Faculty of Computer Science & Information Systems
Universiti Teknologi Malaysia

Final Report

Development of Compound Clustering Techniques Using Hybrid Soft-Computing Algorithms

PROJECT NUMBER: 04 - 02 - 06 - 0093 EA001
VOTE NUMBER: 74252
PROJECT LEADER: Assoc. Prof. Dr. Naomie Salim

ABSTRACT

Databases of molecular structures available to the pharmaceutical industry comprise millions of molecules. With the advent of combinatorial chemistry, a vast number of compounds can be available either physically or virtually, which can make screening all of them infeasible in terms of time and cost. Therefore, only a subset of the entire database that encompasses the full range of structural types of the underlying dataset needs to be selected for screening to maximise the likelihood of finding as many biologically distinct active compounds as possible in a screening experiment. One of most used compound selection method is cluster-based compound selection, which involves subdividing a set of compounds into clusters and choosing one compound or a small number of compounds from each cluster. Selecting only representative compounds from each cluster is based on the assumption that structurally similar molecules have similar properties. A good clustering method groups similar compounds together, to ensure all activity classes are represented, whilst separating active and inactive compounds into different sets of clusters, to avoid an inactive compound being selected as a cluster representative.

Hierarchical clustering methods such as Ward's and Group Average are considered industry standard for compound selection purposes. Previously, there is limited work on the clustering and classification of biologically active compounds into their activity based classes using fuzzy and neural network. Furthermore, it has been found that many of the biologically active molecular structures exhibit more than one activity in which case they can be used as drugs for the treatment of more than one disease. However, previous clustering methods on chemical compounds are mostly limited to hard partitioning, which allows a compound to belong to only one cluster.

In this work, neural, fuzzy and hybrid methods are utilized for the clustering of biologically active molecular structures into their corresponding activity classes. The methods have been evaluated for their performance on MDL's MDDR, NCI's AIDS and IDDB drug databases containing various biologically active classes of molecular structures. The neural network methods use a number of heuristics to find appropriate parametric values. Initially, the heuristics needs user intervention to select optimal values, which give poor results. To overcome this problem, fuzzy memberships have been employed to find optimal parameters. Since fuzzy clustering methods such as the fuzzy c-means and fuzzy G – K are computationally exhaustive in terms of time and memory requirements, a hierarchical approach have also been used in this work for their implementation. The hierarchical fuzzy clustering algorithm developed in this work assign the overlapping structures (structures having more than one activity) to more than one clusters if their fuzzy membership values are significantly high for those clusters.

When compared with industry standard methods, the neural networks show very poor performance when 2-D bit-strings descriptors are used. However, their relative performance improves when used with topological indices as descriptors. The fuzzy and fuzzy neural methods show slightly better results than the industry standard methods. The hierarchical fuzzy clustering method developed here is far better than a similar implementation of the hard k-means method. When used for overlapping structures, its performance improves significantly. Although the neural network methods are not very effective in clustering biologically active structures, their performance is remarkable when used as classifiers. The feed forward and radial basis functions networks show higher learning capabilities than support vector machines and rough set classifier in the classification of datasets comprising more than two classes. However, their performance is slightly inferior to that of support vector machines for binary classification of chemical structures into drug and non drug compounds.

TABLE OF CONTENT

Chapter 1 3
Introduction 3
1.1. Background of the Problem 5
1.2. Problem Statement 7
1.3. Objectives of the Study 8
1.4. Scope of the Study 9
1.5. Milestones of the Project 9
1.6. Research Frame Work 11
1.7. Research Contributions 12
1.8. Report Organization 13

Chapter 2 14
Chemoinformatics and Compound Clustering 14
2.1. Chemical Structures Representation 15
2.1.1. Fragmentation Codes 16
2.1.2. Linear Notations 17
2.1.3. Connection Tables 17
2.2. Molecular Descriptors 18
2.2.1. 2-Dimensional Descriptors 18
2.2.2. 3-Dimensional Descriptors 21
2.3. Classical Clustering Methods 23
2.3.1. Single Linkage Algorithm 26
2.3.2. Complete Linkage algorithm 26
2.3.3. Group Average Algorithm 26
2.3.4. Centroid Clustering Algorithm 26
2.3.5. Median Clustering Algorithm 27
2.3.6. Ward's Clustering Algorithms 27
2.3.7. Single Pass Algorithm 28
2.3.8. Jarvis Patrick's Algorithm 29
2.3.9. K-means Algorithm 29
2.4. Classification of Chemical Compounds 30

Chapter 3 31
Neural Networks and Compound Clustering 31
3.1. Unsupervised Neural Networks 33
3.1.1. Kohonen Neural Network 33
3.1.2. Neural Gas Network 34
3.1.3 Enhanced Neural Gas (ENG) Algorithm 36
3.2. Supervised Neural Networks 37
3.2.1. Multi Layer Perceptron(MLP) 37
3.2.2. Radial Basis Function (RBF) Network 39
3.3. Other Machine Learning Methods 39
3.3.1. Support Vector Machines 39
3.3.2. Rough Set Classifier 41
3.4. Experiments and Results 42
3.4.1. Experiment 1 42
3.4.2. Experiment 2 47

3.4.3 Experiment 3 51
3.5. Summary 53

Chapter 4 .... **55**
Fuzzy Logic & Compound Clustering 55
4.1 Fuzzy Logic 55
4.2. Fuzzy Set Theory 56
4.3. Membership Function 57
4.4. Fuzzy Clustering 60
4.4.1. Hard and Soft Clustering 61
4.4.2 Fuzzy c-means 64
4.4.3 Fuzzy Gustafson- Kessel Algorithm 67
4.4.4 Modified Gustafson- Kessel Algorithm 69
4.5 Experiments and Results 72
4.5.1 Experiment 1 72
4.5.2 Experiment 2 76
4.6. Summary 80

Chapter 5 .... **82**
Hybrid Techniques and Compound Clustering 82
5.1. Fuzzy Kohonen Self-Organizing Feature Map 82
5.1.1. Fuzzy Kohonen Network (FKN) Algorithm 84
5.2. A Hierarchical Fuzzy Algorithm 87
5.3. Validity of Clustering for Fuzzy Hierarchical Algorithm 89
5.4. Genetic Algorithm and Clustering 91
5.5. Experiments and Results 92
5.5.1. Experiment1 92
5.5.2. Experiment 2 95
5.5.3. Experiment 3 103
5.5.4. Experiment 4 106
5.6. Summary 110

Chapter 6 .... **111**
Discussions and Conclusions 111
6.1. Research Objectives and their Achievements 111
6.2. Neural Networks and Chemical Structures 112
6.3. Fuzzy Clustering 113
6.4. Hybrid Clustering Methods 113
6.5. Conclusion 114

References **115**

Chapter 1

Introduction

Clustering and classification of chemical databases has tremendous significance in the process of in silico drug designing and discovery. In the field of drug designing and discovery the researchers want to find a subset of compounds or molecules from a large dataset of compounds that can potentially inhibit, block or stop the activity of a malignant protein or enzyme, for further rigorous studies of each and every molecule as a potential drug candidate. These rigorous studies include, engineering of new compounds chemically or genetically or modification or refinement of the current molecules for enhanced disease fighting activity, prediction and improvement of absorption, distribution , metabolism, excretion and other toxic (ADMET) properties, and studies conducted during clinical and pre-clinical administration of the drug for approval [1].

The computational systems used in drug discovery and designing are based on two approaches: one the structure based approach when the 3D structure of the biological target (protein or enzyme) involved in the disease is available, and second the ligand based approach where structural information of known drugs (highly active molecules) are mapped for mining similar molecules from large molecular database [2], and this second approach is the basis of this work. The molecular structure that comprises a number of atoms, special groups of atoms, the bonding arrangements between various atoms and groups, the bond angles and the bond lengths, is responsible for many activities and properties of the molecule. This whole research is based on the similar property principle which states that structurally similar molecules exhibit similar biological activities [3] .

Classification can be used to select the desired compound structures from a large database. In classification some of the compounds activities are known in advance and based on these compounds the classifier are trained and after training the classifier is used for the selection of similar compounds. Thus classification is a supervised mechanism, where the system is first trained with help of a priori known examples and then used to classify an unknown dataset into specific classes for which the system is trained.

Generally the main objective of Clustering (or cluster analysis) is to organize a collection of data items into some meaningful clusters, so that items within a cluster are more similar to each other than they are to items in the other clusters. This notion of similarity and dissimilarity may be based on the purpose of the study or domain specific knowledge. But one thing must be kept in mind that, from clustering, we always mean unsupervised classification of data, a course where there is no teacher to provide any guidance of the path. In other words, there is no pre notion about the groups and their number (may or may not be) present in the data set. Cluster analysis and classification are widely used in many diverse fields such as engineering, electronics, biology, medicine, archeology, the social sciences and astronomy, and thus a large number of algorithms have been developed.

Although, the research on the development of new and robust clustering methods is an on going process [4, 5], most of the methods used in the field of chemoinformatics are hierarchic in nature which are not good because of their higher time and space complexities but suits the partitioning of chemical structures because of its tree like branching of the datasets into biologically active groups. An important issue related with these methods is their non overlapping nature. In non-overlapping methods each object (molecule) belongs to only one cluster. These methods also uses distance measure like Euclidean which support only same size and shape for all the clusters. The clustering methods that support overlapping clusters, are efficient time wise and allow the clusters to have any independent shape and size are required for the clustering of natural and real databases like molecular structures.

Both clustering and classification are important data analysis approaches. The classification can be used in situation where partly some knowledge about some of the groups in the data is already known whereas clustering is used in situation where there is no knowledge available about the clusters in the dataset. So, classification is a trained process and the training is based on known examples of data and clustering is not based on training.

In this work a number of clustering methods based on fuzzy and neural networks are evaluated for the analysis of molecular databases. A few techniques based on the combination of hierarchical, neural and fuzzy approaches are also considered. For classification of chemical structures a number of approaches have been adopted such as the neural networks and support vector machines.

1.1. Background of the Problem

Chemoinformatics can be defined as the discipline of storing, processing and retrieving chemical information using computers. It is a relatively new discipline of science and technology that flourished well in the last 10 to 15 years. Chemoinformatics in its name encompasses all the fields related to the discovery and design and the processing of chemical compounds with the help of computers such as computational chemistry, computer chemistry, chemometrics, chemical information, and QSAR [6]. This field is differentiated from other data processing fields by such requirements as to work with chemical compounds or molecular structures. This requirement urged the researchers to find new ways and methods to represent, store, process and retrieve chemical compounds in computers.

The representation of molecular structures is very important in clustering and all other aspects of chemoinformatics. The molecular descriptors should be able to accurately represent the chemical properties and biological activities of the chemical compounds. The molecular descriptors can be categorized into three main categories namely 1-Dimentional, 2-Dimentional and 3-dimentional.

The 1-Dimensional descriptor actually refers to the physiochemical properties like molar refractivity (the ratio of the speed of light in a vacuum to its speed in a sample compound) [7], ClogP (Log of the octanol / water partition coefficient), principal moments of inertia, principal axes, volume of the inertial ellipsoid, molecular weight, etc. The 2-Dimentional descriptors are calculated from the 2-D structural graph representation of a molecule. They characterize structures according to their size, number of bonds, orientation of bonds and overall shape. Examples of 2-D fingerprints include topological indices, Kappa shape indices, electrotopological shape indices, and 2-D fingerprints. The 3-Dimensional descriptor is calculated from the 3-D structural graph of molecules and is highly dependent on the molecular conformations. Since, for the same molecule, the 3-D descriptors calculated for one conformation is quite different from the same descriptors of the same molecule for another conformation, one has to keep information for almost all of the conformations. Sometimes, 3-D descriptors for a few conformations of a molecule give misleading results. That is why most of the researchers have used 2-D descriptors.

The term cluster analysis was first used by Tryon in 1939 that encompasses a number of methods and algorithms for grouping objects of similar kinds into respective categories [8]. The aim of a cluster analysis method is to partition a given set of data or objects (compound structures) into clusters (subsets, groups, classes). This partition should have the following properties [9]:

- Homogeneity within the clusters, i.e. data that belong to the same cluster should be as similar as possible.
- Heterogeneity between clusters, i.e. data that belong to different clusters should be as different as possible.

In the last six decades of clustering life starting from Tryon [8] until now a large number of methods and ways had been developed that can cluster the underlying datasets. However the effects of each method are different, keeping in view the objective of clustering, the ways these methods work and the variability of the application area.

In computational chemistry, a large number of crisp non overlapping methods have been effectively used. These clustering methods can be divided into hierarchical and non hierarchical methods. In hierarchical methods in each step of iterative process, a pair of clusters is either merged together to form a new cluster or a bigger cluster is divided into two more homogeneous clusters, with a tree like parent/ child relationship established between clusters at each successive level of iteration. On the other hand, in non hierarchical methods, the data set is divided into a number of overlapping or non overlapping clusters in which there is no hierarchical relationships among the clusters.

The applications and evaluation of these methods in clustering chemical datasets have been discussed by many researchers. Downs and Bernard [10, 11] have given a good review of most of the clustering methods for processing of chemical structures dataset. The most significant work is done by Peter Willet and his students on the clustering of a variety of compounds databases using 2D fingerprint descriptors [12-14]. In [15] a number of methods like Wards [16] and Jarvis-Patrick [17] have been evaluated against a number of 2D and 3D fingerprints like MACCS fingerprints, MDL fingerprints and it has been shown that the performance of Wards hierarchical clustering method is the best.

A number of works [18-22] have been reported on the application of fuzzy c-mean for the clustering of chemical datasets but the work of Rodgers et al [23] gives a comprehensive analysis of the fuzzy c-means applications. Artificial neural networks had been employed for classification and clustering of chemical structures. Most of the works show results for the classification of compound structures into two binary groups of active and non-active structures [24-27].

In this work a number of methods have been developed based on the fuzzy, neural, hierarchical, rough sets, genetic algorithms and the hybrid methods based on the combination of these methods for the clustering and classification of chemical compounds and some of these methods show improved performance over the classical methods. Some of the techniques based on fuzzy memberships such as fuzzy and fuzzy hierarchical developed in this work are better for analyzing overlapping datasets. Many of the algorithms here are evaluated on multi class datasets as opposed to only binary class datasets which have been used in most of the studies so far conducted.

1.2. Problem Statement

In recent years, fuzzy logic and artificial neural networks have attracted considerable attention as candidates for novel computational systems, because of the variety of the advantages that they offer over conventional computational systems. If we think in terms of control systems, these methods can be regarded as model free systems as they do not require an exact mathematical model of the controlled process. These methods have the inherent capability to model non linear systems from the input –output data. They share a common framework of trying to mimic the human way of thinking and provide an effective and promising means of capturing the approximate, inexact and vague nature of the real world processes. If fuzzy logic has efficient and simple modeling abilities on one hand, neural networks have good learning capabilities on the other hand. Here in this work, we will evaluate the neural, fuzzy and traditional clustering methods and combine the merits of these techniques to develop more robust, reliable and intelligent methods for the clustering of chemical compounds.

Currently available compound clustering methods are unable to handle the overlapping, and vague nature of the natural compounds clusters. Since fuzzy logic generally has the ability to represent vague and approximate knowledge and fuzzy classifiers have inherent ability to cluster overlapping datasets, so, their field of applications can be extended to compound clustering. The neural networks on the other hand posses modeling capabilities of nonlinear processes and the unsupervised methods like Kohonen neural network are also efficient due to only two neuronal layers. The combination of the two methods can give better results than the traditional methods. A hierarchical fuzzy hybrid scheme is also developed and investigated for the clustering of chemical structures. The results of the method are very encouraging as compared to a similar implementation of the hard k-means method.

In this work besides neural network and fuzzy logic some other approaches like rough sets, support vector machines and genetic algorithms have been investigated for improved analysis of the chemical structure databases.

In this work it has been tried to find answers to the following questions:

- How can new algorithms be developed that are robust, reliable, efficient and their performance is better than current methods in clustering drug like molecular structures into their activities?
- How well the fuzzy clustering methods like fuzzy c-means, G-K Algorithm, GG Algorithm can be extended for the clustering of multi class and overlapping chemical datasets?
- What are the main neural network techniques that can improve the clustering and classification of chemical structures?
- What are the advantages of fuzzy and neural clustering methods?
- How can fuzzy and unsupervised neural network methods be combined to develop new and refined data analysis tools for chemical datasets?
- How can other methods like rough sets and genetic algorithms be employed to improve the performance of data analysis tools?

