and the centroid of cluster j of value c_j which is represented by the following equation:

$$d_{ij} = \|x_i - c_j\| \tag{4.37}$$

and μ_{ji} denotes the membership of object i in cluster j; the values of μ_{ji} are initialized randomly such that the following condition is valid:

$$\sum_{j:1..K} \mu_{ji}^m = 1, i = 1, .., N \tag{4.38}$$

And c_j is the centroid of cluster j which it is calculated as follows:

$$c_j = \frac{\sum_{i:1..N} \mu_{ji}^m * x_i}{\sum_{i:1..N} \mu_{ji}^m} \tag{4.39}$$

After calculating the objective function J^m by equation 4.36, the value of μ_{ji} is updated by the following equation

$$\mu_{ji}^m = \frac{1}{\sum_{j:1..K} \frac{d_{ij}}{d_{ik}}} \tag{4.40}$$

The calculation of the J^m is repeated iteratively until the difference between the current J^m and the previous one is smaller than a certain threshold value ε such that:

$$J_{current}^m - J_{last}^m < \varepsilon \tag{4.41}$$

The correctness of equation 4.41 indicates the conversion of the objective function.
A popular measure used for calculating the distance between any object and the cluster center, is the Minkowski metric. This measure is required for high dimensional data and it is calculated as follows:

$$d_p(x_i, x_j) = (\sum_{k=1..D} \mid x_{ik} - x_{jk} \mid^p)^{\frac{1}{p}} \tag{4.42}$$

Where x_i and x_j are two objects in the input data set. D is the dimensionality of the input data set, i.e. number of attributes. p is a parameter that represents the type of Minkowski measure. The Euclidean distance (22) is a special case of this measure, where in this case, $p = 1$, while Manhattan metric is the case where $p = 2$. However, there are no general theoretical guidelines for selecting a measure for any given application.
By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the right location within a data set. But increasing the number of features may lead to ambiguity in training so that it would not even converge.
Most of the improvements in the fuzzy c-mean models are applied on the selection the most relevant or informative features. The benefits of feature selection (21) includes a lot of aspects like facilitating data visualization and data understanding, reducing the measurement and storage requirements, providing faster and more cost-effective classifiers, and defying the curse of dimensionality to improve prediction performance. The study in (6) has an addition is that the fuzzy feature clustering and feature selection are based on the interdependence among features using chi-square algorithms. Other feature selection techniques like in (20) has improved the MI formula for feature ranking which is used for feature selection in the pattern classification task. The problem of the current work in this area is that it stresses on the improvement of the feature selection techniques, while ignores the use of the feature ranks of the selected features in the classification models.
In (91), Particle swarm optimization-based Fuzzy Classification Systems is proposed, where each individual is represented to determine a fuzzy classification system. The individual is used to partition the input space so that the rule number and the premise part of the generated fuzzy classification system are determined. Subsequently, the consequent parameters of the corresponding fuzzy system are obtained by the premise

fuzzy sets of the generated fuzzy classification system.

4.3 Proposed Techniques

4.3.1 Pattern-based classification proposed technique

The transition from a clustering technique to a classification technique is a trend that could lead to successful results. An example of such trend is the Learning Vector Quantization (LVQ) algorithm which is based on a standard Self Organizing Maps. The standard LVQ has drawbacks like the instability behavior in the case of overlapped data and the strong dependence on the initial positions of the representatives (92). The proposed technique revealed this idea of transition from clustering to classification technique by using the frequent pattern-based clustering technique to build a new classification technique. The idea behind frequent pattern-based cluster analysis is that the frequent patterns discovered may indicate clusters. Frequent pattern mining can lead to the discovery of interesting associations and correlations among data objects. The proposed technique extracts all the frequent patterns that appear in objects of the same class A in the learning data set. If these frequent patterns are detected in an object in the testing data set this object will be classified as in class A. One of the drawbacks that appear in frequent pattern-based cluster technique is the dependent on a user defined threshold required to detect the frequent patterns (93). An inappropriate user defined threshold value may result in too many or too few patterns, with no coverage guarantees (22). In order to avoid this drawback in the proposed technique, the forward feature selection technique has been used to detect the patterns that produce the minimum classification error percentage.

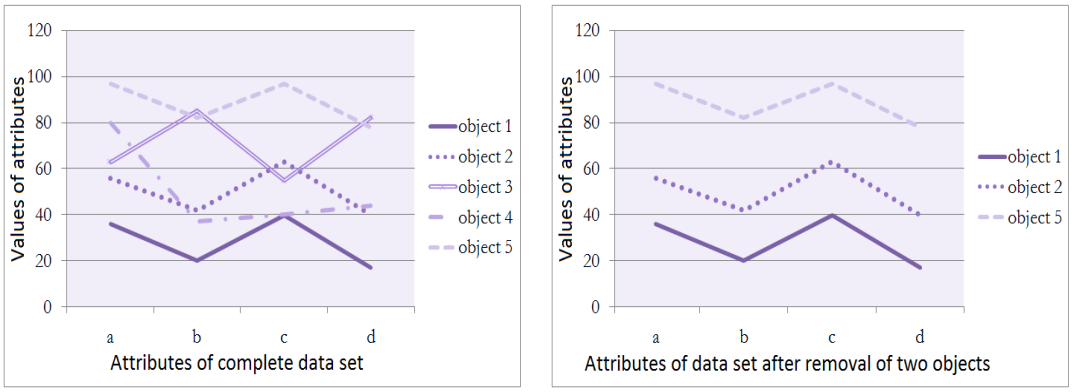
4.3.1.1 Pattern-based Subspace clustering technique

Pattern-based clustering is a kind of subspace clustering algorithm which extracts a subset of the input data set that have similar patterns. Most of the subspace-clustering algorithms rely on a certain distance function to capture the similarity among objects. In high dimensional space, the distance between any pair of objects is nearly the same. Whereas pattern-based clustering is effective in discovering such clusters (8, 27, 94).

To illustrate this clustering algorithm, we give an example in figure 4.3.1.1. Table 4.1 is a data set that consists of five objects with five attributes. figure 4.7(a) shows the values of the objects in full space (five attributes), where no obvious pattern is visible. However, if we just select attributes $\{A, B, D, E\}$ as in figure 4.7(b) for objects 2, 3, 5, we can observe the following pattern: for all the three objects, from attribute A to attributes B; D and E, the values first go down, and then up and finally down. We can assign these three objects into the same subspace cluster as they show similar pattern. Likewise, similar patterns may exist with other objects in other subspaces.

Table 4.1: Data set of 5 objects

Attribute	A	B	C	D	E
Obj1	80	36	55	38	42
Obj2	35	18	26	38	17
Obj3	98	84	45	100	80
Obj4	63	86	72	55	83
Obj5	56	40	50	63	40



(a) Set of five different object

(b) Set of objects of similar patterns (cluster)

Figure 4.7: An Example of pattern-based clustering

To tell whether two objects in D exhibit a coherent pattern in a given subspace S, it is essential to describe how coherent the objects are on these attributes. The following

definition serve this purpose.
Given two objects u, and $v \in D$, then there exists a coherent pattern between u and v in subspace S, if formula (1) and (2) are true.

$$\forall i, j \in S, d_{ij} = (u_i - u_j) - (v_i - v_j) \leq \delta \tag{4.43}$$

$$\forall i, j \notin S, d_{ij} = (u_i - u_j) - (v_i - v_j) > \delta \tag{4.44}$$

Subspace S is defined by the set of bounded dimensions (or subspaces), in which objects u and v have a similar shifting pattern. That is to say, if the rank of the two objects on two arbitrary attributes in S is less than a user-defined threshold "delta δ", then the two objects have a coherent pattern, as illustrated in figure 4.8. The minimal variation of object v on attributes i and j is Δ, while the maximal variation of u is $\Delta + \delta$, and the difference is less than δ. If all pairs of attributes in S satisfy this condition of variation, then objects u and v have coherent pattern.

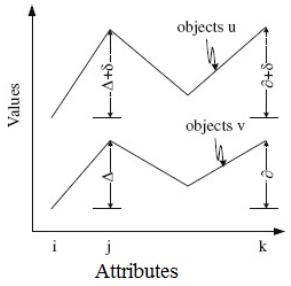


Figure 4.8: Pattern-Based Classifier - A coherent pattern between two objects

4.3.1.2 Pattern-based classification steps

The proposed Pattern-based classification technique is based on the frequent pattern-based subspace clustering technique. The definition of the clustering algorithm is to build clusters of objects that have common patterns, while in the proposed technique, the definition searches for the patterns that exist in objects of the same class. Each of these coherent patterns will be in the form of $[i, j]$, meaning that the change from feature i to feature j is the nearly similar for all objects in the same class. This technique handles the threshold user dependence problem that appears in the pattern-based

subspace clustering technique using an approach similar to the feed-forward feature selection approach. The proposed pattern-based classification technique is composed of four main methods: Extract Patterns, Validating Patterns, Get Best Patterns and Model Testing.
The model of the technique is shown in figure 4.9 where the training part in the input data set is divided into two subparts, The first one P_1 is to extract the patterns for each class in this part in the 'extract pattern' Method. While the second one P_2 is to test these extracted patterns in the validating patterns step. The best set of patterns is used later for testing and classification.

This model shows another addition that not considered before which is to find the pattern through different arbitrary lengthes of cross validation. This will allow to chose a single cross validation that produces the best classification accuracy and hence generate the patterns that produces this accuracy. Then use these patterns in the testing stage. The steps of each method will be discussed as follows:

- **Extract patterns:**
  This method is the training part of the technique. The input set of objects P_1 will be divided according to the assigned class label. Then this method extracts, from the given set of objects of each class, the set of coherent patterns through the modified form of the Pattern-based subspace cluster definition, where the δ_{ij} value for each $[i, j]$ is defined as follows:

  $$\forall u, v \in P_1, \delta_{ij} = max[|u_i - u_j| - |v_i - v_j|] \tag{4.45}$$

  The result of this method is a set of patterns for each class of the form of $[i, j]$, each of a specific distance δ_{ij}, excluded from the patterns those of inverse trends like when $u_i > u_j$ while $v_i < v_j$. Each array is sorted in an ascending order of δ_{ij} such that no $[i, j]$ pattern is repeated among different classes. Since the patterns are sorted in an ascending order, the first pattern should be the most important one. The steps of this method is discussed in algorithm 3

  The removal of misleading values in algorithm 4 is an optional step as it depends on the collection methodologies whether it is accurate or not. This step decreases the sensitivity to outliers by removing the values that are most far away from the average value of the attribute values. This step should remove only a small

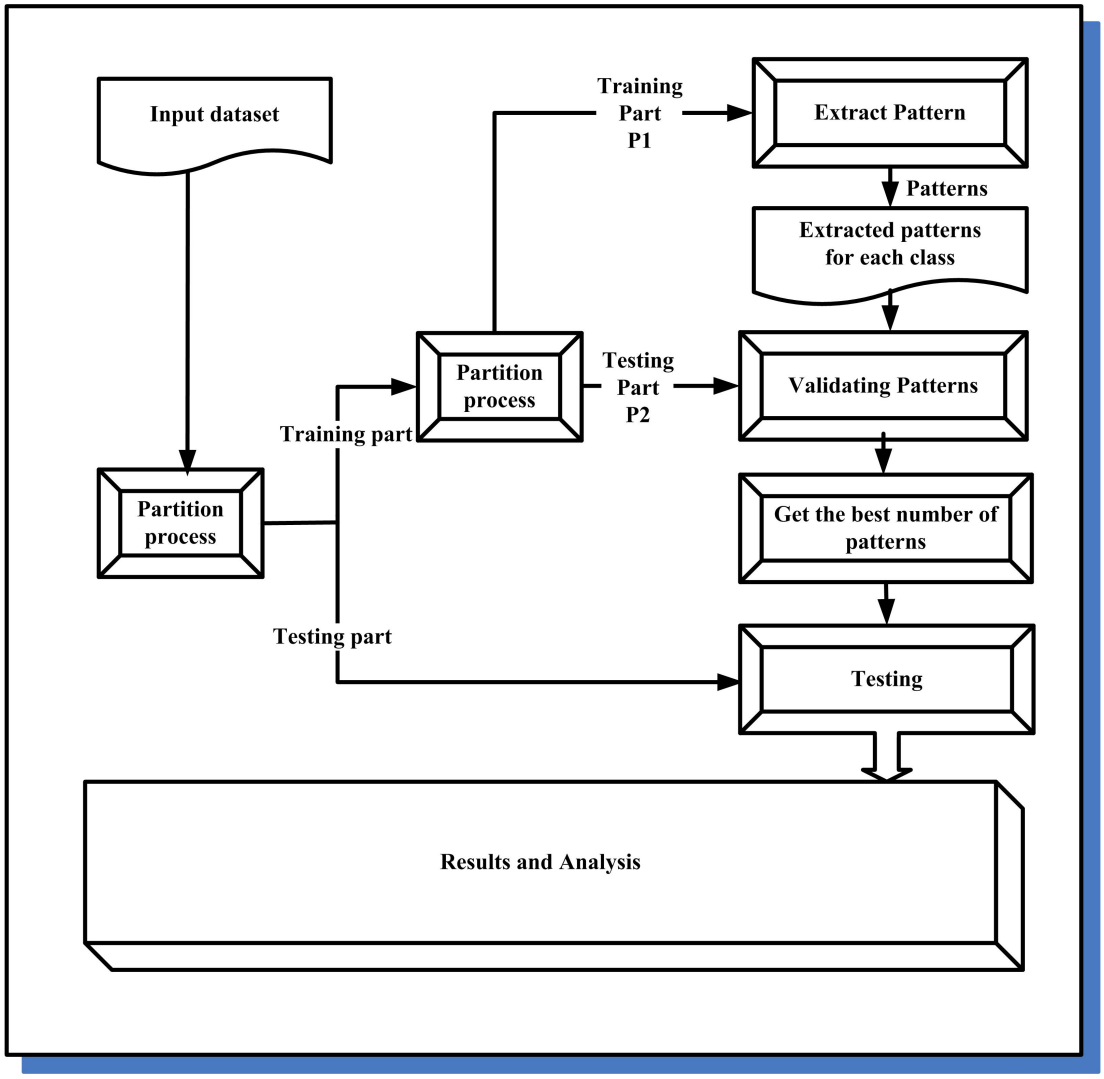


Figure 4.9: Pattern-Based Classifier - Model of the classifier

Algorithm 3 Extract-Patterns algorithm

Input:Data set P_1**;**
for every class c in the data set **do**
  Arr_c empty array for class c
  for every two distinct attributes i and j **do**
    Remove misleading values in attribute i for class c.
    Remove misleading values in attribute j for class c.
    for every two distinct objects u and v in class c **do**
      $d_{ij} = |u_i - u_j| - |v_i - v_j|$
      if $(u_i > u_j \; and \; v_i < v_j)$ **then**
        $d_{ij} = -1$ *and* exit for(u,v)
      end if
    end for
    δ_{ij} = maximum(d_{ij})
    add δ_{ij} to an array Arr_c
  end for
  Sort array Arr_c
  Remove δ_{ij} with -1 from Arr_c array
end for
Compare between the Arr_c array of all classes. For the common $[i, j]$ pattern among different classes, keep only the $[i, j]$ pattern of the minimum δ_{ij} value.
Return Arr_c for each class c

percentage of the values in the attribute in order not to affect the accuracy of the results.

- **Validate patterns:**

  This method tests the part P_2 of the training data set according to the patterns extracted in the 'extract patterns' method. The purpose of this method is to validate the extracted patterns according to their discrimination power among different classes. This method will use one of the objects in each class in the data set P_1 as reference with the objects in P_2 to detect whether there is a similarity between these object. The similarity will be detected according to the corresponding coherent patterns of that class. The result of this method is the

Algorithm 4 Remove percentage x of misleading

Input: $IntervalLength$ number of objects in class c in the training data set
Output: avg average of the values of an attribute i in a class c
for $x * IntervalLength$ values **do**
    Remove the value of max difference from the average avg.
end for

accuracy percentage of classifying objects. The steps of this method is discussed in algorithm 5.

- **Pattern Selection:**
  This method selects a subset of the patterns that produces the best classification accuracy. The Validate-Patterns method is performed iteratively for n times starting from n equals 1, where n is number of patterns of the class c such that Arr_c is of the minimum length. The test will be applied using the first pattern only in the sorted arrays, then again applied using the first two patterns, and repeated until all the n patterns are used. The result of this method is a subset of the sorted set of patterns that shows the maximum classification accuracy in testing P_2. For example, if there exist 2 classes, the first class has an array of patterns Arr_{c1} of length 3 while the second class has an array of patterns Arr_{c2} of length 4, then n will be equal 4. The test will be applied on the first pattern in the arrays Arr_{c1} and Arr_{c2}, then on the first 2 pattern and finally on the first 4 patterns. The subset of patterns array of the maximum accuracy will be returned. Figure 6.5 shows how the classification accuracy varies according to the number of patterns. The accuracy percentage in this figure resulted when applying the technique on. This way of selecting patterns is similar to the PCA feature extraction technique, where the first feature that corresponds to highest eigne value represents the most important feature. It is clear from the figure that the performance increases gradually until it reaches a beak then goes down again, also it is noticed that the chart may contain many local extrema besides the global maximum value (95).

- **Model Testing:**
  Finally test the technique after selecting the patterns for each class that highly

Algorithm 5 Validate-Patterns algorithm

Input: Arrays Arr_c **from algorithm (1);**
$className = noClass$;
$counter = 0$;
for every object u in the P_2 **do**
  for every class c in the data set **do**
    v_c: one object from P_1 of class c;
    for each pattern $[i, j]$ in array Arr_c **do**
      if $d_{ij} = (u_i - u_j) - (v_i - v_j) < \delta_{ij}$ **then**
        $className = c$
      end if
    end for
  end for
  if more than one class satisfies the conditions **then**
    For each class calculate the average of $\frac{d_{ij}}{\delta_{ij}}$ values.
    $className$ = class of the lowest average value.
  end if
  if $className$ = c is the correct class **then**
    $counter = counter$+1
  end if
end for
accuracy percentage = $counter$/(Number of objects in P_2)

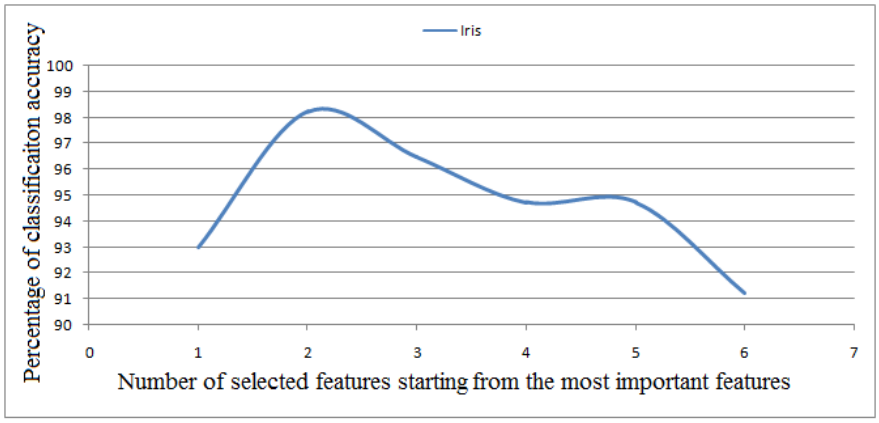


Figure 4.10: Pattern-Based Classifier - Iris classification performance according to the number of patterns

discriminate between objects accordingly. The steps of this method is the same as the testing-patterns phase as the patterns. Then the classification of the testing part of the objects will be performed, and the classification results is returned.

4.3.2 Fuzzification of Euclidean space in machine learning techniques

In the fuzzy c-means, the centroid of a cluster is computed as being the mean of all points, weighted by their degree of belonging to the cluster according to their proximity in feature space. The degree of being in a certain cluster is related to the inverse of the distance to the cluster. Support vector machine is a non-linear binary classification algorithm which is based on the theory of structural risk minimization (SVM). SVM is able to solve complex classification tasks without suffering from over-fitting problems that may affect other classification algorithms. Computationally speaking, the SVM training problem is a convex quadratic programming problem, meaning that local minima are not a problem (96). In Fuzzy SVM, the fuzzy membership values are calculated based on the distribution of training vectors were the outliers being given proportionally smaller membership values than other training vectors (97) (98).
The problem in such fuzzified techniques is that it applies the fuzzy logic concept on the level of objects and ignores the features composing such objects. Each object has a degree of membership to the class labels in the learning problem. For multivariate objects, the features of the objects have different relevance degrees to the target class labels. Feature selection techniques like chimerge are applied on the input data sets to select the features of the highest degree of relevance. Feature selection is a type of feature reduction technique that are required to reduce the number of feature to the minimum. This is applied either by selecting the best features or extracting a lower number of features from the higher ones. The resulted data set from this technique contains only the relevant and informative feature to the classification problem, while ignoring irrelevant features (95). Consequently, applying the classifier on the resulted and reduced features should show a better performance. The problem that appears in such way is that it deals with features in a crisp manner, either the features are selected or not. While these selected features have different degrees of importance that may enhance the classification accuracy results if taken into consideration.
The proposed technique includes such degree of importance inside the machine learning techniques, especially to the techniques like FCM and SVM (29). These techniques

depends on Euclidean calculations between data points in the space. For high dimensional data sets, a popular measure is used for calculating the distance which is the Minkowski metric (99), Euclidean is a special case of such equation. Most of the existing kernels employed in linear-nonlinear SVMs measure the similarity between a pair of data instances-based on the Euclidean inner product or the Euclidean distance of corresponding input instances. The calculation of the Euclidean distance or product ignores the degree of relevance of each feature to the classification problem and treats all the features equally. Then the required is that the fuzziness concept should be lowered from the level of data point membership degree to a specific set to the level of the feature membership degree to the classification problem. To apply this, the crisp dot product between data points in the Euclidean distance calculation is transformed to a fuzzy dot product through multiplying the dot product of each feature to the membership function of the corresponding features. The feature ranks are extracted through a feature selection and ranking technique named Chimerge that calculate the χ^2 value of the features in the input data set (100) (101). The resulted ranks from the chimerge technique will be considered as the membership degrees of the corresponding features, where this step is considered as hybridization between the feature selection technique and these classification techniques.

Neuro-fuzzy hybridization is the most visible integration realized so far. Fuzzy Set theoretic techniques try to mimic human reasoning and the capability of handling uncertainty - (SW). Neural Network techniques attempt to emulate architecture and information representation scheme of human brain - (HW). In other words ANN used for learning and Adaptation and Fuzzy Sets used to augment its Application domain. Rough sets and Fuzzy sets can be integrated to develop a technique of uncertainty stronger than either. While in Rough-Fuzzy Hybridization, the fuzzy Set theory assigns to each object a degree of belongingness (membership) to represent an imprecise/vague concept, and the rough set theory focus on the ambiguity caused by limited discernibility of objects (lower and upper approximation of concept) (102). Neuro-Rough Hybridization where networks consisting of rough neurons. The rough set techniques are used to generate network parameters (weights) (103).

4.3.2.1 Problem definition of Euclidean calculations

As discussed previously, the fuzzy logic concept on the SVM and FCM is applied on the level of objects (104). The proposed here in this technique is to apply such concept on the level of features as shown in figure 4.11. The reason of such enhancement is that features has a degree of relevance to the target class label, where this relevance is not a crisp relation. In the calculations of the distance each object to the centroid in the

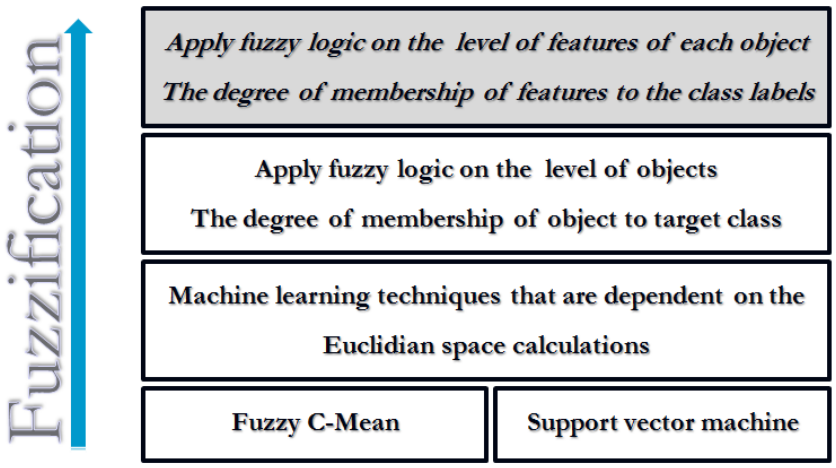


Figure 4.11: Levels of fuzzification - Attributes versus Objects

fuzzy c-mean classifier technique or the dot product in the kernel calculation of SVM, all the features are used in the calculation. Machine learning techniques have used feature selection techniques to eliminate the features that are not relevant to target class labels. After feature selection is applied, the data will be ready for training and testing by the classifier selected. The problem of this procedure is that not all features, even the selected features, are as important as each other to the classifier. In order to solve this problem, the chi2-square ranking technique is used to give a percentage of importance to each feature. Logistic regression is used to rank the features based on the χ^2 values (105) resulted from the chimerge technique. These ranks are processed such that these values range from 0 to 1 and the sum of all values are 1. Then it uses the percentage that corresponds to each feature in the calculation of the Euclidean distance or Euclidean product.

4.3.2.2 Attributes' Rank calculation

Let r_k represents the rank of attribute k, x_i represents the object i and c_j represents the centroid of class j.
The calculation of the rank values r_k will be as follows:

- First, the values that are resulted from the feature selection techniques are saved, these values are the χ^2 values. Then, the degree of importance of each feature is calculated according to the sorting of χ^2 values in an increasing order. The equation of the degree of importance will be in the form:

$$d_k = \frac{D - o_k}{\sum_{k=1..D} o_k} \tag{4.46}$$

  Where o_k represent the order of the attribute from the chi-square technique, for example, if o_k=1, this means that the attribute k is the best attribute, while if o_k= D, it means that attribute k is the worst attribute.

- Then the χ^2 value of each feature k multiplied by d_k is divided by the sum of χ^2 values of all feature, and the result is assigned to the rank value r_k. So, r_k will be calculated as shown in the following equation:

$$r_k = \frac{\chi_k^2 * d_k}{\sum_{i=1..D} \chi_i^2} \tag{4.47}$$

  In the case when the data set produces all χ^2 values have zero value, so equation 4.47 should be replaced by the values resulted in equation 4.46.

The rank value r_k will represent the membership of feature k to the classification problem. The technique of such change can be represented as shown in figure 4.12. The input from the training data set is evaluated by the chimerge technique then the χ^2 values are adjusted as in equation 4.47, and this input is introduced to classifier for training. The evaluation from the chi2-square technique will be used in the Euclidean calculation.

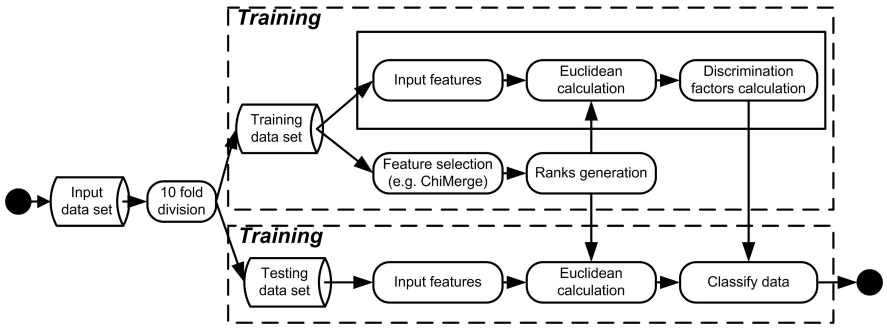


Figure 4.12: Improved machine learning techniques - FCM and SVM learning techniques

4.3.2.3 Using fuzziness in Euclidean calculations

The complete steps of the enhanced FCM technique as shown in figure 4.12, in the steps of the technique, scaling of the input data is important as it showed a better results in case of the scaled data than unscaled one.

- **FCM modification**

  The rank of membership value mentioned in equation 4.47 will be multiplied by the difference between the values of each object and the centroid corresponding to this attribute. So equation 4.48 will be adjusted to be as follows: The equation of Euclidean Distance:

$$d_p(x_i, x_j) = (\sum_{k=1..D} \mid x_{ik} - x_{jk} \mid^p)^{\frac{1}{p}} \tag{4.48}$$

  Where x_i and x_j are two objects in the input data set. And D is the dimensionality of the input data set, i.e. number of attributes. p is a parameter that represents the type of Minkowski measure. The adjusted equation of Euclidean Distance:

$$d_p(x_i, c_j) = (\sum_{k=1..D} r_k * \mid x_{ik} - c_{jk} \mid^p)^{\frac{1}{p}} \tag{4.49}$$

  In figure 4.12, the evaluation from the chi2-square technique will be used in the calculation of the distance to the centroid. When the termination condition is reached; either by convergence or elapsed number of iterations, the calculated centroids and the adjusted χ^2 values will be used in the testing of the classifier.

- **SVM modification**

  Again the rank of membership value mentioned in equation 4.47 will be multiplied by the dot product of the values of each object and the support vectors corresponding to this attribute. So equations 4.29, 4.30 and 4.31 will be adjusted to be as follows:

  - In the case of Linear kernel,

$$K(x_i, x_j) = \prod_k r_k * x_{ik} * x_{jk} \tag{4.50}$$

  - In the case of Polynomial kernel,

$$K(x_i, x_j) = (1 + \prod_k r_k * x_{ik} * x_{jk})^P \tag{4.51}$$

  - In the case of Sigmoid kernel,

$$K(x_i, x_j) = tanh(\beta_0 \prod_k r_k * x_{ik} * x_{jk} + \beta_1) \tag{4.52}$$

4.4 Chapter conclusion

There is an extremely large number of literatures on machine learning techniques, there is yet no clear picture of which technique is better. The reason of such problem is that most of techniques depends on presumptions that are not existing in real-life, medical data sets. As discussed the main problems that are not yet fully solved is the presence of continuous data, curst of dimensionality and the random distribution of the input data sets. In this chapter, two proposed techniques are discussed that are aiming to solve such problems. First, the pattern-based classification technique handles such problems based on the pattern-based clustering technique. This technique does not assume the discreetness of values in the input data and have a good tolerance to the outliers and finally ignore the irrelevant dimensions or features. The second technique uses the ranks of each feature resulted from feature evaluation methods like chimerge and Mutual Information in the calculations of the Euclidean distance as a main parameter in the fuzzy c-mean model. This introduced modification in the fuzzy c-mean model enhanced the performance of the fuzzy c-mean model and increases its capability in handling real-life medical data sets.

Chapter 5

Visualization stage using Formal Concept Analysis and a proposed technique

Representation and visualization of continuous data using the Formal Concept Analysis (FCA) became an important requirement in real-life medical fields. In the medical field, visualization makes it easier for doctors to find the relations and led them to find reasonable results. To apply FCA on numerical data, a scaling procedure should be firstly applied on its attributes. The scaling procedure, as a preprocessing stage for FCA, increases the number of attributes. Hence, it increases the complexity of computation and density of the generated lattice. This chapter introduces a modified modeling technique that uses the chimerge algorithm in the binarization of the input data. The resulted binary data is then passed to the FCA for formal concept lattice generation. The introduced technique applies also a validation algorithm on the generated lattice that is based on the evaluation of each attribute according to the objects of its extent set. To prove the validity of the introduced model, the technique is applied on data sets in the medical field and these data sets show the generation of valid lattices.

5.1 Introduction

Formal Concept Analysis (FCA) is one of the data mining research methods and it has been applied in many fields as medicine. FCA was introduced to study how objects

can be hierarchically grouped together according to their common attributes. This technique is of a great interest for mining association rules in medical data, specially the numerical ones. The basic structure of FCA is the formal context which is a binary-relation between a set of objects and a set of attributes. The formal context is based on the ordinary set, whose elements has one of two values, 0 or 1 (23), (106). A context materializes a set of individuals called objects, a set of properties called attributes, and a binary relation usually represented by a binary table relating objects to attributes. These mappings are called Galois connections or concepts. Such concepts are ordered in FCA within a lattice structure called *conceptlattice* within the FCA. Concept lattices can be represented by diagrams giving clear visualization of classes of objects in each domain. At the same time, the edges of these diagrams give essential knowledge about objects, by introducing association rules between attributes which describe the objects (107). Mostly, the real-world data are not available as binary data. Such data could be either numerical or categorical. To represent a numerical or a categorical data in the form of a formal context, such data should be transformed using conceptual scaling. In FCA, the attribute of numerical values are discretized, then each interval of entry values have to be considered as binary attributes (108). The transformation of such data, i.e. conceptual scaling, allows one to apply FCA techniques. Such procedure may dramatically increase the complexity of computation and representation. Hence, it worsens the visualization of results. This scaling may produce large and dense binary data (24), which are hard to process with the existing FCA algorithms. As it is based on arbitrary choices, the data may be scaled in a lot of different ways that lead to different results. Its interpretations could lead also to classification problems. The study in (109) proposed a scalable lattice-based algorithm ScalingNextClosure to decompose the search space for finding formal concepts in large data sets into partitions and then generate independently concepts (or closed item sets) in each partition.

This chapter replaces the scaling technique by a technique that uses of the chimerge algorithm into binarization of the numerical data attributes and into the validation of the generated formal concept-lattice. The binarization technique is applied using the chimerge algorithm through discretizing the continuous attribute values into only of two values, 0 or 1. Then the resulted binary table is used in generation of the formal concept lattice. Then the chimerge technique here is used to validate this generated lattice. For continuous data sets, the chimerge technique is used to automatically select

proper Chi-square χ^2 values to evaluate the worth of each attribute(110), (111) with respect to the corresponding classes. These χ^2 values are used to compare value of each attribute. Such values were calculated according to a novel formula-based on the generated formal concept lattice. If both evaluations are matched, then the generated lattice is considered to be representing the actual structure of the data. Hence, the binarization algorithm does not corrupt the generated lattice. Fnally, it led to a valid lattice. The conceptual computation and the lattice visualization are performed using a tool for formal concept lattice generation named Conflexplore (112). The introduced technique including the binarization, the visualization and the validation methods is applied on two data sets in the medical field from the UCI database; the Indian Diabetes data set and the Breast Cancer data set. The rest of this chapter is organized as follows: Sections 5.2 and 5.3 give an overview about the formal concept analysis technique. Then section 5.4 shows the proposed model, while section 5.5 shows the conclusion.

5.2 Formal Concept Analysis

FCA is based on a mathematical order theory for data analysis, which extracts concepts and builds a conceptual hierarchy from given data which is represented (113) with a formal context k as follows:

$$K := (G, M, I) \tag{5.1}$$

where K consists of two finite sets of objects G and attributes M, and a binary-relation I between the objects and the attributes(i.e., $I \subseteq (GXM)$). A relationship (g,m) $\in I$ means object $g \in G$ has attribute $m \in M$. The formal context can be easily represented by a cross-table as shown in table 5.1. In this example, the header of columns is an attribute as $M = \{a, b, c, d\}$, and the header of rows is an object as $G = \{O_1, O_2, O_3, O_4\}$. The binary-relation I is represented by putting "X" in the cross-table. For example, object "O_1" has an attribute "c". A formal concept is a pair(A, B), which is combination of a subset A of objects and a subset B of attributes. The set A is called the extent and the set B called the intent of the concept (A, B). The extent and the intent are derived by two functions, which are defined as:

$$intent(A) = \{m \in M | \forall g \in A : (g, m) \in I\}, A \subseteq G, \tag{5.2}$$

$$extent(B) = \{g \in G | \forall m \in B : (g, m) \in I\}, B \subseteq M. \tag{5.3}$$

A formal concept is defined as a pair(A, B) with $A \subseteq G, B \subseteq M$, intent($A$)=$B$ and extent(B) = A. From table 5.1, intent of $\{O_2, O_3, O_4\}$ is a, b and extent of {a,b} is $\{O_2, O_3, O_4\}$, i.e., $(\{O_2, O_3, O_4\}, \{a, b\})$ is a formal concept. Table 5.2 represents a list of all formal concepts, which are extracted from table 5.1. The concepts are partially ordered by super-sub relation which is formalized by

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 (\Leftrightarrow B_2 \subseteq B_1) \tag{5.4}$$

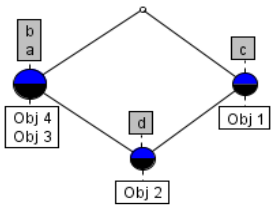
A context K is a set $B(C)$ of all formal concepts of K with the partial order $\leq$, denoted as $\iota := (B(C), \leq)$. In Fig 2, a formal concept C_3 $(\{O_2, O_3, O_4\}, \{a, b\})$ is super-concept of C_1 $(\{O_2\}, \{a, b, c, d\})$.

Table 5.1: Input data for lattice represenation

	a	b	c	d
O_1			X	
O_2	X	X	X	X
O_3	X	X		
O_4	X	X		

Table 5.2: The formal concepts and the corresponding Formal Concept Lattice

	Extents	Intents%
C_1	$\{O_2\}$	$\{a,b,c,d\}$
C_2	$\{O_1,O_2\}$	$\{c\}$
C_3	$\{O_2,O_3,O_4\}$	$\{a,b\}$
C_4	$\{O_1,O_2,O_3,O_4\}$	$\{\}$



Usually the attributes of a real-life data set are not in a binary form, attributes could be expressed in many-valued forms that are either discrete or continuous values. In that case the many-valued context will take the form (G, M, V, I) which is composed

of a set G of objects, a set M of attributes, a set V of attribute values and a ternary-relation I between G, M and V. Then the many-valued context of each attribute is transformed to a formal concepts, the process of creating single-valued contexts from a many-valued data set is called conceptual scaling. The process of creating a conceptual scale must be performed by using expert knowledge from the domain from which the data is drawn. Often these conceptual scales are created by hand, along with their concept lattice, since they are represented by formal contexts often layed out by hand. Such that we choose a threshold t for each many-valued attribute and replace it by the two one-valued attributes "expression value of $g \leq tg$" and "expression value of $g > t$". The threshold value t must be chosen specifically for each many-valued attribute. For instance, table 5.3 contains two different data types, the first attributes is of discrete values and the other is of continuous values.

Table 5.3: Multi-valued context data set

Fruit	color	price	target
apple	yellow	3.1	+
grapefruit	yellow	2.1	+
kiwi	green	1.2	+
plum	blue	3.1	+
toycube	green	6.2	-
egg	white	0.5	-
tennisball	white	3.3	-

This multi-valued context data set can be reduced to a context of the form presented above by scaling as in the table 5.4:

Here, the abbreviation for white is "w", for yellow is "y", for green is "g" and for blue is "b". Also the price values that are smaller than 3 are abbreviated by <3 and the values greater than or equal 3 are abbreviated by ≥ 3.

5.3 Related work

The only way to apply FCA is to binarize the data, and the only way to do so is through the scaling procedure. There are two main problems that faces the scaling

Table 5.4: Scaled context data set

Fruit	w	y	g	b	<3	≥3	target
apple		X				X	+
grapefruit		X			X		+
kiwi			X		X		+
plum				X		X	+
toycube			X			X	-
egg	X				X		-
tennisball	X					X	-

procedure, the first problem is that the discretization algorithm of the data attributes if these attributes are continuous that may imply loss of information, and the second problem is that the scaling algorithm produces a large and dense binary data known as hard to process. The scaling in the previous section is defined as interval-based scaling where interval number and size are chosen by experts and they are hard to determine adequately a priori.