1.3. Objectives of the Study

This research work will investigate theoretically and experimentally the fuzzy methods, neural networks, and the combination of fuzzy logic, neural networks and other data analysis schemes for the resolution of the problem of clustering and classification of multidimensional chemical datasets that exhibit drug like characteristics and are also overlapping. Recently, fuzzy and unsupervised neural methods have been extensively used in many areas of pattern recognition and prediction. This work will result in the development and evaluation of methods that will give more accurate results, will be more reliable, and more robust in predicting natural biological clusters in the underlying chemical datasets. This study will focus on development of clustering methods for finding natural and variable shape and size clusters in chemical datasets.

The main objectives of this study can be summed up in the following lines:

- To develop robust, efficient, reliable and high performance algorithms for the clustering of chemical compounds into biologically active classes.
- To develop fuzzy and neural methods for clustering of chemical datasets.
- To enhance and improve the current fuzzy and neural clustering methods for application in clustering of chemical datasets.
- Since the chemical data is multidimensional, so, our objective is to look for clustering methods that could effectively and accurately cluster chemical datasets.
- The chemical datasets are considered large databases of thousands of compounds structures, so, it is our objective to develop clustering methods based on fuzzy and neural paradigms to be able to cluster large datasets.

1.4. Scope of the Study

This work will focus on the clustering and classification of chemical datasets with the help of traditional hierarchical, fuzzy, neural and hybrid methods such that the objectives of the research outlined in section 1.3 are achieved. The study outlines the different methods used for clustering and classification of chemical compounds in this project. There are a large number of representation schemes available for representing the chemical and biological properties of molecular structures, this work use only 2D descriptors such as topological descriptors and BCI fragment bits screen [28].

In this work chemical compound datasets from sources like MDDR drug database [29], NCI AIDS database [30] and Investigational drug database (IDDb) [31]will be used.

1.5. Milestones of the Project

The project is divided into four main milestones and a number of activities along with a tentative time line for the achievements of the milestones and objectives. The milestone and the timeline are briefly discussed here.

1. Development of Fuzzy based Compound Clustering

Fuzzy logic has been employed in almost all spheres of science and technology and during the course of time a number of robust and reliable clustering methods have been developed. The objective here is to employ these fuzzy based clustering methods for the clustering of chemical structure databases. A number of clustering methods such fuzzy c – means and Gustafson – Kessel will be utilized to improve the clustering of chemical structures. Besides development of algorithms, a number of other tasks such as data collection, data formatting will also be part of the schedule.

2. Development of Neural Network based Compound Clustering

In this part of the project algorithms based on neural networks will be developed for the clustering of chemical structures databases. As discussed in the literature review, a number of neural networks have been employed by the researchers for this purpose, but most of these works deals with problems like separation of chemicals into drug and non- drug classes. However, our approach is to partition the space of drug like compounds into their biologically active classes and so the problem becomes very precise and multi class instead of binary class. Here a number of networks such as Kohonen Self- Organizing Maps (SOMs), Neural Gas, and adaptive resonance theory based networks will be employed.

3. Development of Hybrid Compound Clustering

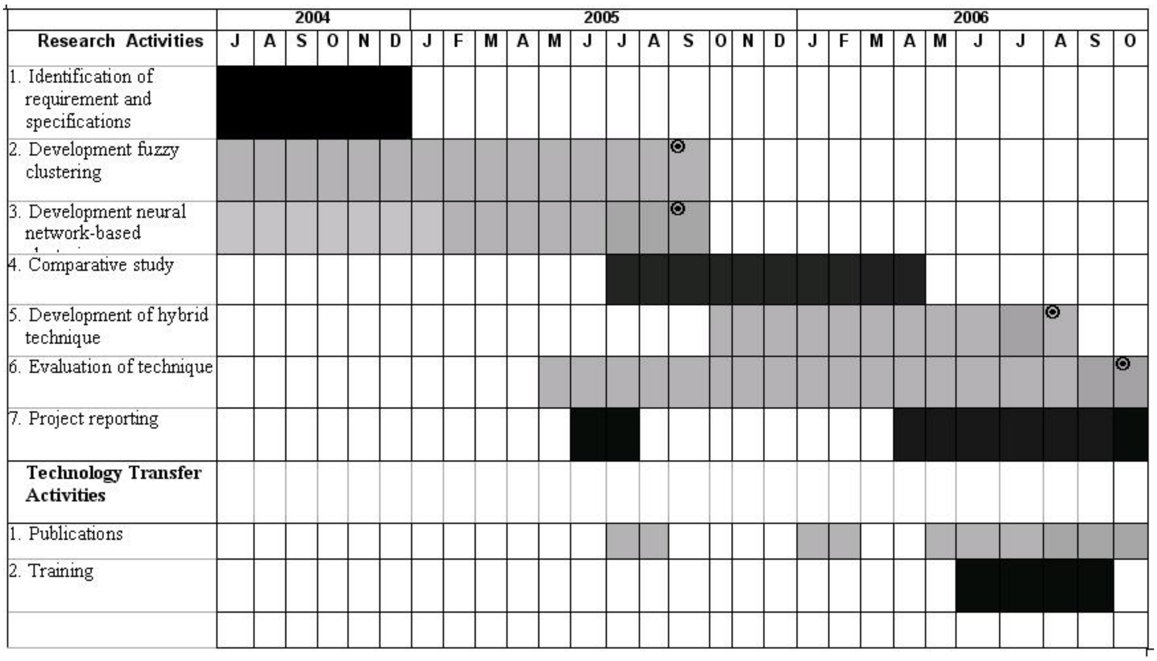
Based on the achievements of the previous two milestones, the good traits of the three approaches, neural, fuzzy and classical hierarchical clustering will be combined to arrive at more robust, reliable and more easy to use compound clustering methods.

4. Evaluation of Compound Clustering Techniques

In this part of the work, the methods developed in the previous three milestones, will be compared with the existing benchmark compound clustering techniques. Besides utilizing some of the benchmark and artificial datasets, a number of drug datasets of compounds with varying characteristics will be developed for the comparison of the techniques.

The timeline of the project is composed of a number of other activities besides these four milestones which are depicted in the project schedule shown in Table 1.1.

Table 1.1 Project Schedule



	2004						2005												2006									
Research Activities	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O
1. Identification of requirement and specifications	■	■	■	■	■	■																						
2. Development fuzzy clustering	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	⊙													
3. Development neural network-based	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	⊙													
4. Comparative study													■	■	■	■	■	■	■	■	■	■						
5. Development of hybrid technique																▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	⊙		
6. Evaluation of technique											▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	⊙
7. Project reporting												■	■									■	■	■	■	■	■	■
Technology Transfer Activities																												
1. Publications													▨	▨					▨	▨			▨	▨	▨	▨	▨	▨
2. Training																								■	■	■	■	

1.6. Research Frame Work

This research had been performed in two stages, stage1 was dedicated to the initial work which composed of the evaluation of the existing important clustering and classification methods based on hierarchical, neural and fuzzy approaches; and in stage 2 some of such techniques were combined to obtain hybrid methods for the clustering and classification of chemical databases.

The first phase of the research concentrated on the development of classical hierarchical methods like Ward’s, Group Average and Jarvis Patrick, fuzzy methods like fuzzy c-mean, fuzzy Gustafson Kessel and neural methods like neural gas, enhanced neural gas, kohonen self organizing maps etc. In this stage a number of classification methods have also been used for the analysis of chemical structures like back propagation neural network, radial basis function neural network and support vector machines.

In stage 2 a number of methods based on the combination of hierarchical methods, neural networks, fuzzy logic, genetic algorithms and rough sets had been employed for the clustering and classification of molecular structures.

This framework is shown in figure 1.1.

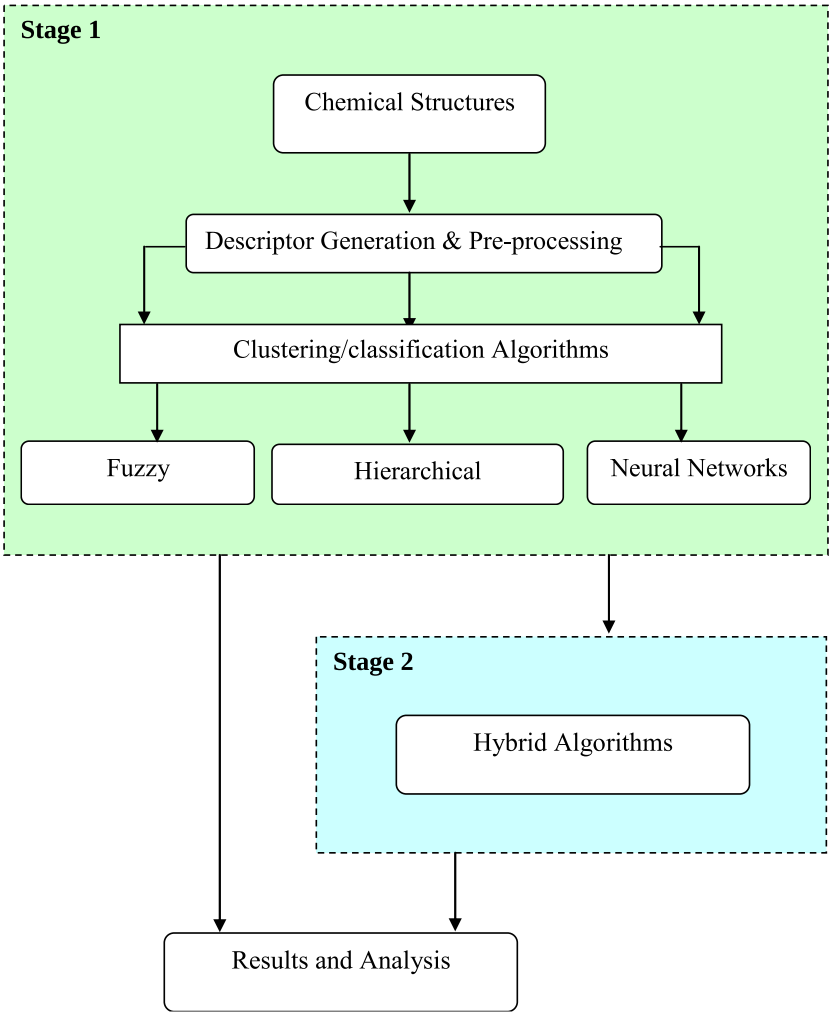


Figure 1.1 Research Frame work

1.7. Research Contributions

Clustering and classification of biologically active chemical compounds into their respective classes has a number of applications in almost all the phases of modern computer aided drug designing and other chemical information processes. Currently, most of the industrial systems are based on the traditional methods of clustering chemical compounds like the hierarchical Ward's and non-hierarchical Jarvis-Patrick methods. These methods suffer from a number of problems as described in detail in the next chapter, due to which the systems are not 100% reliable.

In this work a number of fuzzy based clustering methods which can be referred to as fuzzy objective function based methods, unsupervised neural methods and the combination of the hierarchical, fuzzy and neural methods have been developed and applied to the analysis of chemical structures for the first time. A few classification methods based on neural networks and support vector machines have also been considered for classification.

1.8. Report Organization

The rest of this report is organized into five chapters. The second chapter describes the relevant literature review, the fundamentals of chemoinformatics from the resolution of the problem of clustering and classification point of view. Chapter 3 reviews some of the neural networks methods used in this work for clustering and classification and presents the results. Chapter 4 presents the fuzzy clustering of chemical compounds and the results. In chapter 5 some other methods like support vector machines, rough sets and genetic algorithms are explored for the clustering and classification of chemical space. This chapter also discusses the hybrid methods developed in this work and presents the results. Chapter 6 is the last chapter, which discusses and concludes this work. This chapter also presents some directions for the future work.

Chapter 2

Chemoinformatics and Compound Clustering

The applications of computers in chemistry are almost as old as the arrival of computers themselves. Although first computer, the ENIAC (Electronic Numerical Integrator and Computer) was built for the US Army Ordinance Department in 1943 but its functionality was very limited and the first UNIVAC (Universal Automatic Computer) was delivered to the US Census Bureau in 1951. The first company Quantum Chemistry Program Exchange (QCPE) was formed at Indiana University to distribute quantum chemistry codes in 1963. But in fact the right start in computational chemistry was the foundations of two companies Health Design Inc and Molecular Design Inc in 1978, only three years later than the foundation of Microsoft in 1975, for the development of products in computational chemistry. The first product of the Health Design was TOPKAT a program for the prediction of molecular toxicity and that of Molecular Design Inc was the MACCS chemical database [32].

In drug discovery process, we often need a diverse set of compounds which are tested for their activities against a biological entity such as a protein, some cells, or an animal embryo. If the compounds tested are highly similar to one an other then the success rate of the experiment (which is often known as high-throughput screening) will be very less. Thus the need for diverse compound libraries had been felt and since 1990s a variety of structural processing technologies such as structural similarity searching, clustering and classification of structural datasets, structural descriptors generation, virtual libraries generation, quantitative property prediction, etc have been developed and applied.

Now a day the name of this field of research is chemoinformatics or chemiinformatics which deals with all the techniques for the processing of chemicals using their structural information. It includes the methods for storing and retrieval of chemical structures, ADMET and QSAR prediction, diversified compound selection, and compound docking etc.

In this chapter fundamentals of chemoinformatics which are important from the clustering and classification point of view are discussed. The clustering and classification methods developed and applied to chemical structure analysis are reviewed. When we talk about processing of chemical compounds by computers, the first question that comes into one's mind is how to represent molecules for their computational processing. The next section tries to give some representation schemes used in chemoinformatics. Since, most of the clustering and pattern recognition techniques characterize various objects in the dataset with the help of some features of the objects, known as descriptors in chemoinformatics are also discussed. Another important factor in the clustering and pattern recognition is the use of coefficient of similarity, dissimilarity or a distance measure between every pair of objects (structures) in the dataset. So, a number of important distance and similarity coefficients are described before giving the details about the clustering and pattern recognition techniques used in chemical structures analysis.

2.1. Chemical Structures Representation

Today, many chemistry and drug related organizations have their publicly accessible and proprietary databases of chemical compounds, containing large number of molecules; several hundred thousand is a common figure and some have even billion of compounds. Many have virtual libraries of compounds generated using computational techniques that can be converted to chemicals using combinatorial chemistry techniques.

Although there are 2-dimensional and 3-dimesional representation techniques available for chemical molecules but only the 2-D representation techniques are more popular. Four types of 2-D representation techniques have been used extensively in chemical information systems, these being the systematic nomenclature, fragmentation codes, line notations and connection tables. The systematic nomenclatures were primarily used in manual information retrieval systems, due to the lack of flexibility nomenclatures often requiring to be translated automatically to another type of representation if it is to be used in computerized systems [33, 34]. The connection tables have proved to be the most flexible and generally useful representation due to which it forms the basis of most present day computer libraries and systems.

2.1.1. Fragmentation Codes

Fragmentation codes were the first to be used as structural representation in chemical retrieval systems like structure and substructure searching and are still in use. A fragment code is a set of pre defined substructure attributes, the presence or absence of which is used as characterization of a molecule. Figure 2.1 shows the possible fragments in phenylalanine.

Fragments:
-OH
>C=O
-COOH
-NH$_2$
-Ph

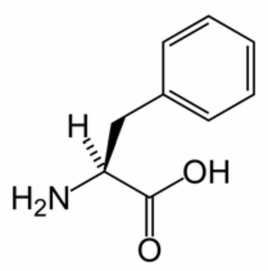


Figure 2.1 Possible Fragments in phenylalanine

A fragment code representation has several problems because of its subjective nature. The system is not standardized and so every organization has their own specific code order for its own database of compounds. For example, a retrieval system for IR spectroscopy data would adopt a very different fragmentation code from a system designed for the retrieval of organo-phosphorous compounds. The coding of new compounds is normally a manual task and sometimes all the molecules need to be recoded if the code is changed because of the addition of a new compound [35]. Fragmentation Codes also results in ambiguous representation for the molecule as the set of codes assigned to a molecule might be interconnected in a different number of ways.

2.1.2. Linear Notations

It is an alphanumeric method to represent the chemical graph of a molecule and is very short as compared to the fragmentation codes, so, well suited for storing and transmittion of molecules. A large number of notational schemes like Wiswesser [36], ROSDAL [37], Sybyl Line Notations (SLN) [38] have been developed before the SMILES notation by Weininger in 1988 [39]. The SMILES notation has obtained a wide spread acceptance because of its easiness to use and comprehend than the Wiswesser Line Notation (WLN) which had been in use for more than three decades since its development in 1954 [36]. SMILES is the only line notation that has the capability to encode the contents of an connection table as opposed to other predecessor line notation schemes. Connection tables are another sophisticated structure representation scheme discussed next. There are a number of enhancements of SMILES (like XSMILES, SMARTS, SMIRKS, STRAPS, CHUCKLES, CHORTLES, CHARTS, etc) [40] designed either to represent special molecular structures or to allow specific applications like database handling, reactions, or polymers. Figure 2.2 shows ROSDAL and SMILES notations for the phenylalanine.