5.3.1 Interordinal scaling

Another way of scaling is the Inter-ordinal scaling that is defined in (114) where it describes all intervals without loss of information. The objective of the Inter-ordinal scaling is to extract the formal concepts (A, B) where A is a subset of objects sharing similar values, i.e. lying in a same interval. Such that all the possible intervals of attribute values can be represented as follows:

$$I_{W_s} = (W_s, W_s, \leq) \mid (W_s, W_s, \geq) \tag{5.5}$$

W_s is the set of all values of attribute s, and the operation of apposition "|" of two identical sets of objects returns a set of attributes, where this set of attributes represents the disjoint union of attribute sets of the original contexts. To give an example, let W_{s_1} $= \{4, 5, 6\}$, then attribute s_1 is replaced by $\{s_1 \leq 4, s_1 \leq 5, s_1 \leq 6, s_14, s_1 \geq 5, s_1 \geq 6\}$. The inter-ordinal scaling creates $2|W_s|$-1 binary attribute for each many-values attribute s; and this makes the derived context dense, large and difficult to process, and on the other hand these binary data show better computational properties.

5.3.2 Pattern concept lattice

Another scaling algorithm defined in which is based on the idea of pattern structures (115). The similarity of two sets of labeled graphs is realized as the similarity between two real numbers (116), between two intervals, where it may be expressed in the fact that they lie within some larger interval, this interval being the smallest interval containing both two. Where the meet of two intervals [a_1,b_1] and [a_2,b_2], with a_1, b_1, a_2, $b_2 \in \Re$, is defined as follows:

$$[a_1, b_1] \sqcap [a_2, b_2] = [min(a_1, a_2), max(b_1, b_2)] \tag{5.6}$$

The definition of $\sqcap$ implies that the meet of several intervals, 2 or more, is the smallest interval containing all intervals, i.e. $[4, 5] = [4, 4] \sqcap [5, 5]$. If each object is described as a p-dimensional vector of intervals, where p is the number of attributes, then all the corresponding intervals are subsumed using $\sqcap$. Finally the set of all pattern concepts gives rise to a pattern concept lattice. To give an example, the following table 5.5 represents five objects. Each of these objects contains 3 attributes. Then the meet of

Table 5.5: Multi-valued context data set

	s_1	s_2	s_3
g_1	5	7	6
g_2	6	8	4
g_3	4	8	5
g_4	4	9	8
g_5	5	8	5

the two objects, 3-dimensional vectors, is $\{g_1, g_2\}^\diamond$ such that:
$\{g_1, g_2\}^\diamond = \langle[5, 5], [7, 7], [6, 6]\rangle \sqcap \langle[6, 6], [8, 8], [4, 4]\rangle$
$= \langle[5, 6], [7, 8], [4, 6]\rangle$
Then g_1 and g_2 belongs to $\langle[5, 6], [7, 8], [4, 6]\rangle$, also g_5 belongs to the same set, the pair $(\{g_1, g_2, g_5\}, \langle[5, 6], [7, 8], [4, 6]\rangle)$ is a pattern concept. Then the set of all pattern concepts gives rise to a pattern concept lattice. This allows to use concept lattices for their knowledge representation and reasoning abilities without transforming data. The problem rises in this technique is that too many repeated patterns are generated.
Another technique is defined in (117) where each relation between objects and attributes

is represented by a membership value. A confidence threshold is defined to eliminate some relations. These relations are out of an interval of values of this threshold from a given fuzzy context. The confidence can be set by user according to the application or the domain knowledge.

5.4 Binarization and validation proposed technique

The required in this proposed model is to generate a binary data that represents the structure of the original data to be available to form a formal concept lattice. On the other hand, a validation on the generated formal concept lattice is applied depending on the evaluation of features. These two steps should avoid the two main problems discussed in the previous sections, which are the loss of information due to scaling of the input and the generation of a dense lattice. Where if the number of attributes is n and the average number of scaled features for each attribute is m, then the order (big O) of the algorithm required to generate lattice after applying scaling algorithm is $O(n * m)$. While the order (big O) of the algorithm required to generate lattice after applying the proposed technique is $O(n)$ which is considered as an enhancement in the performance of the generation of the lattice, lessen the density of the generated lattice and enhancement in understandability of such formal concept lattice.
The feature evaluation could be applied using any feature selection technique, here the chimerge technique will be used to declare the important features and rank them accordingly. And also the chimerge is used to discretize the input values of each attribute until only two discrete values are left, so this discretization process will be named binarization. The steps of visualization is shown in figure 5.1 where scaling is replaced by binarization and a validation step is added.

5.4.1 Binarization of the input

In this step, the chimerge technique is applied on each attribute from the input data, which is considered as any type of numerical data, either continuous or discrete input. In the binarization technique, the merging of intervals goes into a loop until the χ^2 values of each pair exceeds a specific parameter *signlevel*. This terminating condition is modified such that the merging of intervals continues until only two intervals are left. This modification allows to generate only two sets, intervals, of values for each

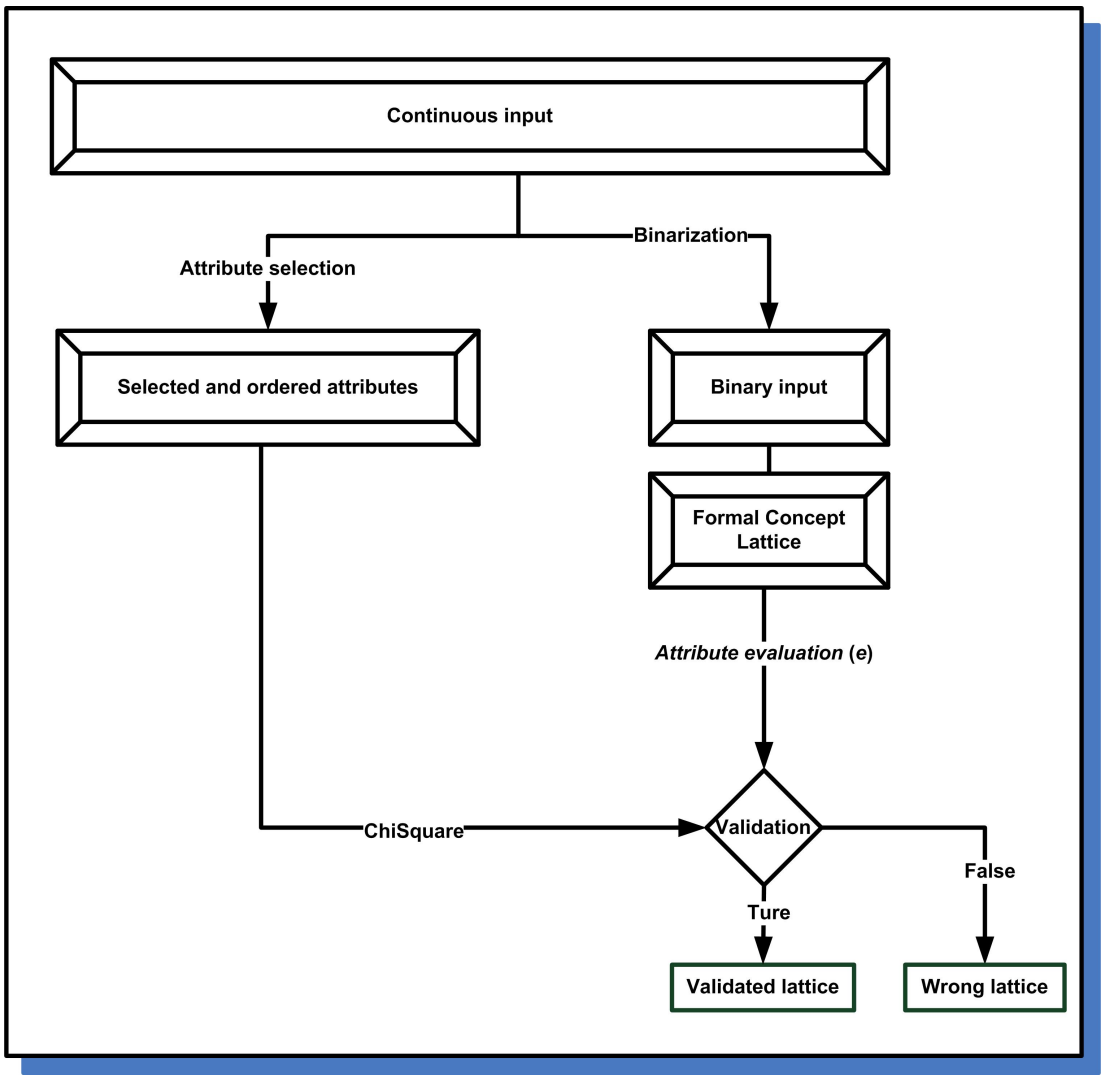


Figure 5.1: The proposed FCM model - Binarization and validation of input

attribute. The next step is to replace all the values of one of these two sets by the value of "0" and the other intervals values are replaced by the value of "1". The intervals of low values could be considered of the "0" value while that of the high values, the second interval, will have the "1" value.

5.4.2 Visualization of the binary data

The next step applies the Formal Concept Analysis on the binary data output from the previous step and generates the formal concept lattice. The generated lattice is readable and available for rule generation and analysis, rather than the previous scaling methods where the number of attributes after scaling doesn't increase.

5.4.3 Validation of the formed lattice

The validation of the generated formal concept lattice is the most important step in this model, where the need is to prove that the claim that the generated lattice reflects correctly the structure of the original input. If the most important features to the classification of the objects appears to be differentiating between objects of different classes in the generated lattice, then the claim is correct. This is clarified in the following steps where it is assumed that objects are classified in only two classes $class0$ and $class1$.

1. One of the attribute selection methods which is the chimerge technique is used to select the most important features. The resulted chisquare χ^2 values are sorted, according to the importance of attributes, in a descending order.

2. Then redesign the generated lattice in a tree form, such that the nodes with first attributes would be in the higher layer, then the nodes of two attributes in the next layer, and so on until the last layer where only one node is left that includes all the attributes.

3. In the higher layer, count the number of objects for each attribute node n, then count the number of objects in each class for this attribute, let n_0 is the number of objects in $class0$ and n_1 is the number of objects in $class1$. i.e. if for on the attribute node, attribute 1, 6 objects obj_2, obj_3, obj_{10}, obj_{11}, obj_{12}, obj_{15}, let obj_2,

obj_3 are objects of class 0 and obj_{10}, obj_{11}, obj_{12}, obj_{15} are objects of class 1, then for attribute 1, n=6, n_0=2 and n_1=4.

4. Then calculate the evaluation e of each attribute as follows:

$$e = \frac{|n_0 - n_1|}{n} \tag{5.7}$$

   i.e. for attribute 1 in the previous step, the e_1 value will be 0.66.

5. Apply the previous two steps on all attributes, Then order the e values for all attribute according to corresponding chisquare χ^2 values.

6. Finally compare the sorted order of the selected attributes resulted in step 1 with that resulted in the previous step, if the selected attributes from the chimerge technique shows the greatest e values then it is proved that the generated lattice is valid.

The validation here depends on a hypothesis that the attributes importance could be extracted from the lattice. The attribute importance can be calculated according to the value of e, where $e = \frac{|n_0 - n_1|}{n}$ as stated in step 3, so the attribute with the highest e value is the most important feature, the attribute of the lowest e value has the lowest importance to the classification problem. The reason of calculating the e value for attribute evaluation in this way, is to show how much this attribute is important in differentiating between objects of different classes. i.e. if an attribute value is 1 for objects of one class and 0 for objects of the other class, then as the number of objects of this first class increase where the attribute value is 1 increases and the number of objects from the other class of attribute value 1 decreases as the e increases, and consequently the attribute importance increases. The comparison between the evaluation of attributes generated in the lattice and the attribute selection methods is applied only on the selected attributes from the attribute selection technique, where the chisquare χ^2 of the selected attributes have values not equal to zero.

5.5 Chapter conclusion

This chapter proposed a new methodology for presenting data in a Formal Concept Analysis. Like support vector machine, the objects in both of these data sets are

categorized into only two classes. Based on the dimensionality reduction concept, only a subset of features are selected and ranked according to their importance to the classification problem. The feature ranking technique used is the chimerge technique in order to split the input into binary data, where this technique is named as the binarization of the input. After the generation of the lattice, an evaluation criterion is proposed to rank the features according to their importance to the target classes. The evaluation criteria indicate that the attributes importance isn't loss after binarization and visualization. This indication could leads to the conclusion that the constructed lattice is valid.

Chapter 6

Experimental work

The diversity in the input data is a critical requirement to prove any proposed technique. The data in the medical field could be extracted from global locations on the internet or directly from hospitals. In this chapter, different medical data sets from different resources are used in the practical and experimental work. Only few data sets from other fields are used to prove the generality of the solutions proposed. A brief description of each data set is introduced to show the importance of such studies in the medical field. The results are shown for the proposed techniques in the three stages of data mining which are the preprocessing, machine learning and visualization stages. Comparisons between the proposed techniques and other techniques including the classification accuracy of input data are discussed. Each of the proposed techniques is applied on two or more of the introduced data sets in the beginning of this chapter as a proof of concept.

6.1 Medical data

Nearly most of the input data sets are medical data sets. The proposed technique is applied on different data sets. Comparison have been made between the proposed techniques and other machine learning techniques to ensure its capability and competency among these other techniques. The data sets used are collected from different resources like UCI machine learning repository (118), Chiba University (119), specialized Ainshams hospital and others. Lets maintain the data sets and maintain the number of objects and attributes in each data set. A brief description about these data can be maintained as follows:

Buba It describes liver disorder and contains two classes: sick and healthy. The number of attributes of each object

BC The data are classifier into two parts, no-recurrence-events, recurrence-events.

Hepatitis This data set contains mostly Boolean. The numeric-valued attribute (discrete) types are BILIRUBIN, ALK-PHOSPHATE, SGOT, ALBUMIN, PRO-TIME. The class distribution is DIE of 32 patient and LIVE of 123 patient.

HCV classified according to the response of some patients to the interferon therapy whether they cured or not.

Indian-diabetes Investigated whether the patient shows signs of diabetes according to World Health Organization criteria. Class value 1 is interpreted as "tested positive for diabetes"

Yeast Its used for Predicting Protein Localization Sites in cells like Gram-Negative Bacteria and Eukaryotic Cells.

Thyroid Diagnosis of thyroid hypofunction. Patient's thyroid has overfunction, normal function, or underfunction.

Thrombosis A database collected at Chiba University hospital from the outpatient clinic of the hospital on collagen diseases (are auto-immune diseases). A Thrombosis is one of the most important and sever complications in collagen diseases. It is important to detect and predict the possibilities of its occurrence. Domain experts are very much interested in discovering regularities behind patients' observations. Thrombosis has four main levels or degrees, which are 0 (negative or no thrombosis), 1 (positive and the most severe one), 2 (positive and sever) and 3 (positive and mild). The experimental analaysis is applied on two data sets which are of 0 and 1 degrees of thrombosis. The training phase will be applied on 12 cases, of 16 attributes, of patients of thrombosis of degree 1 to extract the patterns.

WDBC Wisconsin diagnosis breast cancer to diagnose breast masses based solely on a Fine Needle Aspiration (FNA). The diagnosis could be either (M = malignant, B = benign)

Heart disease Features are extracted from heart sound signals into a data set that is used for the detection of heart valve disease. These extracted feature are 100 features that represents the four stages of a heart signal which are $S1$ signal, systolic period, $S2$ signal and diastolic period (120). These features are divided into six groups as follows:

- F1:F4 are the standard deviation of all heart sounds, $S1$, $S2$ and average heart rate.
- F5:F12 represents signal $S1$.
- F13:F36 represents the systolic period.
- F37:F44 represents signal $S2$.
- F45:F92 represents the diastolic period.
- F93:F100 the four stage of a heart signal are passed from four band-pass frequency filters. The energy of each output is calculated to form these last 8 features.

Iris The data set consists of samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

RingNorm An artificially generated data set of normally distributed attribute values.

WaveForm An artificially generated data set of 3 classes of waves. Where each class is generated from a combination of 2 of 3 "base" waves and each instance is generated f added noise (mean 0, variance 1) in each attribute.

Abalone Predicting the age of abalone from physical measurements like length, hight and weigth. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope, a boring and time-consuming task.

NSL KDD data set for the evaluation of researches in network intrusion detection system. NSL-KDD consists of selected records of the complete KDD'99 data set (121).

Table 6.1 shows a comparison between such data sets according to number of classes, features and objects.

Table 6.1: The data sets used in the proposed models to be compared to other conventional techniques

Data set name	source	Description	number of classes	number of features	number of objects
Medical					
Buba	UCI	liver disorder	Healthy-sick	6	344
BC	UCI	Breast cancer	recurrence-not	9	170
WDBC	UCI	BC diagnosis	malignant-benign	30	424
Hebatitis	UCI	liver Diseases	Healthy-Sick	19	66
HCV	Ain-Sham	liver Diseases	Healthy-sick	13	66
Indian-diabetes	UCI	kidney	Healthy-Sick	8	536
Thromposis	Chiba	collagen disease	Healthy-sick	16	34
Thyroid	UCI	patients	over-normal-under	5	90
Heard sound	Karlovassi	valve disease	murmur-systolic	100	148
Biological					
Iris	UCI	species of plant	3 species	4	147
Yeast	UCI	Cellular Biology	4 classes	8	600
Abalone	UCI	Physics	2 classes	8	1900
Artificial					
Ring Norm	Delve	Generated	2 classes	20	7400
WaveForm	UCI	Generated	3 classes of waves	21	5000
NSL KDD	KDD	Network	attack-normal	41	14870

6.2 Performance measure

The classification effectiveness of the improved model is calculated in terms of *precision* and *recall* (122) which are defined as follows:

$$precision = \frac{a}{a+b}, a+b>0 \tag{6.1}$$

$$recall = \frac{a}{a+c}, a+c>0 \tag{6.2}$$

Where a counts the assigned and correct cases, b counts the assigned and incorrect cases, c counts the not assigned but incorrect cases and d counts the not assigned and correct cases. To evaluate the performance across categories, $F-measure$ is averaged. There are two kinds of averaged values, namely, micro average and macro average. The micro value is the average of the two values, *precision* and *recall*, of all classes, while the macro average is the average of all micro values, this average represents the $F-measure$. The results appear in this chapter are the F-measure multiplied by 100 and the values will be represented as the percentage of classification accuracy.

6.3 Experimental work on proposed preprocessing techniques

6.3.1 Bypass Discretization using Interval-based feature ranking

The selection of the data sets used is based on the need of a data set of continuous attributes and discrete class label. Medical data sets have used which are considered as real-life data sets that have no specific distribution of values and may contain misleading values due to an error in calibrations or collection of data. The first data set used is Indians-diabetes data set. The percentage of error in this data set will be considered zero. 90% of the input data used for training while the rest of 10% is used in testing. The second data set used is a data about HCV where there is a percentage of error that may occur in this data set, where experts indicate that it falls between 2 to 3%. Due to the low number of objects, 70% of the input data only are used for training while the rest of 30% is used in testing. Both data sets are adjusted such that the number objects in every class is equal in both stages of training and testing, the classification accuracy will be measured by dividing the number of correctly classified objects divided by the total number of objects in the testing data set. Another type of data set has been used which NSL-KDD data set, 70% of the input data only are used for training while the rest of 30% is used in testing.

6.3.1.1 Indians-diabetes data set

When different attribute selection is used like information Gain (IG), Chi-Merge (CM) and Gain Ratio (GR), these techniques show the same order of ranked attributes. This is because these techniques are all entropy-based attribute selection techniques. The comparison between the order of features ranks of entropy-based attribute selection techniques and the interval-based (IB) feature evaluation and selection technique is shown in table 6.2. In this table another SVM-based feature selection technique (SVMB) (123) is used, where it shows nearly the same results as Information gain technique.

Table 6.2: The order of attributes according to Information gain IG and Interval-based IB feature selection algorithms

IG	2	8	6	5	1	7	3	4
SVMB	2	6	1	7	8	3	4	5
IB	2	5	7	1	4	6	8	3

Table 6.3, shows the results when perform classification using support vector machine (SVM) and multilayer perceptron (MLP). In the case using MLP classifier, the peak of accuracy has reached with 81.4% when the input data set contain only the first three selected attributes by the IB technique which are 2, 5, 7. While the peak when using the other feature selection techniques is 75.9% and the number of selected attributes is five attributes which are 2, 8, 6, 5, 1. In the case of using SVM both feature selection techniques, the proposed IB and the IG attribute selection techniques, have the same accuracy percentage peak when the first attribute only is selected. It is noticed that both techniques have selected the same attribute 2 as the most relevant attribute.

On the other hand, All the possible combination of attributes are tested in the same way as above using MLP classifier, where 90% of the input data is for training while the rest is for testing. the combination that shows the best results is the set of attributes 2, 5, 7 which is the same set and of the same order generated by the proposed IB .

Table 6.3: Percentage of classification accuracy in the Indians-diabetes case

Classifier technique	SVM	SVM	SVM	MLP	MLP	MLP
Feature selection	IG	SVMB	IB	IG	SVMB	IB
Number of selected features						
1	**64.81**	64.81	**64.81**	74.07	74.07	74.07
2	64.81	61.11	61.11	74.07	74.07	75.92
3	64.81	**66.66**	62.96	74.07407	70.37	**81.48148**
4	53.70	66.66	51.85	72.22222	72.22	79.62963
5	57.40	68.51	57.40	**75.92593**	75.92	75.92593
6	57.47	62.96	57.40	74.07407	**77.77**	74.07407
7	55.55	55.55	53.70	72.22222	77.77	68.51852
8	51.85	51.85	51.85	72.22222	70.37	72.22222

6.3.1.2 Hebatitis C Virus data set

Again in the case Information Gain (IG), Chi-Merge (CM) and Gain Ratio (GR) attribute selection techniques, The attributes are ranked as follows in table 6.4. The SVM-Based attribute selection shows the same results as the previous techniques so it is not useful to maintain them in table 6.4 Table 6.5, shows the results when perform

Table 6.4: The order of attributes according to Information gain IG and Interval-based IB feature selection algorithms

IG	12	13	4	5	1	3	2	9	11	10
IB	3	4	6	13	9	12	1	5	8	7
IG	6	8	7							
IB	11	10	2							

classification using Support vector machine (SVM) and Multi-layer perceptron (MLP) after using both entropy-based attribute selection technique like the IG and the proposed IB techniques. The first row in table 6.5 shows the classification results when using the input data set contains only attribute 12 in the case of using IG and attribute

3 in the case of using IB. The second row the input data set contains attributes 12, 13 in the case of using IG and attributes 3, 4 in the case of IB, and so on until all attributes are used in the input data set.

Table 6.5: Percentage of classification accuracy in the HCV case

Classifier technique	SVM	SVM	MLP	MLP
Feature selection	IG	IB	IG	IB
Number of selected features				
1	25.0	50.0	37.5	50.0
2	37.5	37.5	37.5	37.5
3	37.5	37.5	37.5	37.5
4	37.5	**75.0**	37.5	**62.5**
5	**62.5**	62.5	37.5	50.0
6	37.5	37.5	37.5	37.5
7	50.0	37.5	37.5	37.5
8	50.0	25.0	37.5	37.5
9	50.0	62.5	37.5	37.5
10	50.0	62.5	37.5	50.0
11	50.0	62.5	37.5	37.5
12	50.0	50.0	37.5	37.5
13	50.0	50.0	37.5	37.5

Table 6.5 shows that in the case of the selected attributes using the proposed IB technique, the classification accuracy has the maximum value when using the first four attributes only. Also the peak was 75 % in the case of using MLP, and 65 % in the case of using SVM. So in both classifiers, the selected attributes by IB has a higher peak than those selected by other attribute selection techniques.

6.3.1.3 NSL-KDD data set

In the case of using the NSL-KDD data set, the chimerge and IG technique shows difference in the ranking of the attributes. When applying chimerge technique before MLP,

the peak is 92.96 % when selecting the 41 attribute and when applying IG technique, the peak is 92.96 % when selecting the 33 attributes. On the other hand when applying the proposed IB technique before MLP, The maximum classification accuracy (peak) is 93.2% that appears when selecting 32 attribute. This shows that the proposed IB technique reaches the maximum classification accuracy that is not reached by chimerge and IG , and also with a lower number of features. This is clear in figure 6.1 where it shows that the last variation of the classification accuracy against the last 17 ranked attributes.

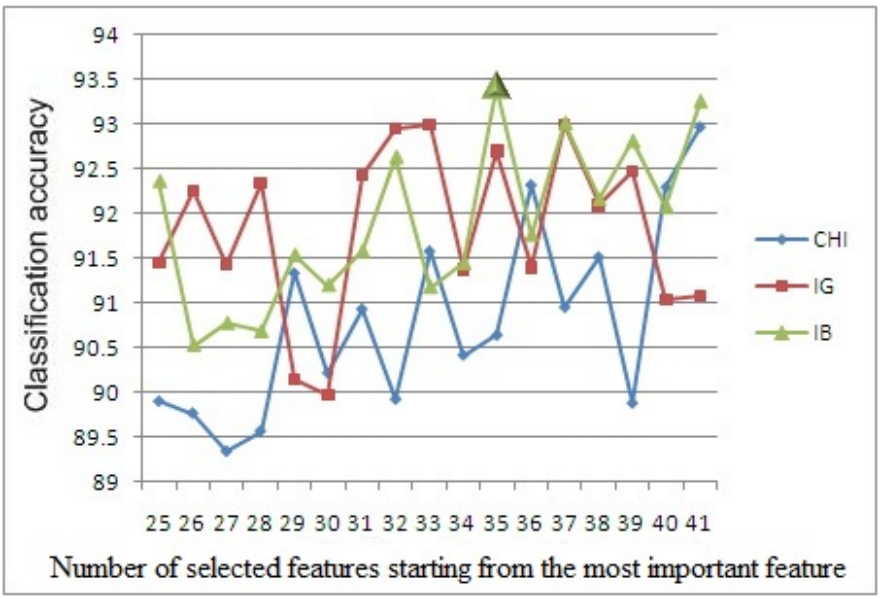


Figure 6.1: NSL KDD - Classification accuracy according to selected attributes

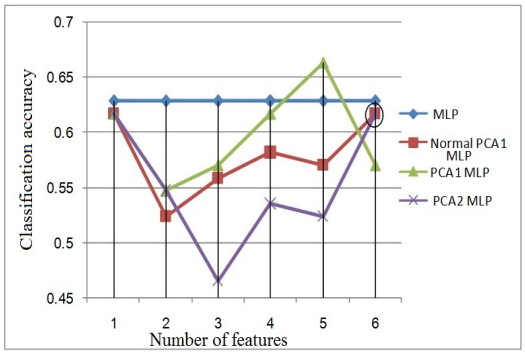
6.3.2 Normalization effect on PCA feature reduction

Different data sets have been used to compare between both techniques (conventional and correlation PCA) in different multivariate distributions of data, either normal or not. The test will be applied to Bupa (liver disorder) and Thrombosis (complications in collagen diseases) data sets which are two medical real-life data sets. Both data sets show a normality percentage that varies from 0.66 to 1 range, the test will be applied also to Abalone of normality 1.

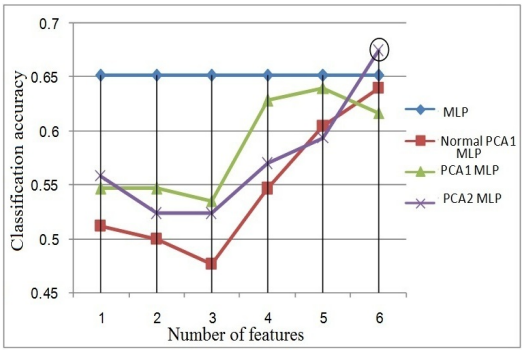
These data sets will be divided into four groups to provide a better analysis variety while Thrombosis (34 objects only) will be tested only once for all the objects due to the small number of objects. Other data sets show nearly the same performance for the three PCA models which will be unhelpful in comparing the models.

Figure 6.3 and 6.3 shows the classification accuracy of the classifier in the four cases discussed in section 4. The figures show 4 different subsets in the sub-figures a,b,c and d for the Buba and the Abalone data sets respectively.

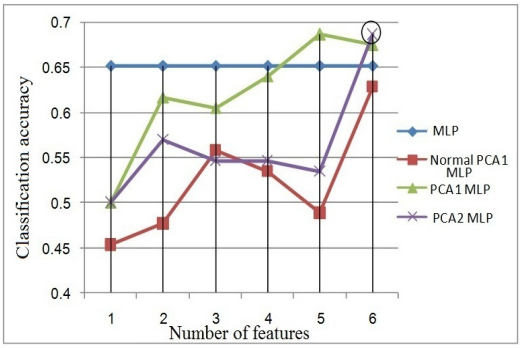
Figure 6.2: Bupa Data: The percentage of classification accuracy versus the number of features selected for classification



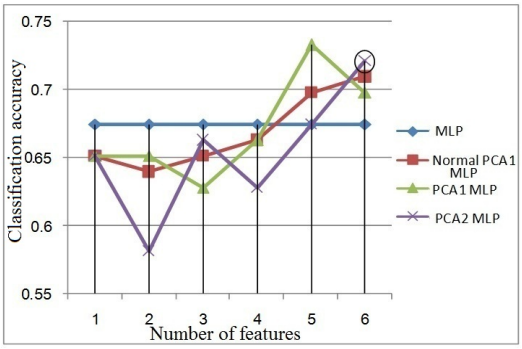
(a) First group



(b) Second group



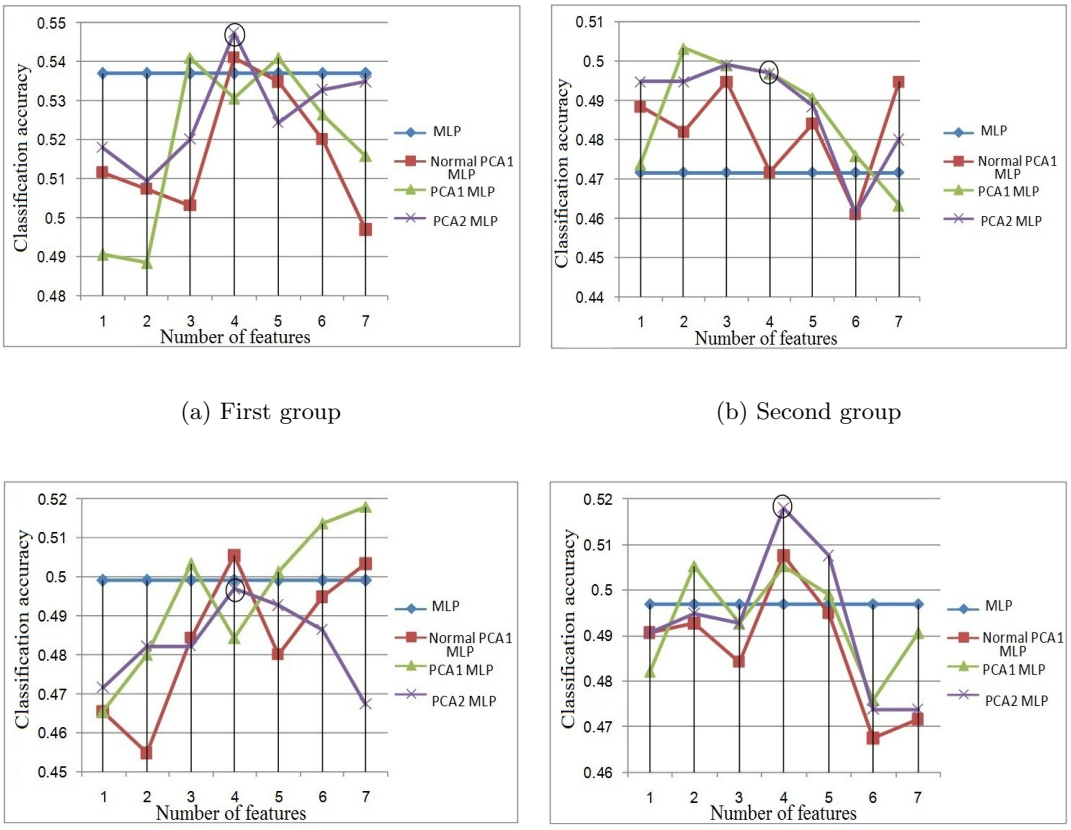
(c) Third group



(d) Fourth group

The number of Principle Components will be defined according to the accuracy of MLP classifier. In other words the number of the required Principle Components is data dependent. This is clarified in figures 6.2 and 6.3 as the first 6 Principle Com-

Figure 6.3: Abalone Data: The percentage of classification accuracy versus the number of features selected for classification



(a) First group

(b) Second group

(c) Third group

(d) Fourth group

ponents is processed (passed to the MLP classifier) out of 6 Principle Components in Buba data set, while the first 4 Principle Components is processed out of 7 Principle Components in Abalone data set. When processing these selected Principle Components, the classifier gets the best performance rather than selecting any other number of Principle Components. Figure 6.2 and 6.3 shows a better performance when applying Preprocessing before the running the MLP classifier rather than running the classifier directly. Also these figures show that the absence normalization before applying PCA has a good effect on classification accuracy. And this confirms the hypothesis that normalization deteriorates the data structure and leads to miss-classification. PCA2 technique as a feature extraction technique shows the best performance only when the

Table 6.6: MLP classifier performance before applying PCA

Data sets	Normality %	number of records	No. of preprocessing
Bupa	0.666667	86	0.651162791
	1	86	**0.627906977**
	0.666667	86	0.651162791
	0.666667	86	0.651162791
Abalone	1	425	0.536842105
	1	425	0.471578947
	1	425	0.498947368
	1	425	0.496842105
Thrombosis	1	34	0.558824

appropriate number of Principle Components is selected. This appears in the figure 6.2, for all the four subsets, as in the four sub-figures of figure 6.2, when 6 features are extracted from PCA2 followed by MLP shows a better performance in all the four cases (subsets). Also in figure 6.3 for the four subset in figures 6.4(a),6.4(b), 6.4(c) and 6.4(d)., when 4 features are extracted from PCA2 followed by MLP shows the best performance. This leads to a very important result which is selecting the correct number of feature extracted guarantee that the PCA2 has better performance without varying according to the used data. Otherwise the power of performance could be change from PCA1 and PCA2 according to the data. The final conclusion from Figures 6.2 and 6.3 is that the normalization process could deteriorate the structure of the data and lead to a low classification accuracy in some cases. This appears in the sub-figure 1-c when the number of Principle Components is 6, and in sub-figures 6.4(a) and 6.4(b) when the number of Principle Components is 4. In these cases both PCA1 and PCA2 without the normalization algorithm shows the best performance. But it also could be important in some other cases like in sub-figure 6.3(a), 6.3(b), 6.3(d), 6.4(d) and 6.4(c) where the PCA1 with normalization shows a better performance than PCA1 without normalization.

Table 6.6 shows the normality percentage and classification accuracy for the three used data set.

Table 6.7 shows the results of using 6 Principle Components (principle components) in MLP classification of Buba data set and 4 Principle Components in Abalone data set and finally 12 Principle Components in Thrombosis data set. And the maximum values appears usually in the PCA2 column.

Table 6.7: MLP classifier performance after applying PCA

Data sets	Normalization + PCA	PCA	Modified PCA
Bupa	0.639534884	0.61627907	0.674418605
	0.61627907	0.569767442	0.61627907
	0.627906977	0.674418605	0.686046512
	0.627906977	0.674418605	0.686046512
Abalone	0.541052632	0.530526316	0.547368421
	0.471578947	0.496842105	0.496842105
	0.505263158	0.484210526	0.496842105
	0.507368421	0.505263158	0.517894737
Thrombosis	0.764705882	0.647058824	0.735294118

The results in table 6.6 and 6.7 shows that the PCA2 preprocessing leads to better or equal classification accuracy with respect to other PCA1 preprocessing methods or applying normalization before PCA1. Also it shows that PCA2 does not depend whether the data set is normally distributed or not. As it shown in table 6.6, Buba data set has only 0.66 normality percentage, which means that 66% of the features are normally distributed, and PCA2 still has a better performance.

6.3.3 The independence of rough set classifier on feature reduction

This part introduces the ability of rough set methodology to successfully classify heart sound diseases without the need applying feature selection. The capabilities of rough set in discrimination, feature reduction classification have proved their superior in classification of objects with very excellent accuracy results. On applying rough set discretization and disease prediction on the heart disease data sets, the need of applying feature selection is not required due to the reducts concept, the rough set produces the highest classification accuracy. The reducts set represents the attributes in the decision rules generated from the input data set in the rough set classifier. It is shown from these resulted reducts are dependents on the attributes selected by features like chimerge and information gain methods. The tool that has been used in this chapter is the rough set exploration system tool (124)

6.3.3.1 The set of reducts in comparison to the chimerge feature selection technique

For each data set, we reach the minimal number of reducts that contains a combination of attributes which have the same discrimination factor for each data set. Table 6.8 shows the final generated reducts for each data set, which are used to generate the list of the rules for the classification.

Table 6.8: Rough reducts sets of the three data sets

Data set	Reduct sets
HS_AR_MS	3, 8, 31, 38,82
HS_AS_MR	3, 6, 36,39
HS_N_S	1, 9, 87, 94, 97

In order to evaluate the proposed rough sets classifier, we will study the lets discuss the reducts in comparison to the feature selection using chimerge approach and find the results of the classifiers after feature selection for each data set.
For the first data set HS_AR_MS, the selected number of features by the chimerge technique are:

{F32, {**F31**}, F30, F29, F100, F28, F33, F27, {**F3**}, F5, F4, F49, F48, F45, F46, F50, F25, F26}.

It noticed that the selected features by the chimerge technique includes feature $F31$ and feature $F3$, where feature $F31$ is the *second* most important feature.
In regard to the second data set HS_AS_MR, the selected features using the chimerge technique are:

{{**F36**}, F13, F12, F14, F15, F35, F4, F2, F18, F17, F11, {**F3**}, F92, F5, F16, F20, F19, F23, F21, F93, F54, F22, F95, F10, F1, F94, F24, F28, F37, F25, F41, F29, F100, F40, F26, F88, F55, F96}.

It noticed that the selected features by the chimerge technique includes feature $F36$ and feature $F3$, where feature $F36$ is the *second* most important feature. Also it is noticed that in both data sets, HS_AR_MS and HS_AS_MR data sets, feature $F3$ is

selected and it is common in chimerge selected features and reducts.
For the third data set "HS_N_S", the selected features using the chimerge technique are as follows:

{F54, F58, F53, F65, F64, F67, F59, {**F97**}, F61, F70, F96, F98, F55, F60, F66, F51, F52, F49, F62, F63, F23, F22, F45, F50, F2, F56, F57, F71, F28, F27, F24, F26, F25, F47, F46, F99, F21, F72, F68, F69, F3, F75, F73, F92, F48, {**F87**}, F7, F76}.

It noticed that the selected features by the chimerge technique includes feature $F97$ and feature $F87$, where feature $F97$ is the $eight^{th}$ most important feature. Also it is noticed that feature $F3$ is appears in all the reducts in the three data sets.

6.3.3.2 The set of extracted rules

The extracted rules from HS_AR_MS, HS_AS_MR and HS_N_S data sets are shown in Tables 6.9, 6.10 and 6.11 respectively.