(a) 1-2-3-4=5-6=7-8=9-4, 1=10O,1-11O, 2-12N
(b) 1-2-3-4-=9-4, 1-11O, 1=10O, 2-12N
(c) C1=CC=C(C=C1)CC(C(=O)O)N

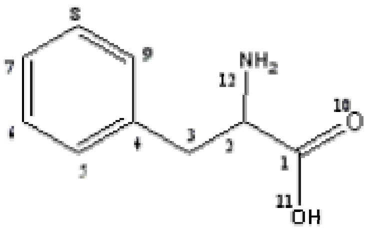


Figure 2.2 A possible ROSDAL and SMILES code for phenylalanine a) a complete ROSDAL, b) a Compressed ROSDAL and c) a SMILES notation.

2.1.3. Connection Tables

These days' connection tables are the primary means of representation for the chemical structures in both public and in house chemical information systems. A connection table contains all the necessary information required for plotting its 2-D graph by any program like SMILES or ChemDraw. The simplest connection table consists of at least two sections: first, a list of the atomic numbers of the atoms; second a list of the pair wise atomic bonds in a molecule. More sophisticated Tables contain the bonding angles information for plotting the bonds. Since connection tables contain an explicit representation of the inter-connections between the atoms in a molecule, they are particularly well suited to manipulations involving such topological information as atom by atom searching, graphical structure input and display, Structure property correlation and reaction indexing. There are a large number of standardized file formats based on connection tables such as the MDL's MOL and SDF formats. For a more detailed understanding of the connection tables and its usage as structure representation medium, the MDL's CTfile formats specifications [41] can be consulted.

2.2. Molecular Descriptors

Molecular Descriptors are numerical values, generated from a mathematical formula, capable of structural properties generation, necessary for the manipulation and analysis of chemical structural information. These numerical values normally characterize most of the physiochemical and biological properties of the chemicals to a large extent. These descriptors can be divided into almost four types 0-dimensional like the atom counts or bond counts in a molecule, 1-dimensional like fragment counts, 2-dimensional like topological indices, 3-dimensional like the 3-D Morse descriptors, and 4-dimensional like the descriptors containing 3-D descriptors along with conformation information. The following paragraph gives some detail about the 2-D and 3-D descriptors which are popular descriptors in clustering, classification and property prediction.

2.2.1. 2-Dimensional Descriptors

There are a big number of descriptors based on the 2-D graph of a molecule. The simplest 2-D descriptors are based on simple counts like number of hydrogen bond donors, the number of hydrogen bond acceptors, the number of ring systems (such as aromatic rings) and the complex 2-D descriptors like 2-D fingerprints, topological indices, kappa Indices, information indices are mainly based on complex mathematical equations. The simple 2-D descriptors are rarely used in chemical systems because of their insufficient discriminating power. The following descriptors based on 2-D graph are popular in chemical information processing:

2-D Fingerprints

A fingerprint can be defined as a string of bits zero or one, describing the presence or absence of some pre-defined substructure or some other features of a molecule. So, the fingerprints can be divided into two types, one based on predefined dictionaries of structural keys and the other one based on some hashing function of the molecule. In case of structural keys, there is always a pre-defined dictionary of molecular substructures and each substructure in a dictionary corresponds to a particular bit in the bit string. Basically 2-D fingerprints have been developed for structural and sub structural searching and for the development of molecular databases and, latter on, have been used in similarity and cluster analysis. So, the aim is to define a structural key when constructing a dictionary that gives the optimal performance in a typical search. In the beginning, the dictionaries were mostly subjective to the type of databases. For example, the dictionary developed for a database of organic molecules might not be useful for a database of solid state materials. But later on there had been much research for finding the set of substructure (fragments), that are most effective in the development of a generalized dictionary [42, 43] 22], and automated methods for screen selection have been developed [44]. The fragments that have higher discriminating power among molecules are selected for a good dictionary and these are neither the most frequently occurring fragments nor the less frequently occurring fragments, rather these are the equifrequent fragments. The most frequently occurring fragments are discarded and the less frequently occurring fragments are combined together so that their usability is increased.

There are three popular fingerprints dictionary systems namely the CAS ONLINE Screen Dictionary for substructure searching [45], Barnard Chemical Information system [28] and MDL MACCS key system[46].

In case of hashed fingerprints, the patterns of a molecule are generated directly from the molecules within a database with the help of some algorithm. The fingerprinting algorithm examines the molecule, generates a pattern for each atom, a pattern for each atom and its nearest neighbor atoms and bonds, and patterns for groups of atoms and bonds connected by various number of paths such as 2 bond paths, 3 bond paths, etc. The number of patterns thus obtained is very large in number and so a fixed length fingerprint can not be assigned. Rather each pattern is used as seed to a pseudo random number generator. The pseudo number generator outputs a number of bits typically 4 or 5 bits per pattern which are latter on added to the fingerprint in a sequence. This method is adapted in the Daylight’s fingerprints. The hashed fingerprint has far more number of patterns for each molecule than a structural key based fingerprint and so is more suitable for large datasets and libraries.

Topological Indices (TIs)

Topological indices are a set of features that characterize the arrangement and composition of the vertices, edges and their interconnections in a molecular bonding topology. These indices are calculated from the matrix information of the molecular structure using some mathematical formula. There are hundreds of molecular descriptors based on the mathematical characterization of molecular structures that play an important role in structure-property and structure-activity relationship, particularly when multivariate regression analysis, artificial neural networks and pattern recognition are used for modeling and analysis. Their advantage over the traditional molecular descriptors, or the descriptors derived from quantum chemical approaches is that they are easily available and can be quickly computed for existing and virtual structures. Among hundreds of possible descriptors, a few have been found useful in characterization of molecular properties and activities, such as the Weiner's index [47-49], Harary index [50], MTI index [51, 52], Balaban index [53, 54], and Zagreb group of indices [55, 56]. Hue Yuan and Chanzhang Cao [57] have recently developed new topological indices based on the vertex, ring and distance of a molecular graph. They are the vertex degree distance index (VDI), odd even index (OEI), ring degree distance index (RDI) and edge distance index (EDI). Although topological indices are being used successfully for more than half a century, still their interpretation is argued. So, Randic (Randic index) and Zupan had recently started work on the structural interpretation of topological indices [58]. There are a number of works predicting structure activity and structure property relationships utilizing the modeling capabilities of TIs. For instance, TI has been used to predict the heats of formation for 60 hydrocarbons and the result show satisfactory predictions [59]. In [60] a QSAR study was carried out for modeling the DNA modeling affinity with the help of distance matrix based TI.

A few of the topological indices are given below as examples:

- **The Wiener Index**. This is half the sum of the bond-by-bond path lengths between each pair of atoms. It can be calculated from each off-diagonal element of the distance matrix, D_{IJ} of a structure with the following formula.

$$\frac{1}{2}\sum_{I,J} D_{I,J} \qquad \forall I, J = 1 \rightarrow n,$$

where *n* is the number of atoms in the molecule.

- **The Balaban Index**. This is the average distance sum connectivity index. It can be calculated as below.

$$\frac{B}{C+1}\sum_{I,J}\frac{1}{\sqrt{D_I \times D_J}} \quad \forall I, J = 1 \rightarrow n,$$

where *n* is the number of atoms in the molecule, D_I is the sum of elements of the *I*th row of the distance matrix, *I* and *J* are neighboring non-hydrogen atoms, *B* is the number of bonds and *C* is the number of rings.

- **Schultz Molecular Topological Index**. It is an important topological index for its interesting applications in chemistry and very high discriminating power for benzenoid graphs [61].

As already mentioned, there are a large number of available topological indices but the problem is that most of them suffer from the degeneracy problem. The first topological index, i.e. the Weiner index is considered to be the most degenerate but this trend among the TIs is decreasing with time and new TIs are added to the list, which are more discriminative.

2.2.2. 3-Dimensional Descriptors

Of course the 3-Dimensional picture of a molecule is closer to its real shape as compared to its 2-dimensional picture and since molecules are 3-dimensional, they can be better represented by descriptors obtained from its 3-D picture. But, the problem is that molecules normally have more than one conformation due to various energy barriers. A chemical conformation is the spatial arrangement of atoms in a molecule. Molecules in which atoms are linked together in the same way, but in which their spatial arrangement are different, are called conformational isomers or simply conformers. These conformers can interconvert by rotation around single bonds, without breaking the chemical bonds. Some of the conformers are more stable and some are unstable. So, the computation of 3-D descriptors are more computational exhaustive and some researchers use the average 3-D descriptors by computing the descriptor for each possible conformer and then averaging it. Two important 3-D descriptors are the 3-D screens and pharmacophore keys.

3-D Screens

3-D fragment screens like 2-D sreens were also initially designed for 3-D substructure searching, but they can also be used as descriptors as we use 2-D screens in similarity measurement and clustering of chemical compounds, and library design. 3-D screens are also strings of bits but encode the spatial distances, angles etc. between different features of a molecule such as atoms, rings, centroids and planes. The ranges of these distances and angles between a pair of features are divided into a number of bins by fixing a constant bin width. Corresponding to each bin of a distance or angle measure, a bit is specified in the bit sring. If the distance or angle calculated for a pair of feature falls in a particular bin , the corresponding bit is set to 1, but intially all the bits in the screen are zeros. For example a distance range of 0-20A° between say two nitrogen atoms can be divided into 10 bins of 2A° , each covering ranges of 0-2A° , 2A°-4A° and so on. Instead of distance between pairs of atoms, in distance based descriptor calculations triplets and quatets of atoms can also be used.

Potential Pharmacophore Point Descriptors

The pharmacophore fingerprints give a common frame of reference for comparing different ligands and for comparing ligands to protein structures using the complementary potential pharmacophores.

3-D pharmacophore fingerprints, consisting of multiple potential 2-, 3- and 4- point pharmacophores can be calculated systematically and with conformational flexibility for structures using software such as the ChemDiverse module of Chem-X [62]. For ligands, the six pharmacophoric features (hydrogen bond donors, hydrogen bond acceptors, acidic centers, basic centers, hydrophobic regions and aromatic ring centroids) are automatically assigned to atoms or dummy atom centroids, whereas for a protein site, complementary site-points with associated pharmacophoric features are first generated and the fingerprint generated from these. A significant increase in the amount of shape information and resolution was found using 4-point pharmacophores, including the ability to distinguish chirality, a fundamental requirement for many ligand-receptor interactions.

Matter et al [63] has compared the performance of 3-D pharmacophore with 2-d fingerprints for diverse compound selection and have found that pharmacophore triplets performed better when the pharmacophores were derived from lower number of conformers but even then they are not comparable to a 2-D fingerprint based design. In [15] Brown et al have used pharmacophore pair and triplets for the clustering of a number of diverse datasets. They have used a number of clustering methods such as Wards, Group Average etc, but their results do not show any superiority of fingerprints based on 3D pharmacophore.

2.3. Classical Clustering Methods

Clustering is a data analysis technique that , when applied to a heterogeneous set of data items, produces homogeneous subgroups as defined by a given model or measure of dis (similarity) or distance. It is an unsupervised process, i.e. there is no pre defined groupings, the clustering job is to find these undefind and unknown clusters. In supervised learning method, there are some known cluster (groups), from which the algorithms learn the underlying relationship among the inputs and their corresponding outputs and so in this way of learning when the model is developed then it is used for the prediction of target groups for new data elements whose groups are unknown. But in case of unsupervised scheme, there is no input output relation in the beginning, only from the input data the groups are predicted. So, clustering can be thought of as an exploratory data analysis technique, that can be used for the selection of diverse compound subsets and data reduction.

Clustering as a methodology for the partitioning of various types of datasets have been in use in almost all fields of social and technical sciences. Everitt [64] has described a number of clustering works in the fields of psychiatry, medecine, social services, education, archialogy, astronomy , market research etc. There are some other good texts available on traditional clustering methods like Numerical texonomy by Sneath and Sokal [65], and classification by Gordon [66]. In case of chemical information a good text is similarity and clustering in chemical information systems by Peter Willet [13]. Later on in the last 15 years, Peter Willet and his group, Barnard, Brown, and Downs have contributed a lot more to the clustering of chemical datasets.

The clustering process for chemical structures is outlined by Brown and Martin [15] as follows:

(1) Select a set of attributes on which to base the comparison of the structures. These may be structural features and/or physicochemical properties.

(2) Characterize every structure in the dataset in terms of the attributes selected in step one.

(3) Calculate a coefficient of similarity, dissimilarity, or distance between every pair of structures in the dataset, based on their attributes.

(4) Use a clustering method to group together similar structures based on the coefficients calculated in step 3.

(5) Analyze the resultant clusters or classification hierarchy to determine which of the possible sets of clusters should be chosen.

There are a number of methods that can be used for the selection of appropriate attributes or descriptors such as the principal component analysis [67] from a set of available descriptors. Although most of the clustering methods uses Euclidean distance as a measure for similarity or dissimilarity among the objects and the centers, in chemical information systems, the Tonimoto coefficient is more useful especially in similarity searching and hierarchical clustering [68, 69].

The classical clustering methods can be divided into two main classes, heirarchical and non-hierarchical clustering methods. A clustering system will be termed as heirarchical if at every iteration a cluster is divided into two clusters or two clusters are merged together to form one relatively homogeneous cluster(s). We find a parent/ child tree like relationship between the clusters at each successive level. Normally these parent/ child relationship in successive levels is visualized with the help of dendograms. In non heirarchical methods, a single partition is divided into more than two clusters and in each iterative step the compounds can go from one cluster to an other cluster untill the compounds in each cluster are more similar than those in other clusters.

Hierarchic methods are, by far, the most popular types of clustering procedure for the clustering of chemical datasets. The methods were developed primarily for applications in life sciences, where the heirarchic clustering could be compared with traditional biological taxonomies in which specimens are grouped into species, and species are grouped into genra [13].

The heirarchical approach can be further divided into two approaches namely the agglomerative and Divisive. In agglomerative methods, the tree is build from the bottom up; first by merging the individual compounds into clusters and then into superclusters by merging more similar clusters based on some dis(similarity) function. On the other hand, the divisive algorithms follows a top down binary tree. Each dataset is divided into two clusters and then successively each cluster is divided into two clusters again and again untill most feasible clusters are obtained. There are a large set of agglomerative methods like single linkage, complete linkage, Group Average, Centroid, median and wards method.

The divisive methods can further be divided into two more subgroups, monothetic and polythetic. The divisive methods where similarity or dissimilarity is computed among various compounds on the basis of only one attribute are called monothetic divisive methods. Similarly, divisive methods based on more than one attribute are called polythetic divisive methods. According to Clifford and Stepheson [70], monothetic divisive methods have several advantages over agglomerative procedures. First, the definition of groups is simple and unambiguous because of the presence or absence of a single attribute. Second, divisive methods should in theory be superior due to the initial split in the data set on the basis of entire data set where the maximum amount of information about attribute frequencies and co-occurrences are available. In contrast agglomerative methods start with individual pair of objects, and so the final clustering may be dependent upon the characteristics of individual data members. In monothetic divisive methods, the selection of proper attribute has significant effects on the final clusters. The attribute selected should be such that it maximizes the dissimilarity among the final clusters and maximizes the similarity among the members of a cluster [13]. But it is also the limitation of monothetic methods that the whole procedure is based on only one variable. It is possible that an object can be more similar to objects belonging to a different group than to members of its own class based on another attribute [66]. There are a number of monothetic divisive methods like association analysis method, Crawford - Wishart method, Information Analysis Method, and Error Sum Squares method. The list of polythetic divisive methods composes of McNaughton-Smith method, Roux method and Minimum Diameter method.

In most of the non-hierarchical methods normally a criterion function is used to measure the goodness of clustering process. Then all the possible partitions are evaluated to see which of them best satisfies the criterion. The problem is that complete enumeration is not possible for all the clusters possibilities. For example, there are about 193 million possibilities to partition 19 objects into only 3 clusters [71]. In order to resolve this problem a number of optimal heuristics has been described on the basis of which a good partition is obtained with much lower computational cost. One of the common approaches is the error sum of squares as the basic clustering criterion; the aim being is to partition the set of N objects into C clusters so as to minimize the total within cluster sum of squares of the distances about the C cluster centroids [13].

Here some of the traditional clustering methods are discussed briefly, which have been in use for the clustering of chemical structures.