Table 6.9: Generated rules for the HS_AR_MS data set

Matches	Decision rules
22	(3="(0.590,Inf)")&(31="(0.004,Inf)")&(38="(0.01,Inf)")=>(D=1[22])
22	(31="(0.00,Inf)")&(38="(0.01,Inf)")&(82="(-Inf,0.00)")=>(D=1[22])
21	(3="(-Inf,0.59)")&(31="(-Inf,0.004)")=>(D=0[21])
20	(8="(0.04,Inf)")&(31="(0.004,Inf)")&(38="(0.012,Inf)")=>(D=1[20])
12	(8="(0.04,Inf)")&(31="(-Inf,0.004)")=>(D=0[12])
11	(3="(0.59,Inf)")&(8="(-Inf,0.04)")&(82="(-Inf,0.00)")=>(D=1[11])
11	(8="(-Inf,0.04)")&(31="(0.004,Inf)")&(82="(-Inf,0.00)")=>(D=1[11])
9	(3="(-Inf,0.59)")&(8="(-Inf,0.04)")&(82="(0.00,Inf)")=>(D=0[9])
9	(3="(0.59,Inf)")&(8="(-Inf,0.04)")&(31="(0.004,Inf)")=>(D=1[9])
8	(3="(-Inf,0.59)")&(38="(-Inf,0.01)")&(82="(0.00,Inf)")=>(D=0[8])
7	(3="(-Inf,0.59)")&(8="(0.04,Inf)")&(38="(-Inf,0.01)")=>(D=0[7])
6	(3="(0.59,Inf)")&(31="(0.004,Inf)")&(82="(0.00,Inf)")=>(D=1[6])
5	(31="(-Inf,0.004)")&(38="(0.01,Inf)")&(82="(0.00,Inf)")=>(D=0[5])
5	(3="(0.59,Inf)")&(8="(0.04,Inf)")&(82="(0.00,Inf)")=>(D=1[5])
4	(8="(0.04,Inf)")&(38="(-Inf,0.01)")&(82="(-Inf,0.00)")=>(D=0[4])
3	(3="(0.59,Inf)")&(8="(-Inf,0.04)")&(38="(-Inf,0.01)")=>(D=1[3])
3	(3="(0.59,Inf)")&(31="(-Inf,0.004)")&(38="(-Inf,0.01)")=>(D=1[3])
2	(3="(0.59,Inf)")&(38="(-Inf,0.01)")&(82="(0.00,Inf)")=>(D=1[2])

Table 6.10: Generated rules for the HS_AS_MR data set

Matches	Decision rules
26	(3="(-Inf,0.31)")&(36="(0.047,Inf)")=>(D=0[26])
25	(6="(0.05,Inf)")&(36="(-Inf,0.04)")=>(D=1[25])
18	(3="(-Inf,0.31)")&(6="(-Inf,0.05)")=>(D=0[18])
14	(3="(0.31)")&(39="(0.001,Inf)")=>(D=1[14])
13	(3="(0.31,0.76)")&(36="(-Inf,0.04)")=>(D=1[13])
11	(3="(0.76,Inf)")&(36="(-Inf,0.04)")=>(D=1[11])
6	(3="(0.31)")&(6="(-Inf,0.05)")=>(D=1[6])
5	(3="(0.76,Inf)")&(6="(0.05,Inf)")&(36="(0.04,Inf)")=>(D=0[5])
3	(3="(0.76,Inf)")&(6="(-Inf,0.05)")&(39="(0.001,Inf)")=>(D=1[3])
2	(6="(-Inf,0.05)")&(36="(-Inf,0.04)")&(39="(-Inf,0.001)")=>(D=0[2])
2	(3="(0.76,Inf)")&(39="(-Inf,0.001)")=>(D=0[2])
2	(6="(0.05,Inf)")&(36="(0.04,Inf)")&(39="(-Inf,0.001)")=>(D=0[2])

Table 6.11: Generated rules for the HS_N_S data set

Matches	Decision rules
20	(94="(0.16,Inf)")&(97="(0.04,Inf)")=>(D=1[20])
12	(94="(-Inf,0.16)")&(97="(-Inf,0.04)")=>(D=0[12])
9	(1="(0.12,Inf)")&(9="(0.00,Inf)")&(97="(-Inf,0.04)")=>(D=0[9])
8	(9="(0.00,Inf)")&(87="(-Inf,0.00)")&(97="(-Inf,0.04)")=>(D=0[8])
8	(1="(0.12,Inf)")&(87="(0.00,Inf)")&(94="(-Inf,0.16)")=>(D=0[8])
8	(9="(-Inf,0.00)")&(87="(-Inf,0.00)")&(97="(0.04,Inf)")=>(D=1[8])
7	(1="(0.12,Inf)")&(87="(0.00,Inf)")&(97="(-Inf,0.04)")=>(D=0[7])
6	(9="(-Inf,0.00)")&(87="(0.00,Inf)")&(94="(-Inf,0.16)")=>(D=0[6])
4	(1="(0.12,Inf)")&(87="(-Inf,0.00)")&(97="(0.04,Inf)")=>(D=1[4])
3	(9="(-Inf,0.00)")&(87="(0.00,Inf)")&(97="(-Inf,0.04)")=>(D=0[3])
3	(9="(-Inf,0.00)")&(87="(-Inf,0.00)")&(94="(0.16,Inf)")=>(D=1[3])

6.3.3.3 Classification accuracy of the rough set model in comparison to the other classification techniques

The comparison shown in table 6.12 contains a comparative study between the proposed model of rough set classifier and other conventional and known classifiers. Feature selection will be applied before the conventional classifier in order to enhance the corresponding percentage of classification accuracy. On the other hand, rough set model will not preceded by feature selection, where it will depend only the set of generated reducts. For example, in the case of HS_AS_MR data set, the classification accuracy of DT and SVM after feature selection are 89.0 and 94.5 respectively, where these results are still less than that of rough set accuracy results. Although, Sequential minimal optimization, (SMO) has the percentage of classification accuracy only slightly greater than the Rough set system in the case of HS_AS_MR data set, rough set model has a greater percentage of classification accuracy in HS_AR_MS and HS_N_S data sets.

Table 6.12: Accuracy results: Comparative analysis among Hidden Naive Bayesian network (HNB), Bayesian network (BN), Naive Bayesian tree (NBT), Decision tree (DT), Sequential minimal optimization (SMO)

Classifier	HS_AS_MR	HS_AR_MS	HS_N_S
SVM	90.54	90.54	85.93
HNB	90.54	91.89	**90.625**
BN	86.48	83.78	84.37
DT	89.18	83.78	82.81
SMO	**93.24**	94.59	87.50
NBT	87.83	89.18	89.06
RS	**92.90**	**97.1**	**90.0**

Figure 6.4 illustrates the overall rough sets classification accuracy in terms of sensitivity and specificity compared with Hidden Naive Bayesian network (HNB), Bayesian network (BN), Naive Bayesian tree (NBT), Decision tree (DT), Sequential minimal optimization (SMO). Empirical results reveal that the proposed rough set approach performs better than the other classifiers.

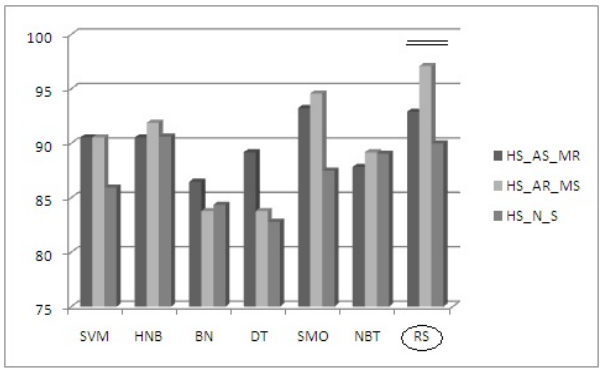


Figure 6.4: Classification accuracy: Comparative analysis among Support Vector Machine (SVM), Hidden Naive Bayesian network (HNB), Bayesian network (BN), Naive Bayesian tree (NBT), Decision tree (DT), Sequential minimal optimization (SMO)

6.4 Experimental work on proposed machine learning techniques

6.4.1 Frequent pattern subspace classification

The model is tested against five different classification techniques which are bayesian network, Support Vector Machine, multilayer neural network, decision table, The instance-based learner(IB1) and classification via clustering. These classification techniques are performed using the Weka software (125). 90 percent of the input is used for the training and the rest is for testing, this partitioning is performed on all the classifiers under comparison. The results appears of Iris, Buba and Thrombosis data sets appears in Tables 6.13, 6.14 and 6.15 respectively.

6.4.1.1 Iris data set

In the case of Iris data set, The proposed pattern-based classifier model shows the best classification accuracy of 98.2% over the other classifier. Also the training time required for classification is better than that of the MLP that shows the second best classification accuracy of 96%. The number of patterns using in the testing data set is 2 patterns out of 6 patterns, where only two patterns shows the best classification accuracy as shown in figure 6.5.

Table 6.13: Comparison of the proposed model with other classification techniques according to the classification accuracy of Iris data set

Classifier	Accuracy	Training Time in milliseconds	Testing Time in milliseconds
PBC	98.24561404	10	10
IB1	94.73684211	0	0
MLP	96.49122807	110	0
NB	94.73684211	0	10
CC	66.66666667	10	0
DT	94.73684211	10	0
BN	94.73684211	10	0

6.4.1.2 Buba data set

Although the proposed model shows the second best classification accuracy of 60.5% where MLP got 75% in case of Buba data set classification, The training time required in MLP was 0.46 while it is only 0.016 in the proposed mode. The number of patterns using in the testing data set is 6 patterns out of 8 patterns, as six patterns shows the best classification accuracy as shown in figure 6.6.

Table 6.14: Comparison of the proposed model with other classification techniques according to the classification accuracy of Buba data set

Classifier	Accuracy	Training milliseconds	Testing milliseconds
PBC	60.52631579	60	17
IB1	52.63157895	4	28
MLP	78.94736842	1004.5	0
NB	42.10526316	20	0
CC	18.42105263	130	0
DT	47.36842105	152	0
BN	50	20	10

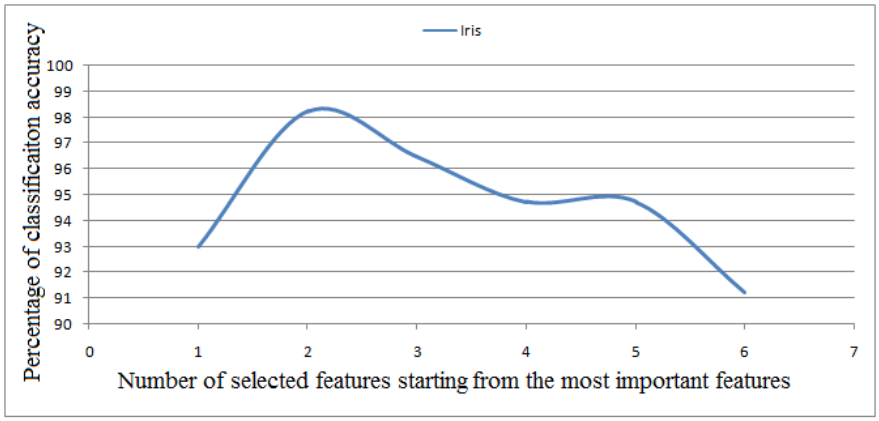


Figure 6.5: Iris Accuracy - classification accuracy according to selected attributes

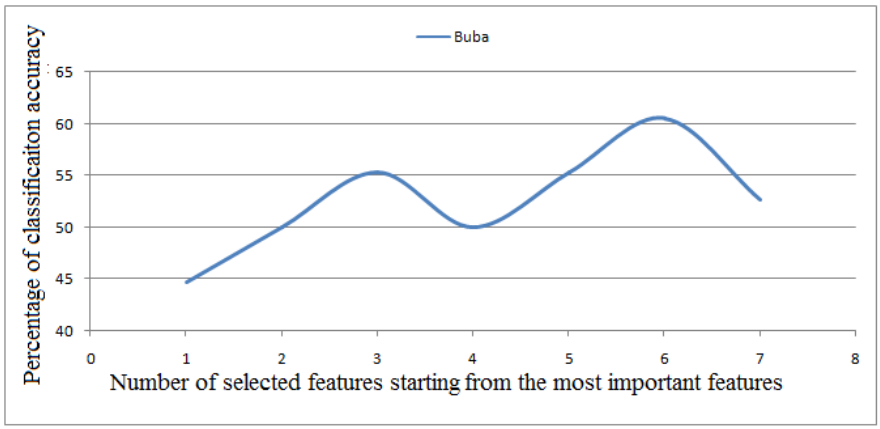


Figure 6.6: Buba Accuracy - classification accuracy according to selected attributes

6.4.1.3 Thrombosis data set

The results of the classification Thrombosis data set appears in table 6.15. The number of patterns using in the testing data set is 60 patterns out of 100 patterns, as sixty patterns shows the best classification accuracy as shown in figure 6.7.

Table 6.15: Comparison of the proposed model with other classification techniques according to the classification accuracy of Thromposis data set

Classifier	Accuracy	Training milliseconds	Testing milliseconds
PBC	75	50	50
IB1	75	1	0
MLP	62.5	100	0
NB	75	0	0
CC	50	0	0
DT	50	20	0
BN	50	0	0

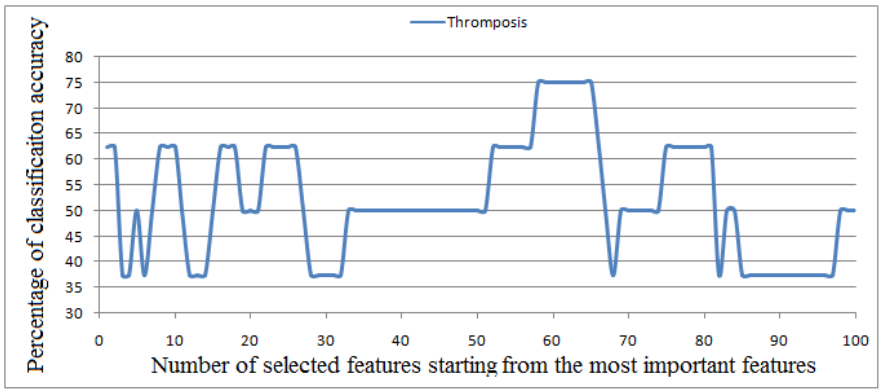


Figure 6.7: Thrombosis Accuracy - classification accuracy according to selected attributes

6.4.2 Fuzzification of Euclidean calculations

The proposed improvement is applied on two data sets and compared to the case when all features are used. Also it is compared to the case when only the important features are used in classification, in order to perform this, the test is applied on the first important feature only, then the first two important features, and so on until all feature are used in classification test. On the other hand the FCM with the proposed modification is compared to other classifiers. This proposed enhancement is applied on two different data sets to ensure its capability and competency among other classifiers. These data sets are the indian diabetes (ID) data set of 536 objects and 8 features, the yeast data set of 600 objects and 8 features, wine and WDLC data set of 569 objects and 30 features, hepatitis data set of 200 objects and 19 features, waveform data set of 300 objects and 21 attributes, HS_AS_MR (Heart sound for valve disease) data set of 60 objects and 100 features, and finally letters data set of 300 objects and 16 attributes.

6.4.2.1 Classification results for FCM

Table 6.16 shows classification accuracy percentage calculated by the Macro averaged $Fmeasure$ multiplied by 100. It shows a comparison between FCM classifier before and after modification, also it compares these results to the FCM classifier after applying feature selection. The modification applied in equation 4.3.2.3 shows in the cases of indian diabetes, yeast, hepatitis and HS_AS_MR an increase in the classification accuracy in comparison to the ordinary FCM. This modification also shows in the cases of indian diabetes and yeast data sets an increase in the classification accuracy in comparison to the ordinary FCM after applying feature selection using chimerge technique, while in the other cases maintained in this comparison, the resulted accuracy is the same. These results proves the importance of applying fuzzification in the Euclidean distance calculation in Fuzzy c-mean clustering as it removes the features of no relevance the classification problem. And of the other hand, it takes into consideration the degree of importance of relevant features which gives it a privilege over the using of feature selection techniques.

The enhancement in the classification performance from the Fuzzy C-Mean to the modified Fuzzy C-Mean model appears to be raised from 57% to 64.2% when being applied

on the indian diabetes data set. When applying the modification on the Fuzzy C-Mean model, the classification accuracy of the yeast data set is raised more to 83.3%. Chimerge ranks leads to a better classification performance than that of Mutual information technique due to the better discretization algorithm embedded in the chimerge technique. This proves that not all of the selected features have the same importance as each other. In the case of indian diabetes and yeast data sets, the chi-square technique produces different values for each attributes and only few attributes have zero values, equation 4.47 will be applied without changes. In cases where all features are ordered with different importance but all have the zero rank values, r_k values could be replaced by d_k values.

Table 6.16: The classification accuracy percentage of six data sets

Data Set Used	ID	Yeast	WDBC	Hepatitis	HS_AS_MR	Waveform
Accuracy using FCM only	57%	77.8%	93.18%	50%	87.5%	90%
Accuracy using modified FCM	64.2%	83.3%	93.18%	75%	100%	90%
Accuracy using feature selection before FCM	53%	79.6%	93.18%	75%	100%	90%

Applying feature selection before classification: A sequential forward feature selection algorithms will be applied on the six data sets used to validate the importance of the proposed modification. The features are ordered according to their χ^2 values in an descending order. The first feature is the most important feature then the importance decreases until the last feature as shown in tables. Table 6.17 shows the attributes followed by ":" followed the chimerge value.

The classification accuracy against the selection of the features according to their importance are shown in figures 6.8, 6.9 and 6.10. The first subset of features used contains the first, most important or the highest χ^2 value, the next subset of features contains the most important two features. The subset of features increases until all

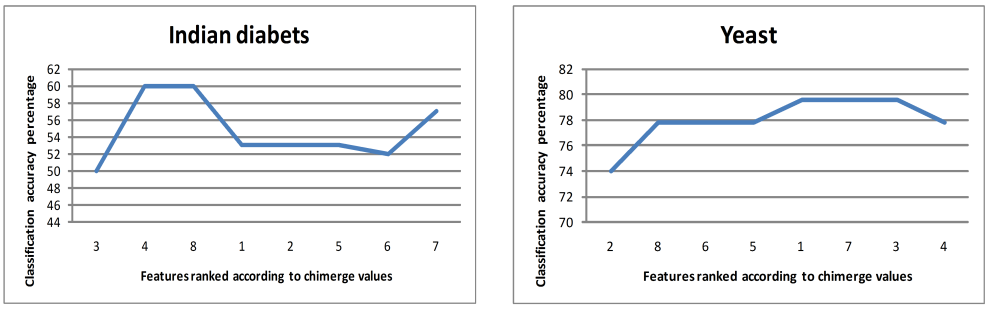
Table 6.17: Attributes are sorted in descending order according to their corresponding chimerge resulted values.

Indian diabetes	3:474.8, 4:173.7, 8:45.5, 1:43.2, 2:36.5, 5:0, 6:0, 7:0
Yeast	2:150.5, 8:69.6, 6:66.2, 5:46, 1:44.1, 7:0, 3:0, 4:0
wdbc	23:327.7599, 21:322.2237, 24:315.0164, 28:307.7802, 8:302.815, 3:274.1966, 7:271.2789, 4:269.8813, 1:265.5033, 27:263.61, 14:247.1274, 13:188.1631, 6:180.7739, 26:179.8408, 11:179.562, 22:147.3341, 2:138.5774, 17:123.2408, 18:116.1362, 16:84.5679, 25:82.1957, 29:72.5917, 5:54.7833, 30:50.2809, 9:48.088, 20:19.4187, 15:0, 19:0, 10:0, 12:0
Hebatites	1:11.591, 13:10.185, 12:9.429, 14:7.543, 6:0, 8:0, 7:0, 4:0, 5:0, 2:0, 3:0, 17:0, 16:0, 19:0, 18:0, 10:0, 9:0, 15:0, 11:0
HS_AS_MR	13:38.83851, 36:37.22024, 12:36.44602, 14:36.44602, 15:33.093, 35:32.29091, 18:31.95592, 17:31.95592, 3:31.30769, 16:29.8466, 4:29.36684, 2:28.69872, 11:28.69872, 19:28.69872, 92:28.69872, 22:26.64, 20:26.64, 95:26.64, 94:25.71631, 93:25.71631, 5:24.48, 21:24.0018, 10:23.71795, 23:22.42424, 1:22.2, 24:21.80735, 54:20.41379, 25:19.98, 37:19.21473, 28:18.61491, 41:18.61491, 88:18.16364, 29:17.83929, 40:17.45796, 26:15.78667, 27:14.194, 96:10.24615, 55:10.24615, 46:0, 39:0, 43:0, 45:0, 42:0, 44:0, 6:0, 30:0, 9:0, 7:0, 8:0, 34:0, 38:0, 31:0, 33:0, 32:0, 79:0, 78:0, 82:0, 80:0, 81:0, 74:0, 72:0, 73:0, 77:0, 75:0, 76:0, 97:0, 90:0, 91:0, 100:0, 98:0, 99:0, 85:0, 83:0, 84:0, 89:0, 86:0, 87:0, 56:0, 53:0, 59:0, 57:0, 58:0, 49:0, 47:0, 48:0, 52:0, 50:0, 51:0, 68:0, 66:0, 67:0, 71:0, 69:0, 70:0, 62:0, 60:0, 61:0, 65:0, 63:0, 64:0
WaveForm	10:103.4748, 9:92.7483, 11:90.5961, 17:76.0242, 16:63.8204, 18:59.8624, 15:58.4475, 8:43.5202, 19:27.1048, 12:22.9167, 20:22.8311, 7:13.9037, 14:11.6402, 3:0, 4:0, 2:0, 1:0, 21:0, 13:0, 5:0, 6:0

features are used in the classification test. For example, in figures 6.8(a) and 6.8(b), the maximum accuracy reached when the features selected are $\{2, 8, 6, 5, 1\}$ and $\{3, 4\}$ respectively.