2.3.1. Single Linkage Algorithm

The single linkage clustering method is the simplest of all hierarchical agglomerative methods, also known as nearest neighbor technique first described by Florek et al [72]. The defining feature of the method is the distance between two clusters defined as the distance between the closest data elements of the two clusters and so the rest of the data elements of the clusters has nothing in the calculation of the inter cluster separation.

2.3.2. Complete Linkage algorithm

The complete linkage clustering methods is also called the furthest neighbor clustering method. It is a hierarchical agglomerative method where the distance between two clusters to be merged is calculated using the distance between the two farthest data elements of the two clusters. It is an exact opposite strategy to that of the single linkage clustering method.

2.3.3. Group Average Algorithm

The group average clustering method is an agglomerative hierarchical method that merges two clusters in a hierarchy if the distance between the two clusters is the minimum among all the clusters. This distance is an average of all the distances of the elements of one cluster to the elements of the other cluster in a pair of clusters.

2.3.4. Centroid Clustering Algorithm

In the centroid clustering method, each cluster in a hierarchy is represented by its mean vector and so the distance between two pair cluster for merging is determined from two mean vectors of the two pair clusters. A drawback of the method is that if the sizes of the two clusters to be merged together are very different then the centroid resulting from the merging, so it will be too close to that of the larger size group and so the characteristics of the smaller group are virtually lost.

2.3.5. Median Clustering Algorithm

The median clustering algorithm was developed to remove the limitations of centroid method. In this method, the mean vector is formed in such away as the median point of a triangle formed by three points, the centroids of the two pair clusters and the centroid of a cluster or group formed by merging the two pairs intended for merging.

2.3.6. Ward's Clustering Algorithms

The Ward's clustering method was suggested by Ward in 1963 which is based on the minimization of the information loss associated with the merging of two groups in a hierarchy. According to Ward's those pairs of cluster should be merged which result in the minimum amount of loss of information. Ward defined the information loss in terms of an error sum of squares criterion given as:

$$ESS \quad = \quad \sum_{i=1}^{n}(x - x')^2$$

Wards method has proved to be an extremely powerful grouping mechanism, and is considered the best of hierarchic methods. However, it is criticized for its circular cluster shapes. A generalized algorithm for all the hierarchical clustering algorithms is given figure 2.2

```
For I=1 to N-1Do
For J=I+1 to N Do
Calculte the distance between cluster I and
cluster J.

Search the distance pairs to identify the closest pairs
of clusters.
Merge the closest pair and set N=N-1
And REPEAT the Algorithm.
```

Figure 2.2: A general Hierarchical Algorithm

2.3.7. Single Pass Algorithm

Single Pass clustering algorithm is the simplest among the non hierarchical algorithms. The algorithm is based on a user defined threshold. If the distance of a data element from a cluster centre is less than the defined threshold the data element is assigned to the corresponding cluster. The algorithm is outlined in figure 2.3:

```
C=0
For I=1 to N  Do
        Compare I^th objectd with the C clusters to identify the closest distance
        If distance < Threshold Then
                Object added to the corresponding cluster
                Centroids are recomputed
        Else
                C=C+1
                Object becomes the C^th cluster centroid
        End
End
```

Figure 2.3: A Single Pass relocation Algorithm

Where C represents the number of clusters and N is the total number of objects.

This algorithm is quite fast but is suffered from being totally dependent on the order in which the objects are processed.

2.3.8. Jarvis Patrick's Algorithm

Jarvis Patrick's clustering algorithm [17] is the best nearest neighbor based non hierarchical algorithm. This method is exclusively used for the clustering of chemical compounds. The method has two stages. In the first stage N-1 lists of the top K nearest neighbors for each of the N objects is generated. The nearest neighbors are determined on the basis of Euclidean distance, but in chemical datasets, the Tonimoto similarity is also common. The typical value for K is 16 or 20 [11].

The second stage scans the lists of nearest neighbors to create clusters from the objects that follow the following conditions:

- Object *I* is in the top *K* nearest neighbors list of object *j*.
- Object *j* is in the top *K* nearest neighbors list of object *i.*
- *I,*and *j* have at least *Kmin* of their top *K* nearest neighbors in common, where *Kmin* user defined has a range from 1 to *K*.

According to Downs and Barnard, the commonality among the nearest neighbors is used in Jarvis Patrick like methods as criterion for cluster formation. Since it is more compounds centered so, it is well taken by the researchers and industries working in the area of chemical databases and other relevant fields of chemo informatics.

2.3.9. K-means Algorithm

The K-means algorithm is relocation based non hierarchical clustering algorithm. It minimizes the sum of the squared Euclidean distances between all the members' objects of a cluster. The basic steps of the algorithm are shown in figure 2.4.

- *Select K seed objects randomly to act as initial cluster centroids*
- *Assign each object to its nearest cluster*
- *Recalculate the cluster centroids*
- *Repeat the previous two steps until the change in the centroids is negligible*

Figure 2.4: Basic Steps of K-means Algorithm

K-means is considered very efficient as it needs only O(Nmk) time to cluster a dataset of N objects into k clusters in m iterations. Since m and k are usually very small as compared to number of objects, so the time complexity of the algorithm is almost O(N).

2.4. Classification of Chemical Compounds

Classification is another approach to the analysis of multivariate databases. The techniques used range from straightforward statistical classification methods, such as nearest neighbor and linear discriminant classifiers to more sophisticated methods, such as decision trees, neural networks and support vector machines.

The classification approach is a supervised one where some knowledge about the underline datasets such as the number of classes and some examples of each class. Classification systems need to be trained prior to its application to an unknown dataset for mining the desired classes for which the classifier is trained. For example, if a number of Angiotensine Converting Enzyme (ACE) Inhibitors are available, a classifier can be trained by them and then the classifier can be used to extract other ACE inhibitors present in an unknown dataset. The performance of the classifier highly depends on the training set.

The classification of drug like compounds in general into their activity groups using computational methods such as neural networks can make the early filtering and screening process of drug design faster and less costly [73]. Godden et al [74] have used a median partitioning based method to classify a small number of compounds containing very diverse set of activities like enzyme inhibitor, receptor agonist and antagonist and synthetic and naturally occurring molecules. In [75] support vector machines and a two layer neural network trained with back propagation and some other learning methods were tested for the prediction of drug- non drug compounds from a pool of around 10,000 compounds of which about half were drugs and half non-drugs, collected from various databases. They also analyzed the performance using various types of descriptors. Their study shows that the performance of the SVM is slightly better than neural networks but could not justify the conclusion that SVM outperform neural networks. In another study a Kohonen based neural network was used to study the classification of substrate and inhibitors of P-glycoprotein [76].

Chapter 3

Neural Networks and Compound Clustering

With the advent of neural networks almost every field of science and technology have undergone revolutionary changes. Neural networks had been applied to various problems of control, system identification, image processing, pattern recognition, and data analysis. The neural network theory has been motivated by the learning abilities of the human brain which is a very complex system capable of thinking, remembering, and problem solving. There have been many attempts to emulate the human brain functions with computer models, and although there have been some rather spectacular achievements coming from these efforts, all of the models developed to date pale into oblivion when compared the complex functioning of the brain. Thus a neural network is a data processing system consisting of a large number of simple, highly interconnected processing elements (artificial neurons) in an architecture inspired by the structure of the cerebral cortex of the brain [77].

The neural networks come in two flavors, the supervised and unsupervised. It is the unsupervised learning neural networks that can be used for clustering purposes where no information is available a priori about the groups or clusters in the dataset.

When no external teacher or critic's instruction is available, only input vectors can be used for learning. Such an approach is learning without supervision or, or what is commonly referred to as unsupervised learning. An unsupervised learning system (or agent) evolves to extract features or regularities in the presented patterns, without being told what outputs or classes are associated with the input patterns are desired [78]. In other words, the learning system detects or categorizes persistent features without any feedback from the environment.

Classification is the process where the system first learns from the available a priori knowledge about the underlying dataset or process that need to be learned and then identify similar knowledge from the unknown source.

In the last one or two decade neural networks had been employed for clustering and classification problems. The ANNS have a number of important features [5] like:

1 ANNs process numerical vectors and so require patterns to be represented using quantitative features only.

2 ANNs are inherently parallel and distributed processing architectures.

3 ANNs may learn their interconnection weights adaptively. More specifically, they can act as pattern formalizers and feature selectors by appropriate selection of weights.

The neural networks used in clustering problems are also called the competitive learning networks as in such networks the output class information of a given data sample is not required. In competitive learning, similar patterns are grouped by the network and represented by a single unit (neuron). This grouping is done automatically based on data correlations. Well-known examples of ANNs used for clustering include Kohonen's learning vector quantization (LVQ) and self-organizing map (SOM) [79], and adaptive resonance theory models [80]. The architectures of these ANNs are simple: they are single- layered. Patterns are presented at the input and are associated with the output nodes. The weights between the input nodes and the output nodes are iteratively changed (this is called learning) until a termination criterion is satisfied. Competitive learning has been found to exist in biological neural networks. However, the learning or weight update procedures are quite similar to those in some classical clustering approaches like hierarchical and k-means methods.

The ANNs used for classification can have more than two layers of neurons and most of the time are fully connected. The large number of connections enables them to learn any kind of linear or non linear processes. These networks are based on procedural learning, where the output for a given input pattern is compared with desired output and based on the error obtained the weights of each and every neuron and layer are updated in successive iteration until the error becomes negligibly small.

In this chapter the neural networks methods used in this work for classification and clustering of chemical structural databases are discussed. The experimental design, results and discussion are also given.

3.1. Unsupervised Neural Networks

3.1.1. Kohonen Neural Network

The Kohonen self organizing neural network is an unsupervised, competitive learning network, where the neurons participate in a competition among themselves for learning [81]. The output layer neuron forms a 2D or 3D map that describes the groupings in the dataset. This network basically implements a non-linear projections from a high dimensional space to a low dimensional feature map that shows a shadow of the hidden classes in the dataset on a two dimensional mirror. This type of network imposes a neighborhood constraint on the output units, such that a certain topological property in the input data can be preserved in the weights of the output neurons. During the training the weights of the winning neuron as well as the neurons in its neighborhood are updated and are brought closer to the input patterns.

The learning procedure of Kohonen feature maps is similar to that of competitive learning networks. That is, a similarity (or dissimilarity) measure is selected and the winning unit is considered to be the one with largest (or smallest) activation. However, for Kohonen feature map, we not only update the weights of the winning neuron but also all the weights of all the neurons in the neighborhood of the winning neuron. The training process for Kohnen self-organizing map can be described in the following two steps:

Step 1: When an input Z_k is presented to the input layer, the output winning neuron is selected as the one with largest similarity measure (or smallest dissimilarity measure) between all weight vector W_i and the input. If Euclidean distance is used as dissimilarity measure then the distance of the winning neuron will be the smallest from the input pattern. This is shown in equation 3.1.

$$\|Z_k - W_c\| = \min_i \|Z_k - W_i\| \qquad 3.1$$

Where the index c represents the winning neuron.

Step 2: If N_c denotes the set of neurons in the neighborhood of winner neuron, then the weights of the winner and its neighboring neurons are updated as follows:

$$W_i(t+1) = W_i(t) + \eta(t)(Z_k(t) - W_i(t)) \quad i \in N_c \qquad 3.2$$

The neurons which are loser and of course not member of the set N_c are not updated. Here η is the learning rate parameter and varies with time. The neighborhood set N_c can be defined by a mathematical function depending on the shape of the clusters in the dataset. But, in most of the real world problem the bell shaped Gaussian function is the choice.

$$\Omega_c(i) = \exp\left(\frac{-\|p_i - p_c\|^2}{2\sigma^2}\right) \qquad 3.3$$

where p_i and p_c are the positions of the output neurons i and c, respectively, and σ reflects the scope of the neighborhood. By using the neighborhood function the update formula becomes:

$$W_i(t+1) = W_i(t) + \eta\Omega_c(i)(Z_k(t) - W_i(t)) \quad i \in N_c \qquad 3.4$$

Initially the value of learning parameter η and size of neighborhood Ω are kept large and as the learning progresses, these are gradually decreased.

3.1.2. Neural Gas Network

The neural gas algorithm is an important neural network, first introduced by Martinez [82] for the prediction of time series and then applied successfully to the clustering of various databases [83], vector quantization [84], pattern recognition [85, 86], and topology representation [87] etc.

According to Martinez [82] the neural gas algorithm has a number of advantages like, 1- converges quickly to low distortion errors, 2- reaches a distortion error lower than that resulting from k-means clustering and maximum entropy clustering (for practically feasible number of iterations), and from Kohonen feature map and 3- at the same time obeys a gradient descent on an energy surface (like the maximum entropy clustering, in contrast to Kohonen's feature map).

The neural gas algorithm generates a list of the ranks of weight vectors W_k corresponding to each input pattern Z_k which gives the weights a descending order based on the closeness to the input pattern with W_{k0} being the closest weight vector to the input pattern, W_{k1} as the second close weight vector and W_{ki} **,** i = 1,2, …, c-1 being the weight vector for which there are i vectors W_j with $\|Z_k - W_j\| < \|Z_k - W_{ki}\|$. If the ranking index number k associated with each vector W_k is denoted by $R(Z_k, W_{ki})$, which depends on Z_k and the whole set W_{ki} = ($W_{\mathbf{k0}}$, W_{k1}, …, $W_{\mathbf{kc-1}}$) of weight vectors, then the adaptation step we employ for updating the W_k 's is given by

$$W_{ki}(t+1) = W_{ki}(t) - \eta(t)h_\lambda(R(Z_k, W_{ki}))\left[Z_k(t) - W_{ki}(t)\right] \qquad \mathbf{3.5}$$

The learning rate parameter $\eta(t) \in [0,1]$ describes the overall extent of modification and usually is taken as an exponentially decreasing function of time

$$\eta(t) = \eta_i(\eta_f / \eta_i)^{t/Maxiteration} \qquad 3.6$$

Where η_i and η_f are the initial and final values of the learning rate, respectively, which are initialized in advance. The *Maxiteration* is also a constant specifies the number of maximum *t* steps, also initialized at the start of the algorithm.

As already stated the ranking index *R(*Z_k*,* W_{ki}*)* which depend on the input pattern Z_k and the whole set of the weight vectors $W_{ki} = (W_{k0}, W_{k1}, ..., W_{kc-1})$ of weight vectors which also serve as the prototypes of the dataset *Z*. The value of *R(*Z_k*,* W_{ki}*)* is zero for the closest weight vector and is the maximum for the farthest weight vector. The ranking adaptation parameter $h_\lambda(t)$ is a function of the ranking index *R* and lies between 0 and 1. Martinez [82] has suggested an exponential decreasing function

$$h_\lambda(R) = \quad \exp(-R/\lambda) \qquad 3.7$$

the value of $h_\lambda(t)$ thus depends on the rank of the weight vector, as the rank increases the the updation rate decreases and vice versa. For $\lambda=0$ the update formula of equation 3.5 becomes that of the simple competitive neural network where only the winner is updated and the rest of the weights remain unchanged.

According to Martinez [82], NG algorithm is closely related to the framework of fuzzy clustering methods [88]. All the fuzzy methods like fuzzy c-mean and Gustafson and Kessel algorithm [89] uses the fuzzy membership functions $\mu_{ij}, \quad 2 \le i \le c, \quad 1 \le j \le n$ which allows the data elements to partially belong to more than one cluster. Instead of fuzzy membership value, the NG algorithm utilizes the uncertainty of belongingness value $h_\lambda(R)/C(\lambda)$ to assign each input Z_k to all of the prototype vectors W_{ki}, i=1,2, …, c.

Like other partitioning algorithms of clustering, the neural gas algorithm also derives from a criterion function, the minimization of which by a stochastic gradient descent method results in the above updating equation 3.5 which gives the absolute minimum value for the cost function. The cost function for NG algorithm is as follows:

$$E_{NG} = \quad 1/2C(\lambda) \quad \sum_{i=1}^{c}\sum_{k=1}^{n} h_\lambda(R(Z_k, W_i))\|Z_k - W_i\|^2 \qquad 3.8$$

with $C(\lambda) = \quad \sum_{i=1}^{c} h_\lambda(R(Z_k, W_i))$

In order to obtain good results that minimize the energy function E_{NG} to its absolute minimum, the value of λ in the beginning is kept large and continuously decreased until the absolute minimum is found or the other conditions of the algorithm like number of maximum iterations are met.

3.1.3 Enhanced Neural Gas (ENG) Algorithm

The neural gas (NG) algorithm as described in the previous section has a number of advantages like faster convergence to low distortion errors, lower distortion errors than k-means, maximum entropy and kohonen's self organizing maps yet the updating formula is highly fragile in an environment having noise and large number of outliers and is also sensitive to the order of input vectors [90].

$$\Delta W_i = \quad \eta(t)h_\lambda(R(Z_k,W_i)).\|Z_k - W_i\|.\frac{(Z_k - W_i)}{\|Z_k - W_i\|}, \quad i = 1,2,...,c \qquad 3.9$$

It is obvious from neural gas updating formula (equation 3.9), if an outlier Z_0 is presented to update all the prototypes, the amplitude $\|Z_0 - W_i\|$ generated along the unit direction $\frac{(Z_k - W_i)}{\|Z_k - W_i\|}$ will be considerably large such that the prototypes will be dragged towards the outliers. Moreover, if the outliers are highly scattered around the dataset, the training process will not be smooth and there will be lot of oscillation. To overcome these problems Qin et al [60] suggested the following rule which is called the enhanced neural gas:

$$\Delta W_i = \quad \eta(t)h_\lambda(R(Z_k,W_i)).\exp(-\frac{\|Z_k - W_i\|}{\beta d_i^m(0)}).\sigma_i(iter).\frac{(Z_k - W_i)}{\|Z_k - W_i\|}, \quad i = 1,2,...,c \qquad 3.1$$

0

Compared with equation 3.9, this formula also does obey the stochastic gradient descent rule and heuristically adjust the amplitude $\sigma_i(iter)$ of the gradient descent. The gradient descent amplitude $\sigma_i(iter)$ is given as:

$$\sigma_i(iter) = \quad \sigma_i^m(t) = \begin{cases} d^m{}_i(t) & if \quad \|Z^m{}_t - W_i^{iter}\| \geq d^m{}_i(t-1) \\ \|Z^m{}_t - W_i^{iter}\| & if \quad \|Z^m{}_t - W_i^{iter}\| < d^m{}_i(t-1) \end{cases} \qquad 3.11$$

with

$$d_i^m(t) = \begin{cases} \left\{1/2\left[1/d^m{}_i(t-1) + 1/\|Z^m{}_t - W_i^{iter}\|\right]\right\}^{-1} & if \quad \|Z^m{}_t - W_i^{iter}\| \geq d^m{}_i(t-1) \\ 1/2\left[d^m{}_i(t-1) + \|Z^m{}_t - W_i^{iter}\|\right] & if \quad \|Z^m{}_t - W_i^{iter}\| < d^m{}_i(t-1) \end{cases} \qquad 3.1$$

2

and

$$d_i^m(0) = \left[\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\|Z_j - W_i^{mN}\|}\right]^{-1} \qquad 3.13$$

where N, Z_t^m, and W_i^{iter} represent the number of input vectors, the input vector given at iteration t of the training epoch m and the prototype vector i at the total iteration step $iter$, respectively. The term $d^m{}_i(t)$ is the restricting distance for the prototype W_i , which includes both historical and current distance information and is used here to limit the large absolute distance due to the outliers. This has been implemented in equations 3.11-3.13. If the absolute distance of an input vector at any time is greater or equal to the historical distance in the previous iteration, they are averaged using the harmonic mean and if absolute distance is less in magnitude than the historical distance, the averaging is done using arithmetic mean.

The objective of the training process of any neural network is to decrease gradually the average absolute distance. Whenever, this absolute distance becomes larger than the historical mean distance, the input to the machine must be an outlier. So, this is the scheme, used to detect the outliers and decrease their influence on the clustering process in general.

Unfortunately, our results do not show the importance of this technique, as the performance of ENG is not better than the simple neural gas for both types of descriptors, the topological indices and the BCI bitstring. These results are shown in figures 3.2 and 3.3.

3.2. Supervised Neural Networks

3.2.1. Multi Layer Perceptron(MLP)

The multilayer perceptron is a static feed forward neural network that can have virtually any number of hidden layers besides the essential input and output layers of neurons. Practically one or two hidden layers are enough to model complex systems [77, 91]. Usually the error backpropagation [92] method is the preferred learning method to train the network. In such learning the error yielded at the output neuron is propagated back along the layers of the network and the weights are corrected. The output is compared with the desired output of the sample presented at the input. The error $E_i(t)$ for the output neuron i and a given input sample is given as:

$$E_i(w_o, t) = \left\| d_i(t) - y_i(t) \right\| \qquad 3.14$$

where d_i is the desired output and y_i is the observed output of neuron i at time instant t of the training process. The observed output y_i of a simple three layer network for the output neuron i can be given as:

$$y_i = g\left(\sum_{j=1}^{n} w_{oj} . h_i\left(\sum_{k=1}^{m} w_{hk} z_k + w_{h0}\right) + w_{o0}\right) \qquad 3.15$$

where *g* and *h* are the activation functions of the output and hidden layer neurons respectively. The exponent *n* and *m* represents the total number of neurons in the hidden and input layers respectively, and w_h*'s*, w_o*'s* are the weights of the hidden and output layers, and z_k is the input example.

Using the theory of gradient descent learning, each weight in the network is updated by correcting the present value of the weight with a term proportional to the error at the weight, given as

$$w_{oj}(t+1) = w_{oj}(t) + \eta\delta_{oi}(t)z^{-}{}_{k}(t) \qquad 3.16$$

where η is the learning rate parameter whose value is between 0 and 1. δ_{oi} is the value of local error propagated from the output error $E_i(t)$ given as follows.

$$\delta_{oi} = -2E_i \frac{\partial g_i}{\partial z^{-}}$$

Similarly the weight update and the local propagated error for the hidden layer can be given as:

$$w_{hl}(t+1) = w_{hl}(t) + \eta\delta_{hj}(t)z_k(t) \qquad 3.17$$

$$\delta_{hj} = -\sum_{i=1}^{p} \delta_{oi} w_{oij} \frac{\partial h_j}{\partial z_k}$$

The back propagation is a gradient descent minimization procedure used to minimize the cost functional of the feed forward neural network which is a function of the weights of the network and these weights are changing with time. So, the backpropagation learning algorithm tries to find minimum point on the surface formed by the weights of the network. Since all the error computations are based on the local information of the dataset and network, it is always likely that the learning process may trap in local minima. In order to avoid local minima, a momentum term can be used

$$w_{kj}(t+1) = w_{kj}(t) + \eta\delta_j(t)z_k(t) + \mu\Delta w_{kj}(t) \qquad 3.18$$

where μ is the momentum constant which can have values between 0 and 1, Δw is the change in weight in iteration *t* and *t-1*.

The search for the parameters η and μ is a trial and error problem. In [93, 94] a method based on the fuzzy inference system is used to change these parameters adaptively as the learning process progress. This method enables the learning process to avoid the local minima as well as results in faster convergence.

3.2.2. Radial Basis Function (RBF) Network

The radial basis function network is a three layer feed forward fully connected network, which uses radial basis functions as the only nonlinearity in the hidden layer neurons. The output layer has no nonlinearity. Only the connections of the output layer are weighted whilst the connections from the input to the hidden layer are not weighted [77, 95]. The activation function of the hidden layer can be expressed as:

$$h_i(z_k, m_i, \sigma_i) = \exp\left\lfloor -\left\| z_k - m_i \right\|^2 / \sigma_i \right\rfloor \qquad 3.19$$

Where z_k , m_i, and σ_i are the input training sample, centre of the i^{th} Gaussian, and width of the i^{th} Gaussian respectively. These functions are called the radial basis functions and the final output is the sum of the connection's weight times these functions.

$$y_j = \sum_{i=1}^{H} w_{ji} h_i \qquad 3.20$$

The training process is similar to the one for back propagation network, where a cost function like 3.14 is iteratively minimized. The cost function is a function of the weights in the output layer, the centroids and widths of the radial basis functions. The learning process is not implemented as single procedure, but rather three step procedures are adapted. First the centroids of the radial basis functions are determined using a clustering method like K-means, second the receptive width σ_i are determined using heuristic p-nearest neighbors method and last the weights of the final layer are determined simply by a linear least square regression [95].

3.3. Other Machine Learning Methods

3.3.1. Support Vector Machines

Support vector machines have recently found considerable attention in classification problems due to its generalization capabilities. These classifiers maximize the distance (margin) between the training examples and the decision boundaries by mapping the training examples to higher dimensional space [96, 97]. The dimension of the new space is considerably larger than that of the original data space. Then the algorithm finds the hyperplane in the new space having the largest margin of separation between the classes of the training data using an optimization technique known as the risk minimization. For a binary classification problem where there are only two classes in the training data $y_i = \{-1, 1\}$, a hyperplane can be defined as:

$$W.x + b = 0 \qquad 3.21$$

where *W* is the normal to the hyperplane and $|b|/\|W\|$ is the shortest distance of the plane from the origin.

For a good classification model the positive and negative examples of the training data should fulfill the following two conditions:

$$\begin{array}{ll} W.x_i + b \geq +1, & \text{for } y_i = +1 \\ W.x_i + b \leq -1, & \text{for } y_i = -1 \end{array} \qquad 3.22$$

These inequalities can be combined into one set of inequalities

$$y_i(W.x_i + b) \geq 1, \qquad \text{for } \forall \text{i} \qquad 3.23$$

The SVM finds an optimal hyperplane responsible for the largest separation of the two classes by solving the following optimization problem subject to the condition in 3.23

$$Min_{w,b} \ \tfrac{1}{2} \ W^T W \qquad 3.24$$

The quadratic optimization problem of 3.23 and 3.24 can be solved using a langrangian function

$$L_p(w,b,\alpha) = \tfrac{1}{2} W^T.W - \sum_{i=1}^{m} \alpha_i (y_i (W.x_i + b) - 1) \qquad 3.25$$

where *α_i* are the constants known as langrange multipliers. The solution of equation 3.25 for α_i determines the parameters *w* and *b* of the optimal hyperplane. We thus obtain a decision function for the binary classification as:

$$f(x) = \text{sgn}(\sum_{i=1}^{m} y_i \alpha_i \langle x, x_i \rangle + b) \qquad 3.26$$

In any classification task only a few langrangian multipliers α_i tend to be greater than zero and the corresponding training vectors are the closest to the optimal hyperplane and are called the support vectors. In nonlinear SVM, the training samples are mapped to a higher dimensional space with the help of a kernel function $K(x_i ,x_j)$ instead of the inner product $<x_i , x_j >$. Some of the famous kernel functions are the polynomial kernels, radial basis function kernels, and sigmoid kernels[97].

3.3.2. Rough Set Classifier

The rough set theory can be used for knowledge discovery from unknown databases and information system development. If U is a universe of discourse and A is an attribute then the information system is defined as A=(U, A). Both the sets A and U are non empty and finite. Some times the information system in rough set theory is also known as decision table. There are two types of attributes in a decision table, one is the condition attribute and the other one is the decision attribute. The condition attribute is the input attribute and the decision attribute is the target output attribute.

An object that contain similar information is called indiscernible and so two objects can be member of the same equivalence class if and only if they can not be discerned from one an other on the basis of same attribute subset. This indiscernability relation *IND(B)* can be given as:

$$IND(B) = \{(x, y) \in \bigcup \mid \forall a \in B, a(x) = a(y)\} \qquad 3.27$$

Reduct is defined as minimal selection of the important attributes that enables the same classification of elements of the universe as the whole set of attributes. It is used to remove from the table all attributes that are repeated or overlapped. In [98], a genetic algorithm implementation has been used for reduction. The algorithm has support for both cost information and approximate solution.

Other important concept in rough set theory is decision rules. Decision rules are often presented as implications and are called "*if..., then*" rules. It was generated from a set of reducts. Reduction is important in order to obtain minimal decision rules. Classification is done using the generated set of rules and a specified classifier such as *Standard Voting* and *Naïve Bayes* [99].

3.4. Experiments and Results

In this section three different experiments are described, where each experiment varies in terms of the dataset, the set of descriptors and the methods used. The first experiment describes the clustering process and the results obtained where the unsupervised neural methods have been used. The second and third experiments are about the classification of bio-active enzyme inhibitors and AIDS activity.

3.4.1. Experiment 1

Data and Descriptors:

The MDL's drug data report (MDDR) is a collection of compounds which covers the patent literature, journals, meetings and congress on chemical compounds. The database currently contain over 132,000 biologically relevant compounds and well defined derivatives, with update addition of around 10,000 compounds each year [29]. This database can be searched online through MDL's DiscoveryGate software. The database provide the researchers with various options based on structure , biological activity, molecular derivative, or based on a combination of a number of properties.

Initially, around 3000 compounds belonging to various biologically active groups from the MDL's MDDR database were selected for this experiment. But it has been observed that many of the compounds are redundant in the database and so need to be removed. Furthermore, for some of the compound structures our descriptor generation software was unable to generate the desired descriptor. So, all those compounds having redundancy or for those the descriptor could not been generated were removed and we were left with 1388 compounds belonging to 7 biologically active groups which can further be divided into 15 biological groups and are shown in Table 3.1.

Table 3.1: Shows the groupings and some characteristics of the Dataset

S.No	Activity	No. molecules
1	Interacting on 5HT receptor Potentially useful in the treatment of depression, anxiety, hypertension, eating disorders, obesity, drug abuse, cluster headache, migraine, obsessive compulsive, and associated vascular disorders, panic attacks, agoraphobia eating, urinary incontinence and impotence.	
	5HT Antagonists	48
	5HT1 agonists	66
	5HT1C agonists	57
	5HT1D agonists	100
2	Antidepressants Potentially useful as an antiepileptic, antiparkinsonian, neuroprotective, antidepressant, antispastic and/or hypnotic agent. Some of the compounds may be useful in the treatment of dopamine-related CNS disorders such as Parkinson's disease and schizophrenia.	
	Mao A inhibitors	84
	Mao B inhibitors	174
3	Antiparkinsonians Potentially useful in the treatment of septic shock, congestive heart failure and hypertension and in the prevention of acute renal failure.	
	Dopamine (D1) agonists	32
	Dopamine (D2) agonists	104
4	Antiallergic/antiasthmatic Most of these are used as antiinflammatory, antiasthmatic and antiischemic agents. However, adenosine (A3) antagonists are useful as a tool for the pharmacological characterization of the human A3 receptor.	
	Adenosine A3 antagonists	73
	Leukotine B4 antagonists	150
5	Agents for Heart Failure Potentially useful as a bronchodilator, smooth muscle relaxant or cardiotonic agent, accelerator of hormone secretion, platelet aggregation inhibitor, etc.	
	Phosphodiesterase inhibitors	100
6	AntiArrythmics Most of the Potassium channel blockers block the cardiac ion channel carrying the rapid component of the delayed rectifier potassium current.	
	Potassium channel blockers	100
	Calcium channel blockers	100
7	Antihypertensives	
	ACE inhibitors	100
	Adrenergic (alpha 2) blockers	100
	Total molecules	1388

In this work two chemical descriptors generation softwares have been use.

(1) Barnard Chemical Inc (BCI) [28] fingerprint generation software, which can generate four types of fingerprints for any compound structure based on BCI dictionaries bci512, bci1024, bci1052, and bci4096; and

(2) Dragon software from MelanoChemometrics [100] with the help of which more than 1500 descriptors can be generated that include upon topological indices, charge indices, 3-D Morse descriptors and many more.

We have generated around 100 topological indices using the Dragon software, out of which only 10 have been selected, accounting for around 98% of the variance in the dataset. The 10 topological indices include upon, the Narumi geometric topological index (GNar), the total structure connectivity index (Xt), Pogliani index (Dz), Schulz molecular topological index (MTI) , path/walk 3 - Randic shape index (PW3), path/walk 4 - Randic shape index (PW4), path/walk 5 - Randic shape index (PW5), 2D Petitjean shape index (PJ2), eccentric connectivity index (CSI), and distance/detour ring index of order 3 (D/Dr3).

The other scheme, we have used for descriptor generation is the BCI bitstring, where we generated bit strings for our dataset using the BCI Makebits utility. Since, all bitstrings are based on a dictionary of molecular substructures; here we have used the BCI1052 dictionary. The size of the bitstring is very high i.e. 1052 bits per molecule, so, we need to reduce the dimension of the feature set. But the problem is that this whole large feature set of 1052 features is only accountable for around 75% variance in the dataset due to which it is very difficult to reduce the size any further.

Methodology:

In this work three neural networks methods namely the Kohonen, neural gas and enhanced gas have been used for the clustering chemical structures and the performance have been compared with the industry standard group average and Ward's methods. Each of the method was repeated for 10, 20… 100 sets of clusters. Then the active cluster subset method [15] has been used to determine the proportion of active compounds from the inactive compounds within an active cluster subset. The active cluster is defined as the superset of all the clusters having at least one active compound. The singletons are counted towards the active cluster subset as it will wrongly show high proportion of actives. Moreover, the objective of the clustering is to combine the active compounds with the active compounds into same group and inactive compounds with inactive. Since, our dataset contain more than two activities, so for each activity the rest of the activities are taken as inactive and this procedure is repeated for each activity group and then an average is taken.