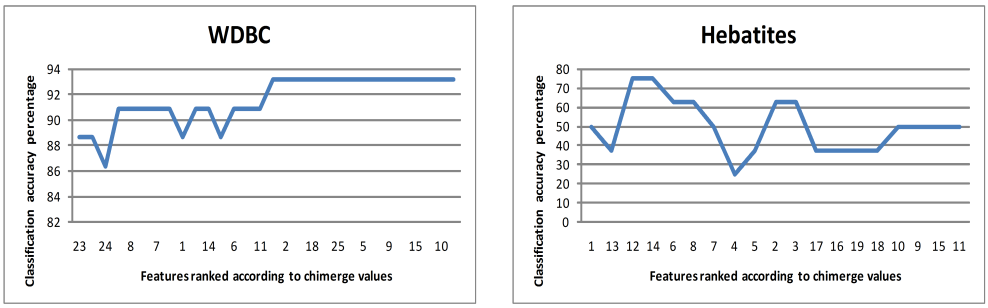
According to the results shown in figure 6.8, 6.9 and 6.10, the classification accuracy percentage reached is either greater than the global maxima or equal. This leads to the

conclusion that the proposed technique in fuzzification of the distance using chimerge ranked values have the ability to avoid completely the cross-validation repetitions in sequential forward feature selection algorithms (39).



(a) Indian diabetes (b) Yeast

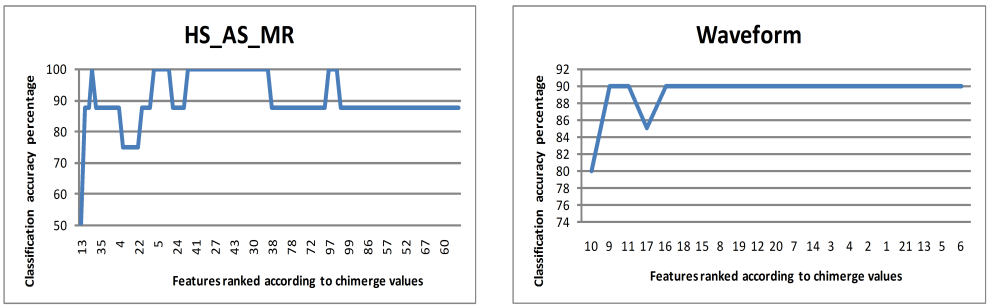
Figure 6.8: Variation of classification accuracy for indian diabetes and yeast data sets



(a) WDBC (b) Hepatitis

Figure 6.9: Variation of classification accuracy for WDBC and Hepatitis data sets

Statistical and classification Comparisons: Table 6.18 shows the results of other classifiers in comparison to the modified FCM technique where it appears to have best results when applied on the indian diabetes data sets. The classification accuracy is measured by the $FMeasure$ value , then multifidly this value by 100. Indian diabetes is a continuous multi-variate features' data set. So Fuzzy C-Mean model shows



(a) HS_AS_MR (b) Waveform

Figure 6.10: Variation of classification accuracy for HS_AS_MR and Waveform data sets

better results of 77.7% than other classifiers. As Decision trees and Decision tables tend to perform better when dealing with discrete/categorical features, while Bayes belief models like Bayesian Network (BN) and Naïve Bayesian Network due to the "independence assumption". Moreover, if feature selection is applied before the Fuzzy C-Mean classifier, the maximum classification performance appeared is 79.6% for five selected features. When applying the modification on the Fuzzy C-Mean model, the classification results is raised more to 83.3%.

Table 6.18: Comparison between different classification techniques and the modified FCM modified (M-FCM)

Classifier	Indian diabetes Classification accuracy
M-FCM	81.4%
Multilayer Perceptron	77.6%
IB1	69.2%
BFTree	75.9%
Decision Table	76.3%
Naïve Bayes	74.6%
Bayes Network	76.1%
LibSVM	75.2%

An experimental study is applied to measure the characteristics of each data set as shown in table 6.19. The first column in this table measures the average of discreteness in each data set. The discreteness value measures the repetition of values in attributes of the data set, where it is measured by the formula 6.3:

$$Disc = \frac{\sum_{i=1}^{n} \frac{max_{RepVal}}{m}}{n} \tag{6.3}$$

Where $Disc$ represents discreteness measure, n is number of attributes, m is the number of objects, and max_{RepVal} is the maximum number of repeated value in attribute i. The second column represents the average of the Kolmogorov-Smirnov test (K-S test) values of the attributes, K-S tests the degree of normality in the attribute (126). The third column represents finally the average of the standard deviation of the attributes in the data sets. Two observations are gained from table 6.19, the first is that indian diabetes and yeast data sets which show increase in the classification accuracy by the modified FCM have high K-S average and the highest standard deviation average, the second observation is that WDBC and waveform data sets which show the same classification accuracy before and after modification have the lowest discreteness average value and nearly the lowest standard deviation average.

Table 6.19: The results of different kernel functions in SVM

Data set	Discreteness	Normalization	Standard deviation
Indian diabetes	0.017	0.62	26.22
Yeast	0.11	0.75	37.73
WDBC	0.002	0.36	0.08
Hepatitis	0.35	1.0	8.80
HS_AS_MR	0.013	0.013	1.09
Waveform	0.001	0.76	1.51

6.4.2.2 Classification results for SVM

Table 6.20 shows the Macro averaged $Fmeasure$ of the modified model when applying fuzziness on the Euclidean calculations in the three types of kernel as shown in equations 4.50, 4.51 and 4.52. The results shows that the modification applied in the three types

of used kernels has a better results over the case when applying the ordinary SVM or even when applying feature selection before applying the SVM.

Table 6.20: The results of different kernel functions in SVM

	Linear Kernel	Polynomial kernel	Sigmoidal kernel
All features	73%	75%	73%
Selected features	75%	73%	73%
Euclidean fuzziness	81.6%	76%	78%

Training a support vector machine requires the solution of a very large quadratic programming (QP) optimization problem. Sequential minimal optimization (SMO) breaks this large QP problem into a series of smallest possible QP problems (127). These small QP problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop. The new SVM learning algorithm is called Sequential Minimal Optimization (or SMO). Instead of previous SVM learning algorithms that use numerical quadratic programming (QP) as an inner loop, SMO uses an analytic QP step. In the SMO iterations, the examples through the entire training set are started at random locations, in order not to bias SMO towards the examples at the beginning of the training set. When applying the comparison between SMO before and after fuzzification, the change in these random locations causes unfair comparison in the results between these two cases. To guarantee the fairness in comparison and to avoid this discrepancy in the results, The same random locations are applied in both cases. Two criteria are used in the comparison applied between SMO and SMO when fuzzification is applied, the first criterion is the classification accuracy resulted and the second criterion is the number of iterations required to achieve convergence. As shown in table 6.21, it is noticed from the results that number of iterations in the case of applying fuzzification on SMO technique is decreased, which means an enhancement in the performance of the SMO technique. Also an enhancement in the classification accuracy percentage is recognized in cases like the indian diabetes, WDBC and Hepatitis data sets. The kernel used in the SMO technique is the Polynomial kernel.

Another three data sets are used in this comparison to ensure the importance of this modification. The data sets are WDBC, Hepatitis and Letters data sets, which are extracted from the UCI data base (118) and HS_AS_MR data set.

Table 6.21: A comparison between sequential minimal optimization before and after fuzzification

Data set used	Number of iterations		Classification accuracy %	
	Conventional	Fuzzified	Conventional	Fuzzified
Indian diabets	29298	1571	56.00	76.00
Yeast	33	28	96.70	96.70
WDBC	40835	38086	88.70	91.93
Hepatitis	7512	16	25.00	58.00
Letters A, B	331	134	99.00	99.00
HS_AS_MR	2691	39	75	87.5
Waveform	437	126	81.66	83.33

6.5 Experimental work on the proposed visualization technique

The proposed technique is applied on two medical which are considered as examples for a numerical and continuous data. The first data set is the Indian-diabetes data set and the second data set used is the Breast Cancer where the number of attributes is 9. The tool used to generate the lattice is called Conflexplore which is a FCA (Formal Concept Analysis) tool for context editing and analysis and visualizing. Conflexplore is web-based application in the openFCA project (128). The Concept lattice construction in this tool is divided into two subtasks namely, the generation of all concepts using the In-Close algorithm, then Alaoui's algorithm can be used to correctly link the concepts together (129), (112).

6.5.1 Visualization results

6.5.1.1 Indians-diabetes data set

Table 6.22 shows the results when perform chimerge technique to evaluate the attributes in case of Indian-diabetes data set. Only four attributes are selected out of eight attributes which are $8, 1, 6, 2$, which are sorted according to their chisquare χ^2 values as follows $28.19, 18.66, 18.462, 16.97$ respectively.

Table 6.22: Indians-diabetes attributes chisquare

Attribute	Chisquare χ^2
8	28.19
1	18.66
6	18.462
2	16.97
3	0
7	0
4	0
5	0

Now the e value of each attribute according to the constructed lattice, shown in figure 6.11, after binarization of the input is calculated as shown in table 6.23, note that n_0 and n_1 are the number of objects that have the value 1 in each attribute in $class0$ and $class1$ respectively.

6.5.1.2 Breast Cancer data set

Table 6.24 shows the results when perform chimerge technique to evaluate the attributes in case of Breast Cancer data set. Only one attribute are selected out of nine attributes which is attribute 1, and its chisquare χ^2 value is 22.175.

Now the e value of each attribute according to the constructed, shown in figure 6.12, lattice after binarization of the input is calculated as follows in table 6.25, note that n_0 and n_1 are the number of objects that have the value 1 in each attribute in $class0$ and $class1$ respectively.

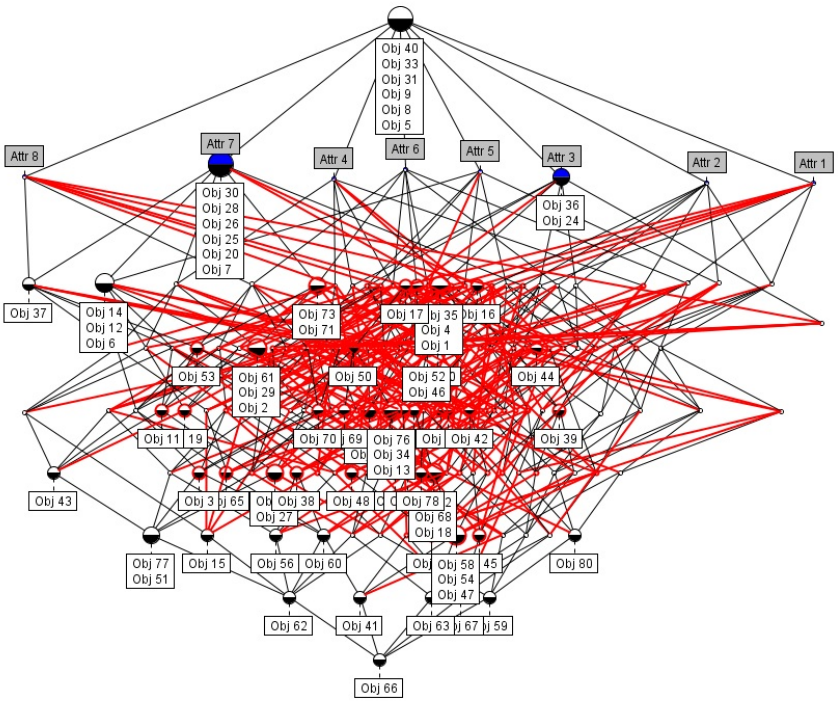


Figure 6.11: Indian-diabetes - Lattice generated

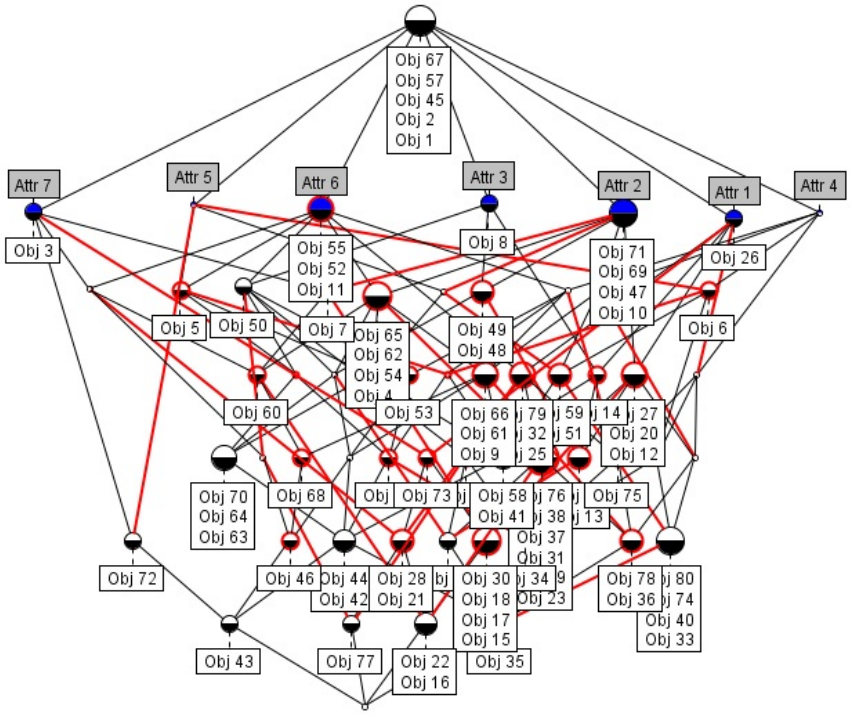


Figure 6.12: Breast Cancer - Lattice generated

Table 6.23: Attribute evaluation according to the constructed Indian-diabetes lattice

Attribute	n_0	n_1	e
8	9	31	0.55
1	10	29	0.45
6	13	31	0.40
2	1	15	0.87
3	16	24	0.2
7	30	30	0
4	9	18	0.225
5	10	14	0.1

6.5.2 Results analysis

Indians-diabetes and Breast Cancer are two continuous multi-variate features data set, Binarization is a must preprocessing before applying Formal Concept Analysis. To prove the validation of the constructed lattices, the e value is calculated for each attribute in both data sets. In case of Indians-diabetes data set, the e of the selected attributes from the chimerge technique, varying from 0.55 to 0.35 are the greatest values over the other attributes where their e are varying from 0.225 to 0. Also the sorting of the attributes according to the e values is exactly the same as according the chisquare χ^2 values. The same as in the case of Breast Cancer where only attribute is selected which is attribute 1. The previous scaling methods generate a huge number of scaled attributes in comparison to the original number of attributes. This is because they depend on the idea of producing for every numerical attribute a set of binary attributes corresponding to the scaled intervals in this attribute. The generated lattice becomes dense in comparison to the case when only the original number of attributes is used in lattice. Also the validity of the lattice should be considered as an important stage in the lattice generation which faces a high difficulty in the lattice generated based on the previous scaling methods. The validation applied in the experiments shows that the attributes of the highest interest of the classification problem appears to be of the highest discrimination among objects of different classes.

Table 6.24: Breast Cancer attributes chisquare χ^2 values

Attribute	Chisquare χ^2
1	22.175
4	0
2	0
3	0
8	0
9	0
5	0
7	0
6	0

Table 6.25: Attribute evaluation according to the constructed breast cancer lattice

Attribute	n_0	n_1	e
1	29	8	0.52
3	8	13	0.125
2	25	9	0.4
6	20	26	0.15
7	10	7	0.0175
4	7	17	0.25
5	6	14	0.35

6.6 Chapter conclusion

The general conclusion appears in the experimental work is that the avoidance of limitations of the data mining techniques increases the accuracy and correctness of these techniques. Presumption in data mining techniques could be presented into three categories which are discreetness, distribution, correlations and number of dimensions. Curse of dimensionality is considered as a critical problem in data mining. Preprocessing is considered as an important step in data mining, but it could affects in a negative way the internal structure of data. The proposed techniques in each of the data mining stages avoid these presumption in order to enhance the classification accuracy of the

results.

Chapter 7

Conclusion and future work

7.1 Conclusion

The objectives of the thesis were to discuss the effect of the medical data characteristics on the data mining techniques and to propose techniques that are capable of handling these characteristics. The results show that performance of data mining techniques increases if these techniques are independent on the characteristics of the data tested. In the case of real-life medical data sets, the presumptions that could be made by data mining techniques are not always true or applicable. Chapter 2 introduces the different stages of data mining, while chapters 3, 4 and 5 show the conventional and proposed techniques in each of the three stages of data mining. Chapter 6 shows the experimental work of the proposed techniques presented in the chapters 3, 4 and 5. The comparisons performed on each of the proposed techniques are applied on two or more data sets that are in the medical field.

Chapter 3 shows different steps of the preprocessing stage in data mining, then shows that some preprocessing algorithms do not preserve the structure of the input data sets. The results in the experimental work show that normalization method as preprocessing method for PCA may alter the original structure of data. Then Interval-based feature evaluation and selection technique is proposed to solve the problem of continuous data sets that is not handled probably in other feature selection techniques like chimerge. The results demonstrate that the proposed interval based method encouragingly outperforms most of the popular feature selection methods. The proposed technique shows

the best classification accuracy results with the least number of features. For example in the case applying MLP classifier on Indian-diabetes data set, the peak of accuracy has reached with 81.4% when the input data set contain only the first three selected attributes by the IB method, while the peak is reached when the number of features are six when using the entropy-based feature selection techniques is 75.9% only.
Chapter 4 shows the conventional machine learning techniques in data mining briefly and how these techniques are enhanced using hybridization and other ways. A frequent pattern-based classification technique is proposed to avoid problems discussed in other classification techniques under investigation. The proposed technique proves its efficiency and its competency according to the classification accuracy and the time of learning. Also a modification is applied on support vector machine and fuzzy c-mean classification techniques that are dependent on Euclidean space calculations. For example, in the case of fuzzy c-mean classifier, the classification accuracy of Indian-diabetes data set has raised from 77% to 83%, and classification accuracy of the yeast data set has raised from 56% to 64%.
Chapter 5 shows the formal concept analysis technique and how it depends on a presumption of the input data to be in a binary form. This presumption is handled by applying scaling method where it resulted in the generation of the highly complicated formal concept lattice. Then a proposed technique is presented to avoid the scaling method named as binarization method. Then novel method for validating the generated lattice based on the chimerge feature selection technique. The proposed technique of binarization has enhanced the generation of the formal concept lattice from $O(n*m)$ to $O(n)$, where n is the number of attributes and m is the increase in the number of attributes caused by the scaling method.