Results and Discussion:

Since, the variance for the BCI generated descriptors is not very good and the size of the bit string is very large, so, first we tried a unique idea to decrease the size of the bits string by combining them into bins of bits and then converting the bits in each bin into decimal numbers. In order to arrive at a good bin length various bin lengths have been tried. We divided the whole set of features into groups of 4, 8, 10, 12 and 16 bits. Each group of bits was converted to decimal and we get a reduced size feature vector of sizes 264, 132, 105, 88 and 66 corresponding to 4, 8, 10, 12 and 16 bits groups respectively. The best clustering results were obtained from the groups of 10 bits which is shown for Wards Clustering in figure 3.1. Here the x-axis plots the no. of clusters and the y-axis shows the corresponding percentage of actives in the active cluster subset.

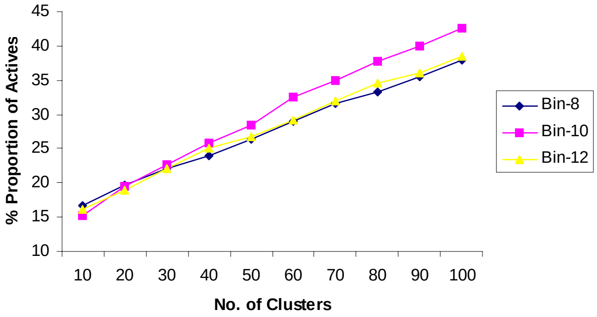


Figure 3.1: Proportion of actives in active cluster subset (BCI)

So, for the rest of the methods the bin method with bin length of 10 bits was used. It should be noted that all 105 variables accounted only for 71% of the variance in the data.

The results of all the three neural network methods are shown in figure 3.2. The Wards, GpAv, SOM, NG, and ENG stands for the Wards, Group Average, Self organizing Kohonen Maps, Neural Gas, and enhanced neural gas respectively. The results of the neural methods are very poor as compared to the other two methods.

The results of the clustering with the neural network methods based on topological indices as shown in figure 3.3, gives some hope of the utilization of neural network methods in chemical structure analysis. As already described the topological descriptors are more suitable for clustering because of their continuous and high discriminative characteristics. The Kohonen SOM method with gaussian neighborhood and exponential learning rate selection gives the best results among the rest of the neural networks. The results of the enhanced neural gas network are very poor for both types of descriptors. The algorithm keeps a history of the average distance of each cluster formed. When a new compound comes for clustering its distance is computed from each cluster centroids. If it is equal to or greater than the historical distance, it is assumed to be an outlier and so the corresponding weight update is suppressed by using the harmonic distance instead of the current Euclidean distance. In other words, the algorithm tries to decrease the number of prototypes so that the influence of the outliers is minimized, which results in many of the clusters as empty clusters.

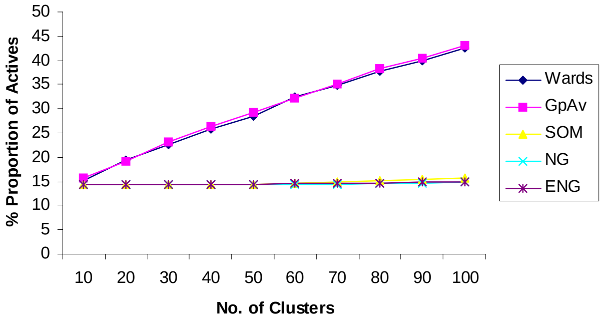


Figure 3.2: Proportion of actives in active cluster subset (BCI)

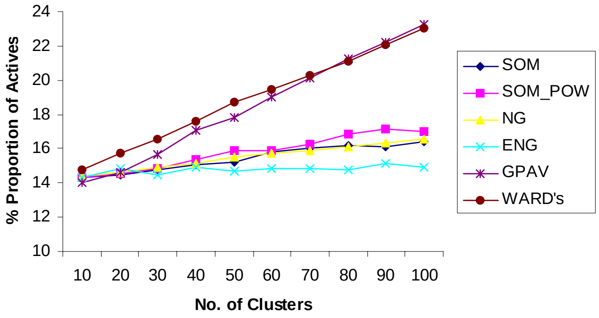


Figure 3.3: Proportion of actives in active cluster subset (TI)

3.4.2. Experiment 2

In this experiment three types of computational intelligence methods have been used for the classification of three types of bioactive enzyme inhibitors. The objective of the work is to determine the effectiveness of the important classification schemes for the analysis of datasets comprising multi class compounds. As outlined in chapter 2, most of the classification works so far reported deals with binary classification problems such as drug / non-drug classification.

Data and Descriptors:

The dataset used in this work is composed of three types of enzyme inhibitors, namely ACE, phosphodiesterase enzyme and steroid 5α reductase enzyme inhibitors. The ACE inhibitors are very useful drugs in heart diseases like high blood pressure, heart failure, and also in diabetes for the preservation of the kidney function. ACE enzymes activates a harmone in human bodies called angiotensine which causes blood vessels to constrict and so results in high blood pressure and a strain on the heart [101].The phosphodiesterase inhibitors can be used for blocking one or more of the various subtypes of the enzyme phosphodiesterase. Currently, they are under active research to be used in humans for the treatment of various diseases and some of them have already been tested on human. A current study show that the use of phosphodiesterase III inhibitors in heart failure patients resulted in increased mortality rates [102]. Defects in the steroid 5α-reductase type 2 enzyme activity cause decreased formation of dihydrotestosterone (DHT) from testosterone (T) which increases the T/DHT ratio, resulting in defective masculinization of external genitalia [103]. A number of inhibitors are available to stop this effect.

These compounds were collected from the MDL's MDDR database. Then a number of filters were applied to remove the compounds which are redundant due to their exhibition of multiple activities. The compounds for which the descriptors could not been generated due to some error in their structural data were also eliminated. In the end, the dataset contained 314 ACE inhibitors, 792 phosphosdiesterase (including subtypes I and III) inhibitors, and 244 steroid 5α-reductase (subtype I and II) inhibitors.

The topological descriptors were used in this work. With the help of dragon software around 99 descriptors were generated, out of which 10 variables have been selected using PCA [67] that accounts for 98% of the variance in the dataset.

Methodology:

In this work three types of classifiers were used for multiple target classifications and the number of targets was the three classes, i.e. various types of inhibitors used in this work.

The dataset have been partitioned into two parts a training part which is used for training of the algorithm and a Test part which is used for Testing. The percentage of training and testing portions of the dataset was varied in order to study the variation of performance caused by change in the ratio of Training to Testing partitions of dataset. The training /testing partitions used contained (10%, 90%), (30%, 70%), (50%, 50%), (70%, 30%) and (90%, 10%) of dataset. For the selection of samples in training and testing portions interleaved method was used to make it sure that percentage of each class in each portion is preserved.

Results and Discussion:

In this work two types of neural networks, feed forward neural network and a radial basis function neural networks with one hidden layer were considered. The networks were tested for a variable number of hidden layer neurons. The number of input layer neurons was the same as the number of inputs which correspond to the number of variables in the dataset and in the output layer there were three neurons corresponding to the three target outputs. The output can range between (0, 1). The training samples were presented to the neural networks, the output was compared with the desired output for a given input sample and the errors were back propagated for the update of weight vectors.

First the experiments were carried out for various numbers of neurons in the hidden layer. In both type of neural networks, it has been observed that the performance enhances with the increase in number of neuronal nodes. It has been found that performance starts degradation after reaching a steady state point. Figure 3.4 and Figure 3.5 shows the behavior of MLP and RBF networks for various values of neurons in the hidden layer. In both the cases the best results were obtained when the number of hidden layer nodes was 20.

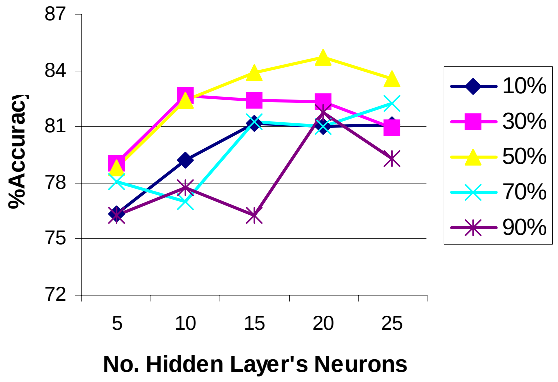


Figure 3.4. Prediction accuracy of MLP neural network. The predictions are shown for variable number of hidden nodes.

The networks have been trained and tested with variable ratio of samples from the dataset. The behavior of both the network is almost similar. As the training data is increased the accuracy of prediction increases, but it is not good when the number of training examples is very large than the testing examples. The MLP network gives the best prediction accuracy when the training/ testing ratio is 50%, whereas the RBF network prediction is the best when it is trained with 10% and tested with 90% of examples in the dataset.

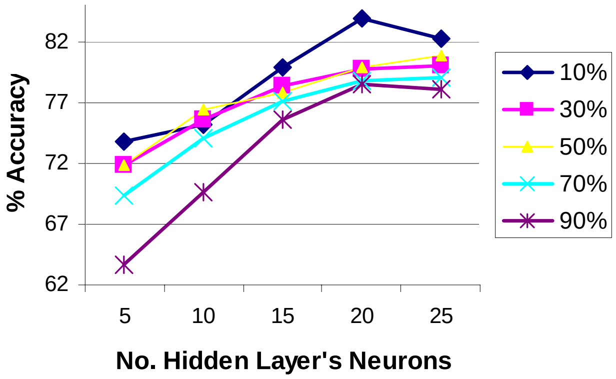


Figure 3.5. Prediction accuracy of RBF neural network. The predictions are shown for variable number of hidden nodes.

For SVM, the polynomial Kernel was used with various degrees, but the best results were obtained for degree 3 and degree 4 as is shown in figure 3.6. As we increase the percentage of

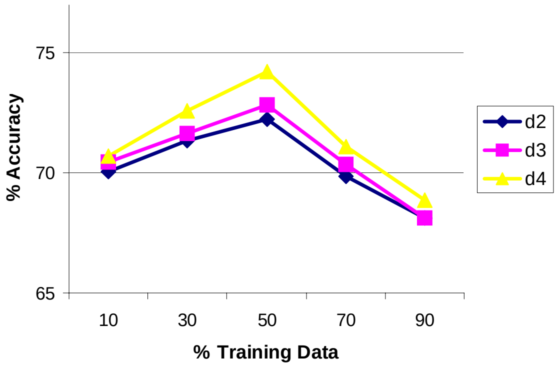


Figure 3.6. Prediction accuracy of SVM. Results are shown for various degrees of the polynomial kernel.

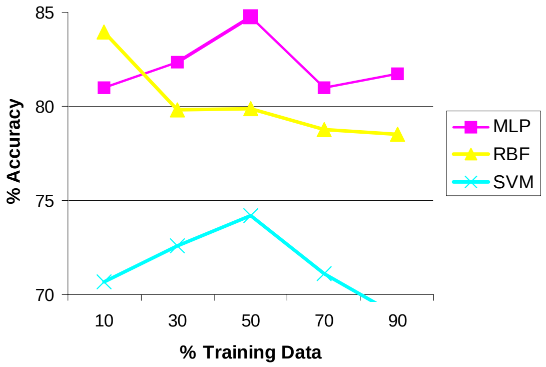


Figure 3.7. Overall Prediction Results for MLP, RBF and SVM. The results shown are the number of neurons are 20 in the hidden layer and degree for the polynomial was 4.

training dataset, the prediction accuracy increases and reaches its highest point at 50% and then decreases. The best average prediction obtained for SVM was 72.83%.

The best results of all the three methods are given in figure 3.7, where the prediction accuracy of neural methods is established over SVM. The MLP gives the highest correct prediction when trained with 50% of the training data and RBF gives its best result for only 10% of the training data and at this point its prediction accuracy is comparable with that of MLP.

The three different classes of enzyme inhibitors have been classified with different accuracy. The ACE inhibitors were classified equally well by all the three methods in comparison with the other two classes. The MLP prediction rate was higher than 90%.

3.4.3 Experiment 3

This experiment compares the classification results of three methods, the multilayer perceptron (MLP), support vector machines (SVM) and a rough set classifier. All the three classifiers have been designed as binary classifier to classify compounds into two classes of active and non active. This experiment uses a subset of the NCI's AIDS database [30].

Data and Descriptors:

The dataset used in this experiment consists of 5772 compounds derived from the NCI AIDS database [30]. These com pounds have been tested against HIV virus and so are divided into three categories, the compounds which shows high activity, the compounds which show moderate activity and the compounds which shows no activity at all. However, here only the compounds showing high activity or no activity at all have been used and the moderately active compounds have been excluded. So, we considered only 880 compounds in this work for our experiments. All structures were characterized by BCI bitstrings standard 1052 fragment dictionary that encodes augmented atoms, atoms sequences, atom pairs, ring components, and ring fusion descriptors.

Results and Discussions:

In order to see the effect of the size of training portion of the dataset, the dataset have been split into two parts training and testing. The ratio of training vs testing partitions were taken as 20:80, 50:50 and 80:20. The comparison was also drawn on the basis of whole 1052 bits and the smaller number of bits selected by PCA.

Table 3.2 Classification Accuracy without Dimension Reduction

Dataset	MLP	RS	Support Vector Machines(SVM)		
			Linear	**Polynomial**	**RBF**
20:80	60.4	42.0	82.3	81.3	82.8
50:50	87.9	42.9	92.0	95.2	97.2
80:20	97.3	44.3	96.5	98.2	98.2

MLP – Multilayer Perceptron Neural Network
RS – Rough Sets Classifier

The classification accuracy of the algorithms was measured as the ratio of the accurately classified no of compounds *n* to the total no of compounds *N* in the testing data. This can be expressed as:

$$A = \frac{n}{N} X100$$

The results show that the dataset with larger training partition gives the best accuracy than the training sets with lesser samples. This result is invariant across all the three methods and other available options. Table 3.2 shows the results of the methods with varying training size for the dataset characterized by all the descriptors (i.e. 1052 bits) and table 3.3 gives the results with the PCA selected descriptors.

The prediction accuracy of neural network improves as we increase the training samples and gives the highest accuracy of 97.3% and 80.5% for the whole descriptor set and PCA selected descriptor set respectively.

Table 3.3 Classification Accuracy with Dimension Reduction

Dataset	BP	RS	Support Vector Machines(SVM)		
			Linear		
20:80	21.5	51.5	75.0	20:80	21.5
50:50	58.9	42.2	84.7	50:50	58.9
80:20	80.5	52.8	91.4	80:20	80.5

MLP – Multilayer Perceptron Neural Network
RS – Rough Sets Classifier

Apart from the size of the training set, the performance of neural network is dependent on a number of other parameters such as the learning rate, momentum rate, number of hidden layer nodes. It has been found that the combination of (0.1, 0.9) and (0.1, 0.1) for the learning rate and momentum rate respectively, gave the best classification accuracy. In this work a single hidden layer based multilayer perceptron was used with the number of hidden layer nodes equal to $\sqrt{mn}$, where m is the number of neurons in the input layer and n is the number of neurons in the output layer. So, the number of hidden layer neurons was 32 for the non PCA descriptors and 26 for the PCA selected descriptors. It should be noted that there were 1052 nodes in the input layer for non PCA descriptors and 686 for the PCA selected descriptors. The output layer had only one neuron.

In theory, the size of hidden layer should enlarge with increase in training data. But if the number of hidden layers is too few, the network will not be able to model the data correctly. Therefore, the right number of hidden layer neurons performs a crucial role in better classification.

The MLP classifier gives the best prediction accuracy of 97.3% without PCA and 80.5% with PCA.

For the support vector machines (SVM) three types kernel i.e. linear, polynomial, and RBF were used. The results obtained with the help of RBF and polynomial kernels were equally good, when all the 1052 descriptors were used. They predicted 98.2% of the test data correctly when trained with the full descriptor set and 80 % of the dataset. When the PCA based reduced descriptor set was used the polynomial kernel show better results than the other two types of kernels. The polynomial kernel classified 91.4% of the test data correctly which was the best prediction among all the three methods using reduced descriptor set.

In general, the prediction accuracy of the rough set classifier was very poor; the maximum prediction accuracy of the algorithm was only 52.8 % when the reduced descriptor set was used. It is very important to note that the rough set classifier performed better on the reduced descriptor set than the full descriptor set. For the full descriptor set the accuracy was only 44.32%. However, these results were for the largest training set containing 80% of the examples of the whole dataset. For the training set containing lesser percentage of examples as training set, the results are worse.