7.2 Future work

The future work needed in the field of medical data can be summarized in the following points:

Intelligent Diagnosis System for Liver Fibrosis in Chronic Hepatitis C Fibrosis liver disease is a deadly disease that adversely affects the lives of many people,

which is the final result of injury to the liver. Accurate assessment of the degree of fibrosis is important clinically, especially when treatments aimed at reversing fibrosis are being evolved. Liver biopsy has been considered to be the gold standardto assess fibrosis. However, liver biopsy being invasive and is not favored by patients or physicians, alternative approaches to assess liver fibrosis have assumed great importance. Moreover, therapies aimed at reversing the liver fibrosis have also been tried lately with variable results. Computer-assisted reading of medical images is a relatively new concept which has been developed during the last 10 years and which is growing into diagnostic radiology. Especially in liver fibrosis, image processing techniques and automated pattern recognition schemes is applied to assist radiologists in the interpretation of liver fibrosis.

An objective is needed to be handled which is to diagnose the liver disease using an application of the bio-inspiring technology so that it can shorten the medical diagnostic process and help the physician in the complex cases which are otherwise difficult to perceive. As well as it could be needed to assess performance and promise of radiologic modalities and techniques as alternative, noninvasive assessment of hepatic fibrosis.

An Integrated Intelligent System to HCV Chronic CV is the main cause of liver cirrhosis and liver cancer in Egypt and, indeed, one of the top five leading causes of death. The magnitude of the terrifying Hepatitis C disease in Egypt has motivated us to propose the development of data mining and machine learning tools that are able to handle and analyze the medical data of the patients suffering from HCV with the intent of incorporation into the learning model the domain-specific knowledge that our consortium team have. The application of the machine learning techniques to the HCV problem may help optimize costly trials by guiding key decisions such as dose selection and length of dosing. Egypt is in real need of advanced multidisciplinary research in order to combat the hepatitis C epidemic. It is well-known that there are many factors that affect a successful treatment outcome. When people are trying to make a decision about whether or not to be treated it is important to take many of these predictors of treatment response into consideration. However, it is also important to remember that the predictors to treatment response are there to help guide people in the decision making process;

they should never be used to deny or discourage treatment for anyone. Also, just because someone does not fall into these categories, it doesn't mean that they will not have a successful treatment outcome. Many people who have achieved a successful treatment outcome do not fall into any of these categories.

The objective now is to develop an integrated intelligent system based Data Mining technology for treatment response of Hepatitis C virus (HCV) treatment predictors in Egypt. The proposed system will utilize the new technology of data mining and knowledge discovery in data bases. These technologies will be integrated with machine learning algorithms, optimization models, statistical models and predictive models in the proposed intelligent system, to discover most important factors for treatment response which affect to HCV patients profiles.

Data evaluation methodology Investigation on producing evaluation techniques that are applied on the input data. This is required in order to select the most applicable classifier. The selection of the appropriate classifier is a data dependent problem that needs some kind of methodology that can successfully determine which method is required.

Automatic classifier selection Recently, we have seen a new era of machine intelligence emerging that is focussing on the principles, theoretical aspects, and design methodology of algorithms gleaned from nature. The required to accomplish in our future work is develop a methodology that is capable to determine the behavior of the hybrid classifier and expect the accuracy results according to the input data. The behavior of the classifier and the nature of the input data sets and their relation still a problem that needs a lot of research.

Data repository Implement a data repository web site that is specialized in the medical field. This repository is not only specialized to gather medical data about different diseases, but also make a kind of unifying the data retrieved from different resources about the same disease. This unifying procedure could open a very important research area as it could help in gathering and increasing the data sets and hence increases the ability to accurately applying classification. Also it helps in comparison between different experiences of treatment and analysis of diseases from different resources.

Medical social network Implement a kind of social network that is capable to gather different types of groups including patients, doctors and medical or health institutes. This social network could help in the sharing of experience of different terminals and different perspectives. Also it could be considered as a help to distribute medical advices in different cases in a more friendly way that is global to different and scaled audience.

Diseases analysis Analysis on the symptoms of different diseases to discover the relation between them. An example of such relations between diseases are like the cause-and-effect, similarities. This could provide an important contribution to the medical field.

References

[1] Blaz Zupan Riccardo Bellazzi. **Predictive data mining in clinical medicine**. *International Journal of Medical Informatics, vol. 77, no. 2, pp. 81-97*, 2008. 2

[2] H. Yokoi T.D. Nguyen S. Kawasaki S.Q. Le T. Suzuki K. Takabayashi, T.B. Ho and O. Yokosuka. **Temporal Abstraction and Data Mining with Visualization of Laboratory Data**. *Proceedings of the MedInfo, pp. 1304-1308*, 2007. 2

[3] B. Uma Shankar Ashish Ghosh and Saroj K. Meher. **A novel approach to neuro-fuzzy classification**. *Neural Networks, vol. 22, no. 1, pp. 100-109*, 2009. 2

[4] Chuan-Liang Chen Yun-Chao Gong. **Semi-supervised method for gene expression data classification with gaussian and harmonic Functions**. *19th International Conference in Pattern Recognition, pp. 1-4*, 2008. 2

[5] Madhavi Latha T. Padma and K. Jayakumar. **ecision making algorithm through LVQ neural network for ECG Arrhythmias**. *ICBME 2008, Proceedings 23, pp. 85-88*, 2008. 2

[6] Jose-Luis Sancho-Gomez Pedro J. Garcia-Laencina and Anibal R. Figueiras-Vidal. **Pattern classification with missing data: a review**. *Neural Computing and Applications, vol. 19, pp. 263-282*, 2009. 3, 55

[7] Larry Hatcher Norm O'Rourke and Edward J. Stepanski. **A step-by-step approach to using SAS for univariate & multivariate statistics**. *Second Edition, SAS Institute Inc, USA, ISBN 1-59047-417-1*, 2005. 5, 9, 22

[8] Paul R. Rosenbaum. **Causal Inference in Randomized Experiments**. *Springer Series in Statistics, Design of Observational Studies, part 1, pp. 21-63*, 2010. 5, 22, 56

[9] Peter Riben Mila Kwiatkowska and Krzysztof Kielan. **Interpretation of Imprecision in Medical Data**. *Advances in Data Management, Studies in Computational Intelligence, vol. 223, pp. 351-369*, 2009. 6

[10] C. Shang and Q. Shen. **Aiding classification of gene expression data with feature selection: a comparative study**. *Computational Intelligence Research, vol. 1, pp. 68-76*, 2006. 6, 23, 54

[11] Nickolas Savarimuthu Sarojini BalaKrishnan, Ramaraj NarayanaSwamy and Rita Samikannu. **Feature Selection using FCBF in TYPE II Diabetes Databases**. *Proceedings of 7th Annual Conference on Information Science, Technology and Management New Delhi*, 2009. 8

[12] Zhi-Hua Zhou. **Rule extraction: using neural networks or for neural networks?** *Journal of Computer Science and Technology, vol. 19, no. 2, pp. 249-253*, Mar. 2004. 8

[13] Cecilio Angulo Haydemar Nunez and Andreu Catala. **Rule extraction from support vector machines**. *Proceedings of the European Symposium on Artificial Neural Networks Bruges (Belgium), pp. 107-112*, Apr. 2002. 8

[14] Basilis Boutsinas Nikolaos Mastrogiannis and Ioannis Giannikos. **A method for improving the accuracy of data mining classification algorithms**. *Computers & Operations Research, vol. 36, no. 10, pp. 2829-2839*, 2009. 9, 22

[15] Huan Liu and Rudy Setiono. **Chi2: attribute Selection and Discretization of Numeric Attributes**. *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, Virginia, USA, pp.388*, Nov. 8, 1995. 9, 25

[16] Lev Klebano1 Xing Qiu, Andrew I Brooks and Andrei Yakovlev. **The effects of normalization on the correlation structure of microarray data**. *BMC Bioinformatics, vol. 6, pp. 120-131*, 2005. 9

[17] C.A.Cumbaa M.Maziarz J.Glasgow M.S.Tsao M.Sultan, D.A. Wigle and I. Jurisica. **Binary tree-structured vector quantization approch to clustering and visualization microarray data**. *Bioinformatics, vol. 18, pp. s111-s119*, 2001. 9

[18] Thair Nu phyu. **Survey of classification techniques in data mining**. *Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2009, Hong Kong, vol. 1, pp 727-731*, 2009. 9, 21, 37

[19] Stefano Monti and Gregory F. Cooper. **Learning hybrid Bayesian networks from data**. *Proceedings of the NATO Advanced Study Institute on Learning in graphical models, Erice, Italy, pp. 521-540*, 1998. 9

[20] Marcel J. T. Reinders Carmen Lai and Lodewyk F. A. Wessels. **Random subspace method for multivariate feature selection**. *Pattern Recognition Letters, vol. 10, pp.1067-1076*, 2006. 9, 55

[21] Pierre Geurts. **Pattern extraction for time series classification**. *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, pp. 115 - 127*, 2001. 9, 22, 55

[22] Vanling Li and Li Song. **Threshold determining method for feature selection**. *Proceedings of the 2009 Second International Symposium on Electronic Commerce and Security, vol. 2, pp. 273-277*, 2009. 9, 55, 56

[23] Sergei O. Kuznetsov Mehdi Kaytoue, Sbastien Duplessis and Amedeo Napoli. **Two FCA-Based Methods for Mining Gen Expression Data**. *Lecture Notes in Computer Science, vol. 5548, pp. 251-266*, 2009. 10, 72

REFERENCES

[24] Zdenek Hork Včlav Snšel and Ajith Abraham. **Understanding Social Networks Using Formal Concept Analysis**. *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 03, pp.390-393*, 2008. 10, 72

[25] Aboul Ella Hassanien Mostafa A. Salama, Kenneth Revett and Aly A. Fahmy. **Interval-based attribute evaluation algorithm**. *Proceedings of the Fedrated conference on computer science and information systems, Szczecin, Poland, pp. 153-156*, Sep. 2011. 10, 30

[26] Aboul Ella Hassanien Mostafa A. Salama and Aly A. Fahmy. **Reducing the Influence of Normalization on Data Classification**. *Proceedings of the 6th International Conference on Next Generation Web Services Practices, Gwalior, India, pp. 609-703*, Nov. 2010. 10, 28

[27] Aly E. Fahmy Mostafa A. Salama, Abol Ellah Hassanien. **Uni-Class Pattern-based Classification Model**. *Proceeding of the 10th IEEE International Conference on Intelligent Systems Design and Applications, Cairo, Egypt, pp.1293-1297*, Dec. 2010. 10, 56

[28] Aly E. Fahmy Mostafa A. Salama, Abol Ellah Hassanien. **Pattern-based Subspace Classification Model**. *Proceeding of the second World Congress on Nature and Biologically Inspired Computing (NaBIC), Kitakyushum, Japan, pp. 357-362*, Dec. 2010. 10

[29] Aly Fahmy Mostafa A. Salama, Aboul-Ella Hassanien. **Feature Evaluation Based Fuzzy C-Mean Classification**. *Proceeding of the IEEEFuzzy Systems Conference, Taibai, Taiwan, pp. 2534-2539*, 2011. 10, 64

[30] Gerd Stumme Bernhard Ganter and Rudolf Wille. **Using Formal Concept Analysis for the Extraction of Groups of Co-expressed Genes**. *Communications in Computer and Information Science, vol. 14, pp. 439-449*, 2008. 14

[31] Sbastien Duplessis Mehdi Kaytoue-Uberall and Amedeo Napoli. **Formal Concept Analysis: Foundations and Applications**. *Lecture Notes in Artificial Intelligence, no. 3626, Springer-Verlag*, eds. 2005. 16

[32] Joseph M. Juran and A. Blanton Godfrey. **Juran's Quality Handbook**. *Fifth Edition, McGraw-Hill*, 1999. 17

[33] G. Karypis A. Gupta and V. Kumar. **Highly scalable parallel algorithms for sparse matrix factorization**. *Parallel and Distributed Systems, IEEE Transactions on, vol. 8, no. 5, pp. 502520*, May 1997. 18

[34] Niall Hurley and Scott Rickard. **Comparing Measures of Sparsity**. 2009. 18

[35] Bruce cooil Sajeev varki and Roland T. Rust. **Modeling Fuzzy Data in Qualitative Marketing Research**. *Journal of Marketing Research, vol. 37, pp. 480489*, 2000. 19

[36] Charu C. Aggarwal and Philip S. Yu. **Outlier Detection for High Dimensional Data**. *Proceedings of the 2001 ACM SIGMOD international conference on Management of data, pp. 37-46*, 2001. 20

[37] M. I. Petrovskiy. **Outlier Detection Algorithms in Data Mining Systems**. *Programming and Computer Software, vol. 29, No. 4, pp. 228237*, 2003. 20

[38] Yu L. Liu, H. **Toward integrating feature selection algorithms for classification and clustering**. *IEEE Trans. on Knowledge and Data Engineering, vol. 17, pp. 112*, 2005. 23

[39] J. Doak. **An Evaluation of Feature Selection Methods and Their Application to Computer Security**. *University of California at Davis, Tech. Rep. CSE-92-18*, 1992. 23, 107

[40] A. Prieto J. Cabestany and D.F. Sandoval. **Heuristic Search over a Ranking for Feature Selection**. *LNCS, vol. 3512, pp. 742749*, 2005. 24

[41] Abdelwadood Moh'd. **Chi Square attribute Extraction Based Svms Arabic Language Text Categorization System**. *Journal of Computer Science, ISSN 1549-3636, vol. 3, pp. 430-435*, 2007. 25

[42] Fuhui Long Hanchuan Peng and Chris Ding. **Feature selection based on mutual information: criteria of max dependency, max relevance, and min redundancy**. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, pp.1226-1238*, 2005. 25

[43] Yun Ze Cai Jin Jie Huang and Xiao Ming Xu. **A parameterless attribute ranking algorithm based on MI**. *Neurocomputing, vol. 71, pp. 16561668*, 2008. 25

[44] Isabelle Guyon and Andre Elisseeff. **An Introduction to variable and feature selection**. *Journal of Machine Learning Research, vol. 3, pp. 1157-1182*, 2003. 26

[45] Javier Bussi Maria G. Borgognone and Guillermo Hough. **Principal component analysis in sensory analysis: covariance or correlation matrix**. *Food Quality and Preference, vol 12, pp 323-326*, 2003. 28

[46] Xiang Sean Zhou Thomas Ira Cohen, Qi Tian and S. Huang. **Feature selection using principal feature analysis**. *Proceedings of the 15th international conference on Multimedia table of contents, Augsburg, Germany, vol. 51, pp. 301-304*, 2007. 28

[47] Zdzisiaw Pawlak. **Rough set approach to knowledge-based decision support**. *European Journal of Operational Research, vol. 99, pp. 48-57*, 1997. 32

[48] Nguyen H. Son and Skowron A. **Discretization methods in data mining**. *Rough Sets in Knowledge Discovery, Heidelberg: Physica-Verlag, pp. 451-482*, 1995. 32

[49] Swiniarski R.W. and Skowron A. **Rough set methods in feature selection and recognition**. *Presented at Pattern Recognition Letters, pp.833-849*, 2003. 32

[50] Zhang M. and Yao J.T. **A rough sets based approach to feature selection**. *In Proceedings of The 23rd International Conference of NAFIPS, pp. 434-439*, 2004. 32

[51] Aboul Ella Hassanien Omar S. Soliman and Nashwa El-Bendary. **A rough clustering algorithm based on entropy information**. *In proceedings of the 6th International Conference on Soft Computing Models in Industrial and Environmental Applications SOCO 2011, Series of Advances in Intelligent and Soft Computing, vol. 87, pp. 213-222*, 2011. 32

[52] Dong J.Z. Zhong N. and Ohsuga S. **Data mining: A probabilistic rough set approach**. *Rough Sets in Knowledge Discovery, Heidelberg: Physica-Verlag, vol. 2, pp. 127-146*, 1998. 32

[53] S. B. Kotsiantis. **Supervised Machine Learning: A Review of Classification Techniques**. *Informatica, vol. 31, pp. 249-268*, 2007. 36

[54] J. R. Kender H. M. Malik. **Classification by Pattern-based Hierarchical Clustering**. *Local Patterns to Global Models Workshop, Belgium*, 2008. 37

[55] R. Lin. **An intelligent model for liver disease diagnosis**. *Artificial Intelligence in Medicine, vol. 47, pp. 53–62*, 2009. 37

[56] M. Kamber J. Han. **Data Mining**. *Belgian-Dutch Conference on Artificial Intelligence, Netherland*, 2001. 40, 46

[57] Aly Fahmy Mostafa A. Salama, Aboul-Ella Hassanien. **Deep Belief Network for clustering and classification of a continuous data**. *IEEE International Symposium on Signal Processing and Information Technology, Luxor, Egypt, pp. 473-477*, 2010. 41

[58] A. K. Noulas and B.J.A. Krse. **Deep Belief Networks for Dimensionality Reduction**. *Belgian-Dutch Conference on Artificial Intelligence, Netherland*, 2008. 41

[59] H.Larochelle and Y.Bengio. **Classification using discriminative restricted boltzmann machines**. *In Proceedings of the 25th international conference on Machine learning, vol. 307, pp. 536-543*, 2008. 41

[60] L. McAfee. **Document Classification using Deep Belief Nets**. *CS224n, Sprint*, 2008. 42

[61] T. Verwoerd and R. Hunt. **Intrusion detection techniques and approaches**. *Computer Communications, vol. 25, pp. 1356-1365*, 2002. 43

[62] H.Larochelle and Y.Bengio. **Classification using discriminative restricted boltzmann machines**. *In Proceedings of the 25th international conference on Machine learning,vol. 307, pp. 536-543*, 2008. 44

[63] G. Dahl A. R. Mohamed and G. E. Hinton. **Deep belief networks for phone recognition**. *NIPS 22 workshop on deep learning for speech recognition*, 2009. 44

[64] J. Louradour H. Larochelle, Y. Bengio and P. Lamblin. **Exploring Strategies for Training Deep Neural Networks**. *Journal of Machine Learning Research, vol.10, pp.1-40*, 2009. 44

[65] A. K. Noulas and B.J.A. Krse. **Deep Belief Networks for Dimensionality Reduction**. *Belgian-Dutch Conference on Artificial Intelligence, Netherland*, 2008. 45

[66] H.Larochelle and Y.Bengio. **A statistical method for profiling network traffic**. *In proceedings of the First USENIX Workshop on Intrusion Detection and Network Monitoring (Santa Clara), CA. pp. 119-128*, 1999. 45

[67] R. A. Olshen L. Breiman, J. H. Friedman and C. J. Stone. **Classification and Regression Trees**. *Belmont, CA: Wadsworth*, 1984. 46

[68] O.T. Yildiz and O. Dikmen. **Parallel univariate decision trees**. *presented at Pattern Recognition Letters, pp. 825-832*, 2007. 46

[69] Papageorgiou E. I. **A new methodology for Decisions in Medical Informatics using fuzzy cognitive maps based on fuzzy rule-extraction techniques**. *Applied Soft Computing, ELSEVIER, vol. 11, pp. 500-513*, 2011. 46

[70] Kruse R. Nauck D., Klawonn F. **Foundations of Neuro-fuzzy Systems**. *Wiley, Chichester*, 1997. 46

[71] Endika Bengoetzea Teresa Miquelez and Pedro Larrånaga. **Evolutionary Computation Based on Bayesian Classifiers**. *Int. J. Appl. Math. Computer Science, vol. 14, No. 3, 335-349*, 2004. 47

[72] D. Geiger Friedman and M. Goldszmidt. **bayesian network classifiers**. *Machine Learning, vol. 29(2-3), pp. 131163*, 1997. 47

[73] Jagagev A. k. Dehuri S. Devi, S. and R. Mall. **Knowledge Discovery from Bio-medical Data Using a Hybrid PSO/Bayesian Classifier**. *International Journal of Computing Science and Communication Technologies, vol. 2, no. 1, pp. 364-371*, 2009. 48

[74] E.C. Tsang Defeng Wang, D.S. Yeung. **Weighted Mahalanobis Distance Kernels for Support Vector Machines**. *Proceeding of the IEEE Transactions on Neural Networks, vol. 18, pp. 1453 - 1462*, 2007. 50

[75] V. N. Vapnik. **The Nature of Statistical Learning Theory**. *Springer, NewYork*, 1995. 50