3.5. Summary

In this chapter a number of neural network techniques, supervised and unsupervised, and some other machine learning techniques have been employed to see their effectiveness in analyzing the biologically active chemical datasets. The algorithms considered here were included upon Kohonen neural network, neural gas network and enhanced neural gas network for the clustering and multilayer perceptron network, radial basis function neural network, support vector machines, and rough set based classifier for the classification of chemical compounds into their biologically active groups.

For the purpose of evaluation a number of datasets have been developed from various drug databases such as NCI Aids and MDL’s MDDR. The first dataset contained 1388 molecular structures from 15 distinct biological classes and was used for the evaluation of the clustering methods. The second dataset used in this work comprised of 1350 important inhibitors. They include three distinct classes’ ACE inhibitors, PDE inhibitors and 5-α SR inhibitors. It was used for the evaluation of three classification methods the MLP neural network, RBF neural network and support vector machines to find their suitability for classification of precise and multi-class datasets. The third dataset used in this work was derived from the NCI Aids database and was used for the evaluation of MLP neural network, support vector machines and rough set classifier for binary classification.

Among the unsupervised neural networks, the kohonen self organizing network showed the best results but generally the results of neural networks were poor than that of industry standard hierarchical clustering algorithms Wards and group average. The topological descriptors perform better than the BCI bitstrings in clustering multi-class datasets. The support vector machine with polynomial kernel showed better results than the MLP neural network and rough set classifier for binary classification but they were not as good when used for the classification of multi-class and precise datasets.

Chapter 4

Fuzzy Logic & Compound Clustering

Fuzzy logic based clustering is some times referred to as soft clustering due to its inherent ability of assigning an object to more than one cluster based on the extent of its similarity with the clusters. This approach of clustering is the subject of this chapter, first a few basics about the fuzzy logic are introduced, next a few important fuzzy clustering algorithms are presented and last the results of the fuzzy logic clustering algorithms employed here for the clustering of chemical structures are given and discussed.

4.1 Fuzzy Logic

Fuzzy logic have been devised by Lutfi Ali Zadeh [104] to define mathematically the imprecise but yet correct judgment of human. The theory explains the real life situations mathematically which are imprecise, vague and uncertain. There is a misconception about fuzzy logic that it is itself imprecise or inexact (or approximate) reasoning, which is not correct as the fuzzy logic itself is precise and can describe and handle in a more simple and easy way what is imprecise.

In general terms, conventional logic rests on a black and white model whereas real life contains several shades of grey. Fuzzy logic can be considered a superset of conventional (Boolean) logic that can cope with partial truths, using a continuous range of truth-values between true (1) and false (0). Thus, fuzzy logic is a multivalued logic that incorporates the concept of vagueness in mathematical formulations.

There are many products currently available that have been manufactured using fuzzy logic such as washing machines, microwave ovens and video camcorders. One of the most successful fuzzy based systems is the Sendai Subway Train in Japan, considered the most advanced subway system in the world [105].

4.2. Fuzzy Set Theory

The conventional set theory developed by George Cantor (1845-1918) in the late 19th century was afflicted by a paradox known as Wang's paradox which can be given with the following inductive argument [106]:

0 is small
If a number, *x*, is small, then *x*+1 is small.
Similarly, if *x*+1 is small, so is *x*+1+1.
Therefore every number is small

This is a paradox and so every number is small even the number 5 trillion and infinity. Obviously 5 trillion is not really a small number and nor is infinity, thus breakpoints were implemented to overcome this paradox. For the above example a specific breakpoint will be selected where *x* stops being small, e.g. at 10 (Figure 4.1). Numbers less than 10 are full members of the set of small numbers, numbers greater than 10 are not. However this still does not seem to describe the situation sufficiently because if 10 is small how can 11 not be small with a difference of only 1?

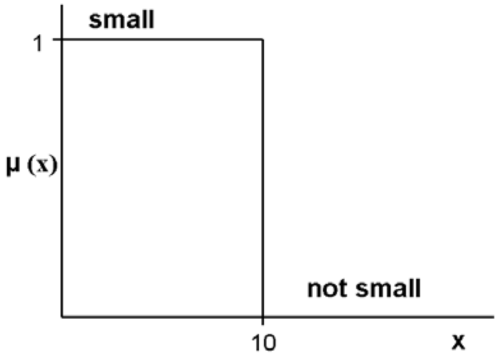


Figure 4.1. Conventional set

The real situation is likely to be a more gradual change, as *x* increases it gradually becomes not small. Fuzzy sets are able to describe this gradual change. Figure 4.2 displays how this situation would be described by fuzzy sets. There are two sets: *small* and *not small*, with fuzzy sets objects can be members of both sets. As a number increases in size its membership of the set of small numbers decreases.

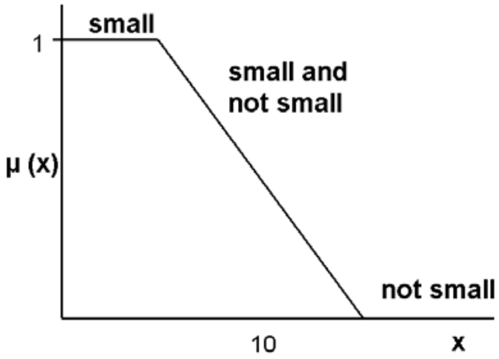


Figure 4.2. Fuzzy sets

The first publications on fuzzy set theory were produced by Lotfi Zadeh in 1965. In the same way that Boolean logic is closely related to conventional set theory, fuzzy logic is strongly linked to fuzzy set theory. Members of each set are only members to a predefined degree, and may also be members of other fuzzy sets. This membership will be described in full in the next section.

Fuzzy sets have been shown to be particularly useful in pattern recognition, this is because objects will often not be clearly a member of one set or another. With conventional sets, objects that do not fit in to any particular set will be assigned to one set leading to imprecision. Fuzzy sets allow the degree to which the object is a member of the set to be specified, thus allowing the classification of ambiguous data.

4.3. Membership Function

The membership function of an object describes to what degree that object is a member of a given set. In traditional set theory an object is either a member of a set, corresponding to a membership function of 1, or not, membership function 0. The Law of the Excluded Middle, defined by Aristotle, states that

"X must be either Z or not Z."

Fuzzy logic extends this notion and allows X to be both Z and not-Z. X will have a membership function for both of these fuzzy sets (Z and not-Z), describing the degree to which X is a member of each. This notion can be seen in Figure 4.2 where the numbers between 8 and 12 would be described as both small and not-small to differing degrees. The integer 8 has a high membership function for the set 'small' and a low membership function for the set 'not-small'.

In conventional set theory given a set *A,* in a space of points, X, with a generic element of X denoted as x, i.e. $X = \{x\}$. The membership function assigns a value $\mu_A(x)$ to each x, i.e.

$$\mu_A(x) = \begin{cases} 1 & if \quad x \in X \\ 0 & if \quad x \notin X \end{cases}$$

or,

$$\mu_A(x) : X \rightarrow$$

The membership is either 1 or 0, depending on whether the element is a member, or not. In fuzzy set theory an object is assigned a value for its membership function of a given set anywhere between these two,

$$\mu_A(x) : X \rightarrow [0, 1]$$

The sum of membership for each object is 1, this is for mathematical tractability. The closer the value $\mu_A(x)$ to unity, the greater the degree with which x is a member of the fuzzy set A. Fuzzy set theory is a generalization of traditional set theory, with a traditional set being the extreme case of a fuzzy set with all objects having membership functions of either 0 or 1. A simple example of a fuzzy set and the notation used is given:

Example 1: A = "real numbers close to 10"

$A = \{ (x, \mu_A(x)) \}$

Where

$\mu_A(x) = \{(1 + (x - 10)^2)^{-1}\}$

This would give

$A = \{(7, 0.1), (8, 0.2), (9, 0.5), (10, 1), (11, 0.5), (12, 0.2), (13, 0.1)\}$

The object is given first, followed by the membership function of the object to the given set. The membership pattern of this fuzzy set forms a symmetrical graph (Figure 4.3). Obviously, the membership function of 10 is 1, because the fuzzy set is defined as being real numbers close to 10.

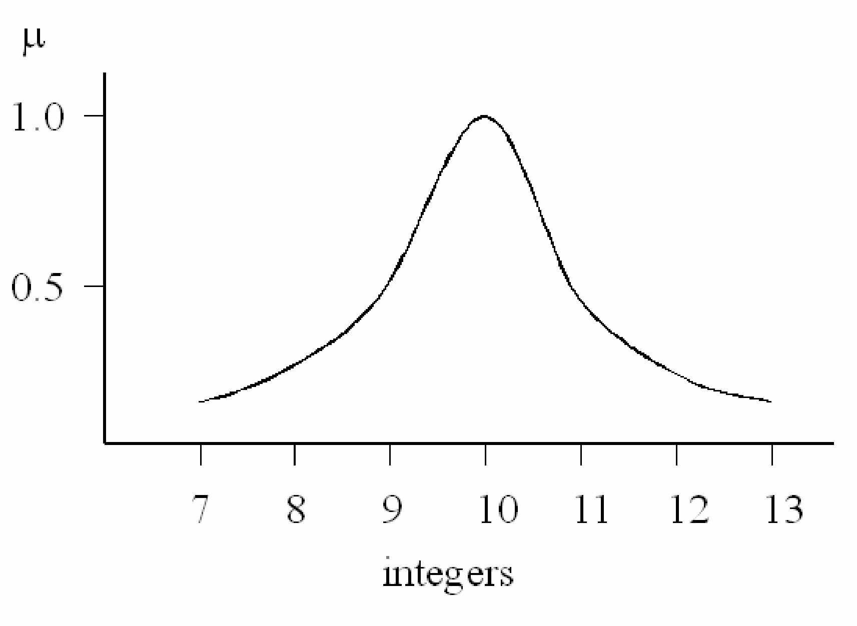


Figure 4.3. Fuzzy set A: real numbers close to 10

Fuzzy sets are functions from some feature space of objects onto [0, 1], the range of the membership function [107]. The membership functions are a fuzzy subset of the feature space. There are an infinite number of fuzzy subsets that can be associated with a set.

Fuzzy set theory provides more information than conventional set theory by assigning degrees of membership to each set for all of the objects. Therefore objects displaying very similar characteristics to the set prototype will be described as such (by having a membership function close to 1). Additionally, the outliers in a particular fuzzy set will be assigned lower membership functions.

The operations on fuzzy sets are an extension of the operations for conventional sets [104], for a good summary refer to [108]. In addition, many algebraic operations can be performed on fuzzy sets but are beyond the scope of this report.

4.4. Fuzzy Clustering

Most traditional cluster analysis methods are crisp (or hard) partitioning, in which every given object is strictly classified into a certain group. So, the boundaries defined for the objects (data elements) are very sharp and so they go to only and only one cluster. However, the features or attributes of the objects in practice are not sharp and they may be having some tendency to be the part of some class to some extent. Fortunately, the fuzzy sets theory proposed by Zadeh [104] provides a powerful tool for such soft partitioning of the data set. Thus, people began to deal with clustering with fuzzy fashion and named them fuzzy cluster analysis. Since fuzzy clustering obtains the degree of uncertainty of samples belonging to each class and expresses the intermediate property of their memberships, it can more objectively reflect the real world. Thereby, it has become the main content of studies on cluster analysis. In the last two decades a large number of variations of fuzzy based clustering methods had evolved starting with the Fuzzy C-means [88] in 1984.

Fuzzy clustering has been shown to be advantageous over crisp clustering in that the total commitment of a vector to a given class is not required in each iteration. Fuzzy methods have shown spectacular ability to detect not only hyper volume clusters, but also clusters which are actually thin shells that is curves and surfaces [109]. We can see a number of examples in the literature [110-112] dealing with shell like volume curves and surfaces.
When compared to their crisp counterparts, fuzzy methods are more successful in avoiding local minima of the cost function and can model situations where clusters actually overlap.

The variants of FCM are themselves emerged from the FCM, as researchers have worked to improve its performance. If some researchers have worked to decrease the time consumptions others have worked to improve its accuracy. This method has been applied to various data types especially in the area of image segmentation with slight variations.

4.4.1. Hard and Soft Clustering

In order to derive the objective function and other relevant mathematics for fuzzy c-means and its other variants, it is better to see the same for the hard (crisp) partitioning technique, so that we may be able to understand the difference between the two approaches. If we look into these issues all of them appears to be objective functional minimization problems. If the constrains are relaxed we get the possibilistic partition scheme. So, the clustering algorithm is nothing but a minimization problem which may be constrained or unconstrained.

These kind of methods are based on classical set theory and defines the presence or absence of a data point in a partition subset on strict logic, that is the object either belong to a subset or not. So, such kind of methods divides a data set strictly into disjoint subsets.

Let us suppose that we have a data set *Z* and the objective is to partition (group or cluster) it into *c* clusters. If we suppose that *c* is known as a priori, then the hard partition of *Z* is a set

$$\{\mathrm{A_i} \mid 1 \le \mathrm{i} \le \mathrm{c}\} \subset \mathrm{P(Z)} \qquad 4.1(a)$$

$$\bigcup_{i=1}^{c} \mathrm{A_i} = \mathrm{Z} \qquad 4.1(b)$$

$$\bigcap_{i=1}^{c} \mathrm{A_i} = \phi \quad \text{for } \mathrm{i} \neq \mathrm{j} \qquad 4.1(c)$$

$$\phi \subset \mathrm{A_i} \subset \mathrm{Z} \quad 1 \le \mathrm{i} \le \mathrm{c} \qquad 4.1(d)$$

$$\mathrm{U} = [\mu_{\mathrm{ik}}]_{cXn} \quad \begin{matrix} 1 \le \mathrm{i} \le \mathrm{c} \\ 1 \le \mathrm{k} \le \mathrm{n} \end{matrix} \qquad 4.1(e)$$

$$\mu_{ik} \in \{0, 1\} \quad \begin{matrix} 1 \le \mathrm{i} \le \mathrm{c} \\ 1 \le \mathrm{k} \le \mathrm{n} \end{matrix} \qquad 4.1(f)$$

Where A_i is the subset of *Z*, and composed of compounds belong to the cluster *i*, which is also a subset of the *P(Z)*. The union of all the clusters should be equal to the set *Z* and the intersection of any two different clusters should be an empty set (this is true only for hard partitioning) as each compound can belong to only one of the *c* clusters. A set *U* is a matrix of size *cxn* which stores the membership values of each compound in each cluster. An important aspect of the hard partitioning system is that membership μ_{ik} can have only two values either 1 or zero. This aspect is evident from equation 4.1(f).

In contrast to the hard partitioning, the fuzzy partitioning allows the compounds to be part of more than one cluster to some degree. However, the fuzzy partitioning restricts the membership of a compound by incorporating the sum of the memberships to 1. These two aspects of the fuzzy partitioning can be given by the following two equations:

$$\mu_{ik} \in [0,1] \quad \begin{matrix} 1 \le \mathrm{i} \le \mathrm{c} \\ 1 \le \mathrm{k} \le \mathrm{n} \end{matrix} \qquad 4.2(a)$$

$$\sum_{\mathrm{i}=1}^{\mathrm{c}} \mu_{ik} = 1 \qquad 1 \le \mathrm{k} \le \mathrm{n} \qquad 4.2(b)$$

The fuzzy partitioning thus gives the following benefits over the classical partitioning:

1 a realistic representation of data

2 clusters are not very restrictive

3 Allows the compounds to be part of more than one cluster and so clusters are no more disjoint sets of compounds

4 possibility of detection of noise and outliers in the data

These benefits can be observed from a simple example of a space of objects illustrated in figure 4.4. The objects would conventionally be classified into two clusters with the middle point A which is equidistant from both the cluster centres, being clustered to either 1 or 2, or as a singleton. This would not be a realistic picture of the data because A is a mid-point. In addition, whichever cluster object *A* is assigned to, it will be very close to objects from the other cluster.

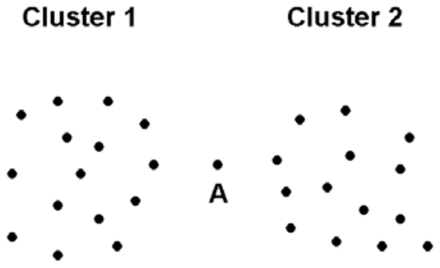


Figure 4.4. An example of overlapping data

This is generally referred to as the ties in proximity problem [113]. Compounds such as *A* in Figure 4.4 that are equidistant from clusters can result in ambiguous cluster assignments. Large data sets, focussed data, binary representations, short fingerprints, specific algorithms such as complete link and measures such as the Euclidean distance can increase the probability of encountering ties and thus produce more ambiguous clusters. The probability of producing ties can thus be reduced by incorporating the good options of fuzzy partition. MacCuish et al. [113] have also suggested treating ties as overlapping points between clusters.