[76] S. D. Wang C. F. Lin. **Fuzzy support vector machine**. *IEEE Trans. Neural Networks, vol. 13, pp. 464471*, 2002. 50

[77] Y. Y. Wang J. H. Zhang. **A rough margin based support vector machine**. *Inf. Sci. vol. 178, pp. 22042214*, 2008. 50

[78] M. Sarkar. **Ruggedness measures of medical time series using fuzzy-rough sets and fractals**. *Pattern Recognition Letter, vol. 27, pp. 447454*, 2006. 50

[79] Qiang Heb Degang Chena and Xizhao Wangb. **FRSVMs: Fuzzy rough set based support vector machines**. *Fuzzy Sets and Systems vol. 161, pp. 596607*, 2010. 50

[80] Min-Ling Zhangm. **A k-Nearest Neighbor Based Multi-Instance Multi-Label Learning Algorithm**. *Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, vol. 2, pp. 207-212*, 2010. 51

[81] Min-ling Zhang Zhi-hua Zhou. **M.: Multi-Instance Multi-Label Learning with Application to Scene Classification**. *Proceedings of the Advances in Neural Information Processing Systems 19*, 2006. 51

[82] P. Lingras. **Applications of Rough Set Based K-Means, Kohonen SOM, GA Clustering**. *Transactions on Rough Sets, Lecture Notes in Computer Science, vol. 2, pp. 120-139*, 2007. 51

[83] R. Slowinski W. Ziarko Z. Pawlak Z., J. Grzymala-Busse. **Rough sets**. *Communications of the ACM, vol. 38, No. 11, pp. 89-95*, 1995. 53

REFERENCES

[84] Skowron A. Pawlak Z. **Rough Sets and Conflict Analysis**. *Studies in Computational Intelligence (SCI), vol. 37, pp. 3574*, 2007. 53

[85] H. Zedan H. S. Own1, W. Al-Mayyan. **Biometric-Based Authentication System Using Rough Set Theory**. *LNAI, vol. 6086, pp. 560569*, 2010. 53

[86] J.L. Breault. **Data mining diabetic databases: are rough sets a useful addition**. *Proceedings of the Computing Science and Statistics, vol. 33*, 2001. 53

[87] E. Eslami P. K. Dehkordy F. Kyoomarsi, H. Khosravi. **Optimizing Machine Learning Approach Based on Fuzzy Logic in Text Summarization**. *International Journal of Hybrid Information Technology, vol. 2, pp. 105-116*, 2009. 53

[88] S. B. Kotsiantis. **Supervised Machine Learning: A Review of Classification Techniques**. *Informatica, vol. 31 pp. 249-268*, 2007. 53

[89] Barabara Hammerc Thomas Villmann Tina Gewenigera, Dietlind Zlkeb. **Median fuzzy c-means for clustering dissimilarity data**. *Neurocomputing, vol. 73, pp. 1109-1116*, 2010. 54

[90] Abdelwadood Moh'd. **Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System**. *Journal of Computer Science, vol. 3, pp. 430-435*, 2007. 54

[91] c. Chen. **Design of PSO-based Fuzzy Classification Systems**. *Journal of Science and Engineering, vol. 9, No 1, pp. 63-70*, 2006. 55

[92] Sebastien Paris Abderrahmane Boubezoul and Mustapha Ouladsine. **Application of the cross entropy method to the GLVQ algorithm**. *Pattern Recognition, vol. 41, pp. 3173-3178*, 2008. 56

[93] M. Kamber J. Han. **Data Mining: know it all**. *Morgan Kaufmann, Second Edition*, 2009. 56

[94] W. Wang X. Zhang. **Mining coherent patterns from heterogeneous microarray**. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Virginia, USA, vol.17, pp.838-839*, Nov. 2006. 56

[95] M. Demel G.F. Ecker A.G.K. Janecek, W.N. Gansterer. **On the relationship between feature selection and classification accuracy**. *Proceeding of the JMLR Workshop and Conference, pp. 90-105*, 2008. 62, 64

[96] V. Vapnik C. Cortes. **Support vector networks,**. *Machine Learning, vol. 20, pp. 273297*, 1995. 64

[97] J. S. Taur G. H. Lee. **A Robust Fuzzy Support Vector Machine for Pattern Classification**. *International Journal of Fuzzy Systems, vol. 8, pp. 76-86*, 2006. 64

[98] D. Lai A. Shilton. **Iterative fuzzy support vector machine classification**. *Proceeding of the IEEEFuzzy Systems Conference, pp. 1391-1396*, 2007. 64

[99] Tai hoon Kim. **Procedure of Partitioning Data Into Number of Data Sets or Data Group**. *Communications in Computer and Information Science, vol. 78, pp. 104-115*, 2010. 65

[100] Zamzeer M. Hadi W. Thabtah F., Eljinini M. **Naïve Bayesian based on Chi Square to Categorize Arabic Data**. *11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, pp. 930-935*, 2009. 65

[101] C.S.P. Rao Surendra Kumar. **Application of ant colony, genetic algorithm and data mining-based techniques for scheduling**. *Robotics and Computer-Integrated Manufacturing, vol. 25, pp. 901-908*, 2009. 65

[102] S. K. Pal and A. Skowron. **Rough Fuzzy Hybridization: A New Trend in Decision Making**. *Springer-Verlag New York, Inc. Secaucus, NJ, USA*, 1999. 65

[103] A. Skowron S. K. Pal, W. Pedrycz and R. Swiniarsky. **Neurocomputing**. *Special Issue on Rough-Neuro Computing, vol. 36, pp. 1-4*, 2001. 65

[104] Fernando de la Torre Minh Hoai Nguyen. **Optimal feature selection for support vector machines**. *Pattern Recognition, vol. 43, pp. 584-591*, 2010. 66

[105] S. Mukkamala A.H. Sung Tamilarasan, A. and K. Yendrapalli. **Feature Ranking and Selection for Intrusion Detection Using Artificial Neural Networks and Statistical Methods**. *Proceedings of the International Joint Conference on Neural Networks, pp. 4754-4761*, 2006. 66

[106] Peter Eklund Peter Eklund Don Walker Richard Cole, Richard Cole and Don Walker. **Using Conceptual Scaling In Formal Concept Analysis For Knowledge And Data Discovery In Medical Texts**. *Proceedings of the Second Pacific Asian Conference on Knowledge Discovery and Data Mining*, 1998. 72

[107] Sergei O. Kuznetsov and Sergei A. Obiedkov. **Comparing Performance of Algorithms for generating Concept Lattices**. *Journal of Experimental and Theoretical Artificial Intelligence, vol. 14, pp. 189-216*, Apr. 2002. 72

[108] Beatrix Versmold Susanne Motameny and Rita Schmutzler. **Formal Concept Analysis for the Identification of Combinatorial Biomarkers in Breast Cancer**. *Lecture Notes in Computer Science, vol. 4933, pp. 229-240*, 2008. 72

[109] H. G. Fu and E. M. Nguifo. **Partitioning Large Data to Scale up Lattice-based Algorithm**. *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2003), Sacramento, California, USA, pp. 537-544*, Nov. 2003. 72

[110] Zamzeer M. Thabtah F., Eljinini M. and Hadi W. **Naïve Bayesian based on Chi Square to Categorize Arabic Data**. *Proceedings of the 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, pp. 930-935*, Jan. 2009. 73

[111] Surendra Kumar and C.S.P. Rao. **Application of ant colony, genetic algorithm and data mining-based techniques for scheduling**. *Robotics and Computer-Integrated Manufacturing, vol. 25, issue 6, pp. 901-908*, Dec. 2009. 73

[112] **Open FCA**. *OpenFCA Project, http://code.google.com/p/openfca*. 73, 111

[113] Sergei O. Kuznetsov. **Machine Learning and Formal Concept Analysis**. *Lecture Notes in Computer Science, vol. 2961, pp. 3901-3901*, 2004. 73

[114] Bernhard Ganter and Sergei O. Kuznetsov. **Pattern Structures and their Projections**. *Lecture Notes in Computer Science, vol. 2120, pp. 129-142*, 2001. 76

[115] Amedeo Napoli Mehdi Kaytoue, Sergei O. Kuznetsov and Sebastien Duplessis. **Mining Gene Expression Data with Pattern Structures in Formal Concept Analysis**. *Information Sciences, vol. 181, pp. 1989-2001*, May. 2011. 77

[116] Sergei O. Kuznetsov Bernhard Ganter, Peter A. Grigoriev and Mikhail V. Samokhin. **Concept-based Data Mining with Scaled Labeled Graphs**. *Lecture Notes in Computer Science, vol. 3127, pp. 94-108*, 2004. 77

[117] Suk-Hyung Hwang Eung-Hee Kim and Sung-Hee Choi. **Conceptual Analysis of Fuzzy Data using FCA**. *Proceedings of the 8th WSEAS International Conference on applied computer science (ACS'08), pp. 37-42*, 2008. 77

[118] **UCI machine learning repository**. *http://archive.ics.uci.edu/ml/datasets.html*. 83, 111

[119] **Chiba University hospital DataBase**. *http://lisp.vse.cz/pkdd99/*. 83

[120] Zafiropoulos E Maglogiannis I, Loukis E and Stasis A. **Support vectors machine-based identification of heart valve diseases using heart sounds**. *Computet Methods Programs Biomedical, pp. 47-61*, 2009. 85

[121] **KDD'99 dataset, Irvine, CA, USA, July, 2010**. *http://kdd.ics.uci.edu/databases*. 85

[122] Yiming Yang. **An evaluation of statistical approaches to text categorization**. *Inform Retrieval, vol. 1, pp. 69-90*, 1999. 86

[123] M. Pontil T. Poggio J.Weston S. Mukherjee, O. Chapelle and V. Vapnik. **Feature Selection for SVMs**. *Proceedings of Neural Information Processing Systems, pp. 668-674*, 2000. 88

[124] **Rough Set Exploration System (RSES)**. *Group of Logic, Institute of Mathematics, Warsaw University, Poland, http://logic.mimuw.edu.pl/ rses/*, 2004. 95

[125] **Weka: Data Mining Software in java**. *http://www.cs.waikato.ac.nz/ml/weka/*. 100

[126] Tsang W. W. Wang J. Marsaglia, G. **Evaluating Kolmogorovs Distribution**. *Journal of Statistical Software, vol. 8, no. 18, pp. 14*, 2003. 109

[127] John C. Platt. **Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines**. *Advances in Kernel MethodsSupport Vector Learning, vol. 208, pp. 1-21*, 1998. 110

[128] Ovidiu Sabou Paul Valentin Borza and Christian Sacarea. **OpenFCA, an Open Source Formal Concept Analysis Toolbox**. *Proceedings of IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), vol. 3, pp. 1-5*, 2010. 111

[129] Andrews S. **In-Close, a Fast Algorithm for Computing Formal Concepts**. *Proceedings of International Conference on Conceptual Structures (ICCS), Moscow*, 2009. 111

Declaration

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This thesis has not previously been presented in identical or similar form to any other Egyptian or foreign examination board.

The thesis work was conducted from Mostafa A. Salama to Cairo University under the supervision of prof. Aly A. Fahmy and prof. Aboul-Ellah Hassanien at faculty of computer science and information technology, Cairo University.

Cairo, Egypt

- **الفصل الرابع** (مرحلة التعليم الألي): يناقش هذا الفصل التقنيات المختلفه في التعليم الألي و يعرض مجموعه من الحلول المُقترحه لتجنب مشكلة الإفتراضات حول البيانات المُدخله. الحل الأول هو مقُترح بتقنية تصنيف ألي للبيانات معتمده على تقنية التجميع بناءاً على الأنماط المتكرره. و الحل الثانيه هو تعديل للتقنات على الحسابات الإقليدسيه (Euclidian calculation) مثل تقية جهاز الدعم الناقل (Support Vector Machine) . و الحل الأخير هو دراسه عمليه تظهر أهمية إستخدام تقنية المجموعات الخام (Rough set) لتجنب الحاجه لتقليل السمات الكونه للبيانات.

- **الفصل الخامس** (مرحلة التصوير): يشرح هذا الفصل إحدي تقنيات تصوير البيانات المهمه و هي تقنية (Formal concept analysis) و تعرض إحدي مشاكل هذه التقنيه و هي إفتراض أن البيانات ثنائية الشكل هو ما يتعارض مع معظم البيانات الطبيه. ثم يناقش الطريقه المُقترحه للتعامل مع البيانات التي تحتوي على قيم فيها تقارب و إستمرارية. و تحل هذه التقنيه المُقترحه مشكلة ال (Scaling) و التي تزيد من تعقيد الشعرية (lattice) و تجعله أكثر وضوحاً و سرعة في الأداء.

- **الفصل السادس** (التجارب العملية) : يناقش هذا الفيصل النتائج من التجارب العمليه لإثبات صحة التقنيات المُقترحه في كل مرحة من مراحل التنقيب في البيانات وإثبات قدرة هذه التقنيات المقترحة في التعامل مع خصائص البيانات الطبيه. و تتضمن النتائج مقارنة التقنيات المقترحه مع التقنيات المعتاده في مجال التنقيب في البيانات.

- **الفصل السابع** (الإستنتاجات و الخطة المستقبليه) : يوضح هذا الفصل الإستنتاجات الناتجه من البحث و من التجارب العمليه للحلول المقترحه. ثم يعرض الخطه المستقبلية المبنية على هذا البحث. وتشمل على بعض النقاط المهمه داخليا مثل التعامل مع فيرس سي لحل مشكلة التكاليف الباهظه للعلاج من هذا المرض. و كذلك تشمل على بعض النقاط المهمه خارجيا مثل التعامل البيانات ذات المصادر المختلفة و توحيد و عرض هذه البيانات على الشبكه الدوليه للمساعده في البحث العلمي في هذا المجال.

ملخص الرسالة

إكتشاف المعرفة في قواعد البيانات (KDD) يصف عملية البحث تلقائيا في كميات كبيرة من البيانات , و تتم عملية البحث عن الأنماط التي يمكن أن تُعبر عن المعرفة حول البيانات. ويعتبر التنقيب في البيانات هو الخطوة التحليلية لاكتشاف المعرفة في قواعد البيانات العملية. إن تقنية التنقيب في قواعد البيانات (Data Mining) يهدف إلى استخلاص المعلومات المخبأة فيها، حيث أن استخدامها يوفر للمؤسسات وأجهزة الأمن في جميع المجالات القدرة على إستكشاف أهم المعلومات في قواعد البيانات والتركيزعلى هذه المعلومات. تحتوي خطوة التنقيب في البيانات على ثلاث مراحل أساسيه, المرحلة الأولى هي تجهيز البيانات المُدخلة مثل تنقية البيانات المُدخلة. و المرحلة الثانية هي التعليم الألي مثل تصنيف (Classification) البيانات المُدخلة أما المرحلة الأخيرة فهي مرحلة التصوير . التعلم الألي (Machine learning) هو أحد فروع الذكاء الاصطناعي التي تهتم بتصميم وتطوير خوارزميات وتقنيات تسمح للحواسيب بامتلاك خاصية "التعلم".

التنقيب في البيانات التي هي من واقع الحياة مثل البيانات الطبية هو التحدي الرئيسي في تطبيقات اكتشاف المعرفة. تنوع خصائص البيانات المُدخلة من الصعب أن يتم التعامل معها عن طريق تقنيات التنقيب في البيانات الموجوده مثل تصنيف البيانات. معظم أساليب التنقيب في البيانات تحتوي على افتراضات مسبقه حول المدخلات مثل افتراض أن قيم البيانات في شكل منفرد (Discrete)، أو أن القيم قد وزعت في الشكل الجاوسي (Gaussian) أو افتراض الاستقلال بين السمات المميزه للبيانات. أيضا من أدوات التنقيب في البيانات هو التصوير , ومن أمثلة التصوير إستخدام التقنيه (Formal concept analysis) حيث أن هذه التقنيه أيضاً تحتوي على إفتراض أن البيانات المُدخلة تكون في شكل ثنائي.

قد تكون هذه الإفتراضات حول خصائص البيانات المُدخلة غير موجودة في معظم الأحيان في واقع الحياة مثل مجموعات البيانات الطبية. الأمر الذي أدى إلي فكرة أنه ينبغي على تقنيات التنقيب في قواعد البيانات ألا تكون مُحتوية على أية إفتراضات قد تنتهك. و بناءً عليه, تم إقتراح مجموعة من التقنيات في الثلاث مراحل الأساسيه للتنقيب في البيانات و هم تحضير البيانات (Preprocessing) و التعلم الألي و التصوير. هذه التقنيات المقترحه تتفادي أية إفتراضات حول البيانات المدخله مما يساعد على تحسين النتائج.

و تحتوي الرسالة على سبعة فصول:

- **الفصل الأول** (مقدمة): يحتوي هذا الفصل على تعريف للبيانات الطبيه و مصادرها و الخصائص المختلفه لهذه البيانات و قدرة تقنيات التنقيب في التعامل مع البيانات من مختلف الفروع الطبيه و يناقش المشكلات التي تواجه تقنيات التنقيب في التعامل مع خصائص هذه البيانات و يعطى نبُذه عن الحلول المُقترحه. ثم يعرض هذه الفصل نموذج تصويري لبقية الفصول في الرساله.

- **الفصل الثاني** (مراحل التنقيب في البيانات): يُناقش هذا الفصل عملية التنقيب في البيانات و المراحل الثلاثه المكونه لها. تنقسم هذه العمليه إلى ثلاثة مراحل مختلف, مرحلة تحضير البيانات و مرحلة التعليم الألى و مرحلة التصوير.

- **الفصل الثالث** (مرحلة تحضير البيانات): يشمل هذا الفصل التقنيات المختلفه في مرحلة تحضير البيانات لإتاحة البيانات للمرحلتين الأخيرتين في عملية التنقيب .وهذه المرحلة تحتوي على ثلاثة خطوات هي تنظيف البيانات و تحويل البيانات و إختصار السمات المميزه للبيانات. التقنيات المختلفه في هذه المرحله لها إفتراضات حول البيانات المدخله. من أمثلة هذه التقنيات , تقنة تحليل العنصر المبدئي المستخدمه في إستخراج السمات اللمميزه للبيانات حيث تفترض هذه التقنيه أن البيانات لها التوزيع الطبيعي. ثم يناقش هذا الفصل الحلول المقتحره في هذه الفصل , حيث تدور الحلول حول تجنب إستخدام خوارزميات التفريد (Discretization) و التطبيع (Normalization). و ذلك من خلال طرح تقنيه مقترحه بإسم (interval-based feature evaluation) و هي تقوم بتقييم السمات المميزة للبيانات حيث تتجنب إفتراض تفرد البيانات.

جامعة القاهرة

كلية الحاسبات و المعلومات

قسم علوم الحاسب

التنقيب في بيانات المعلومات الطبية

مقدمه من

مصطفى سلامة عبدالهادي محمد

رسالة مقدمة الى قسم علوم الحاسب ضمن متطلبات الحصول علي درجة

الدكتوراه في علوم الحاسب

يعتمد من لجنة الممتحنين: الإمضاء

1_ أ. د/ إسماعيل عبد الغفار إسماعيل .....

2_ أ. د/ أمير عطية .....

3_ أ. د/ على على فهمي .....

4_ أ. د/ أبوالعلا حسنين العطيفي .....

ديسمبر 2011, القاهرة

جامعة القاهرة

كلية الحاسبات و المعلومات

التنقيب في بيانات المعلومات الطبية

رسالة مقدمة الى قسم علوم الحاسب ضمن متطلبات الحصول علي درجة

الدكتوراه في علوم الحاسب

مقدمه من

مصطفى سلامة عبدالهادي محمد

ماجستير في علوم الحاسب
مدرس مساعد في
الجامعة البريطانية في مصر

المشرفون

الاستاذ الدكتور / علي علي فهمي
قسم علوم الحاسب
كلية الحاسبات و المعلومات
جامعة القاهرة

الاستاذ الدكتور / أبو العلا حسنين
قسم تكنولوجيا المعلومات
كلية الحاسبات و المعلومات
جامعة القاهرة

أكتوبر 2011, القاهرة