This is the approach taken by fuzzy clustering, clusters overlap to the degree appropriate to cluster members. In a fuzzy cluster representation molecule *A* in Figure 4.4 will be assigned to both clusters 1 and 2, thus accounting for the similarity it shares with both.

However, there exists another approach known as the soft or possibilistic approach where the memberships of the compounds are not probabilistic as imposed by the fuzzy membership principle (equation 4.2b) rather it is possibilistic and depends on the similarity or closeness of a compound to a certain class. The approach was formally introduced by Krishnapuram [114] by allowing the compounds to have any membership between 0 and 1 and the sum of all memberships of a compound is not confined to unity. According to Krishnapuram, the membership should reflect a typicality of the distance of a compound from a class centre. Although the approach is criticized for giving coincident clusters and computational expense by many researches [115-117].

Non-hierarchical clustering methods are sometimes considered not well suited to cluster analysis of data containing background noise [118]. This is because most objects will be assigned to clusters, even when they are outliers as singletons are undesirable. These compounds lie far away from the cluster centre and therefore have a large effect on the cluster centroid calculations. This problem can be overcome to an extent by weighting data points according to the proximity of other data points [118]. Fuzzy clustering takes this idea a step further, as will be explained below.

The rationale behind extending existing clustering algorithms to incorporate fuzzy logic is that fuzzy logic is better able to describe and deal with data that forms non well-separated clusters which are typical of real life situations [119].

Fuzzy set theory has enabled the extension of some of the conventional clustering algorithms, in particular the *k*-means method. The main advantage of fuzzy set representation in cluster analysis is that objects may naturally have characteristics in common with more than one cluster; fuzzy clustering allows these objects to belong to all the clusters with which they have similarity. There are many situations where groups of objects will not form natural exclusive clusters, therefore fuzzy clusters have been proposed.

4.4.2 Fuzzy c-means

The fuzzy c-mean (FCM) is the most important and fundamental algorithm of the fuzzy clustering first proposed by Raspuni [120], improved by Dunn [121] and later generalized by Bezdek [88]. The membership functions are defined based on some distance function such that the membership values reflect the association of a data element to various cluster prototypes. FCM method has the advantage over its counter part crisp CM that it can recognize overlapping clusters in a data set. FCM in essence is a Lyponove function minimization problem which is normally dealt as a subject of variational calculus. The objective of the calculus of variation is to find a solution which leads to the minimization of the functional. A function which has more than one independent variable is often termed as functional or Lyponove function [122].

The functional of the FCM Algorithm has three independent variables, *U* the membership matrix, *Z* the data space and the vector of prototypes *V* and is given as:

$$J(V, U, Z) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^{m}\|Z_k - V_i\|^2 \qquad 4.3$$

where

$$V_i = [v_{i1}, v_{i2}, ..., v_{iN}], \quad v_j \in \Re, \text{ and } j = 1, 2, ..., N$$

$$Z_{kj} = [z_{k1}, z_{k2}, ..., z_{kN}], \quad z_{kj} \in \Re, \text{ and } j = 1, 2, ..., N$$

and μ_{ik} is the same as defined by equation 4.2.

The functional *J* represent a hyperplane whose shape varies with the variations in the three independent variables (which are themselves functions of time) the dataset *Z*, the prototype (or cluster centroid) *V*, and the membership matrix *U*. The fuzzy c-mean algorithm is an iterative minimization procedure for finding the appropriate values for all these variables that result in the minima of the functional. In the minimization procedure since the dataset *Z* can be regarded as a constant function for a particular problem, so, the derivatives of the functional with respect to membership matrix and the prototype matrix will result in the desired steps required for finding the minima.

The steps of the fuzzy c-means algorithm can be summarized as follows:

Step1

Initialize *m*, the partition matrix *U*, the number of clusters *C* and tolerance ε.
Repeat the following steps

Step2

Compute the cluster prototypes

$$V_i^{(l)} \quad = \quad \frac{\sum_{k=1}^{n} (\mu_{ik}^{(l-1)})^m Z_k}{\sum_{k=1}^{n} (\mu_{ik}^{(l-1)})^m} \quad 1 \le i \le c \qquad 4.4(a)$$

Step3

Compute the distances between compound Z_k and cluster centre V_i

$$D^2{}_{ik} = \quad (Z_k - V_i^{(l)})^T (Z_k - V_i^{(l)}) \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \qquad 4.4(b)$$

Step4

Update the partition matrix

If $D_{ik} > 0$

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} (D_{ik} / D_{jk})^{2/(m-1)}} \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \qquad 4.4(c)$$

else

$$\mu_{ik} = \quad 0 \qquad 4.4(d)$$

Step5

$$\text{if} \quad \left\| U^{(l)} - U^{(l-1)} \right\| \quad \le \quad \varepsilon \ \textbf{Stop} \qquad \textbf{4.4(e)}$$

else go to Step 2

The algorithm has two important parameters that the fuzzyfication parameter (index) *m* and the number of clusters *c*. In principle when choose *m = 1*, the fuzzy c-means algorithm becomes a generalization of its hard c-means algorithm if the membership functions are assigned the values o and 1 only. But it can not be tested as a value for *m = 1* will result in a divide by zero error in equation 4.4(c).

As the value of *m* is increased from 1.1 ahead, everything inside the algorithm becomes more and more fuzzy. The data points within the clusters becomes more closer to the centre and results in more compact clusters but at the same time the clusters also becomes closer and closer and results higher overlapping. So, an appropriate value is normally obtained with error in trial for which the algorithm has to run for a number of values to find the most suitable value. Some researchers have suggested a value of 2 [9, 123], but it should be noted that it strongly depends on the dataset and the nature of clusters. The higher values for *m* also decrease influence of outliers and noise.

The number of clusters *c* is also very important and the quality of clustering is highly influenced by it. Although there are available a number of methods for the estimation of number of clusters in a dataset but many a times the algorithm needs to be run for a number of choices and then select the better partition based on some criterion. The initialization of one of the two matrices the membership matrix *U* or the prototype matrix *V* is necessary and different initialization can lead to slightly different results. However, this influence can be minimized by allowing the algorithm to lower least squared errors.

The FCM algorithm is criticized for its high computational cost, due to large number of calculations for the prototype and membership matrix updating in each iteration of the algorithm [124]. The effect gets worsened when accompanied by large number of clusters. As the number of clusters increases size of prototype matrix, distance and membership matrix increases which needs more calculations.

The algorithm is also suffering from the issue of producing same size and shape clusters [114, 125]. The FCM algorithm produces good clusters only when they are well separated, have similar sizes and spherical shapes.

In FCM the memberships of a compound depends on the number clusters and constrained to have a sum of 1. So, the memberships of the compounds are not allowed to have membership more or less than 1, which is not typical of the distance of the compound from the centers. The memberships should show a typicality of their distance from the centers and should not be under probabilistic constraints [114]. For example consider two clusters as is shown in figure 4.5 and give attention to two points A and B are almost at equal distances from the center of cluster1 (left side cluster) and so they should have almost equal membership in cluster1. But since point B have to give a considerable amount of its membership to cluster2, its membership in cluster1 will be smaller than that of point A. So, the memberships are not typical of their distances rather they are giving the degree of sharing. This will cause the algorithm to be sensitive to the presence of noise and outliers in the data. The other approach which tries to decrease the influence of noise and outliers is called the possibilistic fuzzy approach [114]. However, these possibilistic approaches are criticized for their coincident clusters [115] and their higher immunity to noise.

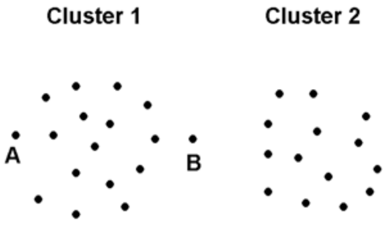


Figure 4.5. A dataset with two clusters

There are a number of solutions have been proposed. In [117] a repulsive term is added to the possibilistic objective functional, so that when the clusters become closer the objective functional increases. It has been shown that the problem of coincident clusters has been effectively removed but with the addition of some parameters that complicates the algorithm. In [126] an improved pssoibilistic approach to clustering have been discussed with an improved objective function based on important partition validity indexes such as partition coefficient (PC) and partition entropy (PE) which gives good results in presence of noise and outliers.

4.4.3 Fuzzy Gustafson- Kessel Algorithm

Gustafson and kessel fuzzy algorithm [89] is an important solution to the problem of same size and similar and spherical shape clusters encountered in fuzzy c-mean type of algorithms. They had brought some improvements into the original FCM by incorporating adaptive distance norm, in order to detect different shape clusters in a data set. The algorithm is capable of detecting ellipsoidal shapes of clusters of dissimilar sizes and orientations to various degrees [114].

This is achieved by allowing each cluster to have its own norm inducing matrix $A_{i,}$ independent from that of other clusters, which results in the following distance norm:

$$D^2{}_{\mathrm{ikAi}} \quad = \quad \left\|Z_k - V_i\right\|_{Ai} \quad = \quad (Z_k - V_i)^T \; A_i \; (Z_k - V_i) \qquad 4.5$$

The matrix A_i must be symmetric and positive definite, otherwise the objective functional can not be minimized with respect to A_i. For this purpose the volume of A_i is restricted to be a constant by $\det(A_i) = \rho$, where the value of ρ is usually selected to be unity but it can vary to allow different sizes of the clusters which will need a priori knowledge about the size of each cluster. So, with the addition of this constraint the objective function with the Lagrange multipliers λ_i will become as given below [9]:

$$J(U, V, Z, \{A_i\}) \; = \; \sum_{k=1}^{n}\sum_{i=1}^{c} \mu^m{}_{ik} \left\|Z_k - V_i\right\|^2 A_i \text{ - } \sum_{i=1}^{c} \lambda_i (\det(A_i) - 1) \qquad 4.6$$

After differentiating the objective functional with respect to λ_i and then putting the result equal to zero, the following relation is obtained.

$$A_i \; = \; \left[\rho_i \; \det(F_i)\right]^{1/p} \; F_i^{-1} \qquad 4.7$$

where F_i is the fuzzy covariance matrix for cluster i given as:

$$F_i = \frac{\sum_{k=1}^{n} (\mu_{ik}{}^{(l-1)})^m (Z_k - V_i^{(l)})^T (Z_k - V_i^{(l)})}{\sum_{k=1}^{n} (\mu_{ik}{}^{(l-1)})^m} \quad 1 \le i \le c \qquad 4.8$$

For, A_i to be symmetric and positive definite, we must have p linearly independent vectors (objects) in the data Z.

The Gustafson Kessel algorithm has the same structure as the FCM, with the only difference of the distance D_{ik} which is calculated with the help of fuzzy covariance matrices as given in equations 4.7 and 4.8.

The GK algorithm has one additional parameter to those of FCM algorithm that need to be initialized in advance. It is the volume parameter ρ whose will principally vary for cluster to cluster in the dataset, but since it is not always possible to know about the volumes of clusters a priori, it is just equated to 1 for all clusters.

The covariance matrix F_i is responsible for the shape and orientation of the clusters and the information regarding the shape and orientation of the clusters can more easily be obtained from its eigenstructure. The length and width of the hyper ellipsoid of the cluster is obtained from the square roots of the eigen values whereas the directions can be obtained from the eigen vectors. It is shown in figure 4.6.

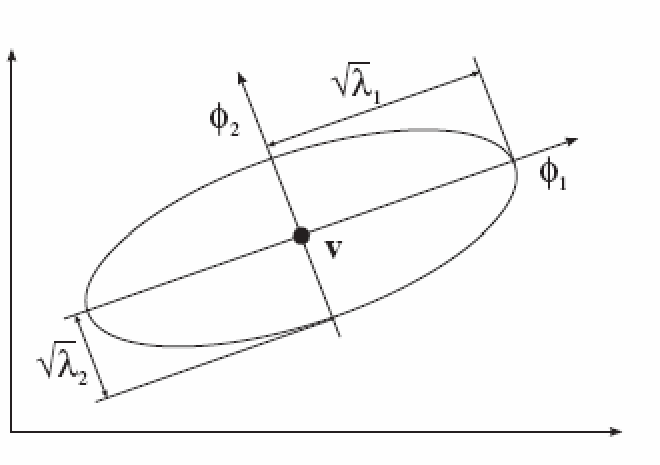


Figure 4.6. The length of the ith axes is given by $\sqrt{\lambda_i}$ and the direction by the vector ϕ_i

4.4.4 Modified Gustafson- Kessel Algorithm

The application of Gustafson Kessel algorithm is not always possible especially in two cases when the number of data samples within the clusters is significantly less and when the data samples within the clusters are highly correlated [127]. In such a situation it is not possible to compute the inverse of the covariance matrix as it become singular and even the volume of the cluster can not be kept fixed as the determinant of the covariance matrix becomes zero.

As shown in figure 4.6, the eigen values and eigen vectors of the covariance matrix gives us information about the shape and orientation of the cluster. When the ratio between the large eigen value and the smaller eigen value is very large (ratio of large to small eigen value is called condition number of the matrix) the matrix is nearly singular and so the norm inducing matrix in equation 4.7 can not be computed because of the singularity of the covariance matrix. The condition number for which the matrix becomes singular varies from dataset to dataset, and so need to be found before using error and trial method.

One solution to this problem of covariance matrix singularity is to constrain the ratio between the large and small eigen values to some predefined threshold. So, when the minimal eigen value become very small it need to be enlarged such that the condition number is less than or equal to this threshold [127]. When the condition number is very large it means the ratio of large to small eigen value is high and from figure 4.6 it can be argued that the cluster will be very close to a linear shape cluster. A linear shape cluster always possesses high correlated data.

We have observed that the above mentioned solution increases the minimal axes of the ellipsoidal and so the shape becomes more and more close to a spheroid which have little relationship with natural distribution of data. This problem specially occurs when the number of data points in the cluster is too few. This problem is tackled by adding a fraction of the identity matrix to the covariance matrix as given in equation 4.9.

$$\mathrm{F}_{\mathrm{i}}^{\mathrm{new}} = (1-\gamma)\,\mathrm{F}_{\mathrm{i}} + \gamma\,\det(\mathrm{F}_0)^{1/p}\,\mathrm{I} \qquad 4.9$$

where F_0 is the covariance matrix of the whole dataset and $\gamma \in [0,1]$ is a tuning parameter. The value of γ should be as small as possible so that the covariance matrix is a close as possible to the natural covariance matrix.

The Gustafson Kessel algorithm in modified form can be given as follows:

Step1

Initialize m, the partition matrix U, the number of clusters C and tolerance ε.

Repeat the following steps

Step2

Compute the cluster prototypes

$$V_i^{(l)} \quad = \quad \frac{\sum_{k=1}^{n}(\mu_{ik}^{(l-1)})^m Z_k}{\sum_{k=1}^{n}(\mu_{ik}^{(l-1)})^m} \quad 1 \le i \le c \qquad 4.10(a)$$

Step3

Compute Covariance Matrix

$$F_i = \frac{\sum_{k=1}^{n}(\mu_{ik}^{(l-1)})^m (Z_k - V_i^{(l)})^T (Z_k - V_i^{(l)})}{\sum_{k=1}^{n}(\mu_{ik}^{(l-1)})^m} \quad 1 \le i \le c \qquad 4.10(a)$$

Step4

Modify the Covariance Matrix

$$\mathrm{F}_\mathrm{i}^{\mathrm{new}} = (1\text{-}\gamma)\,\mathrm{F}_\mathrm{i} + \gamma\,\det(\mathrm{F}_0)^{1/\mathrm{p}}\,\mathrm{I} \qquad \text{(as equation 4.9)} \qquad 4.10(c)$$

Step5

```
If   $\mathrm{cond}(\mathrm{F}_\mathrm{i}) \ge \beta$   then
        Find max Eigenvalue (maxλ) using EigenvalueDecomposition
        For j=1:p
                If maxλ / $\lambda_{jj}$ >= β  then

                        $\lambda_{jj}$ = maxλ / β
                end if
        end for
end if
```

Step6

Reconstruct the covariance matrix F_i from the eigen vector φ_i and newly formed eigen value matrix ψ_i

$$\mathrm{F}_\mathrm{i} = \phi_i\, \psi_i \phi_i^{-1} \qquad 4.10(d)$$

Step7

Compute the norm inducing matrix A_i

$$\mathrm{A}_\mathrm{i} = [\rho_\mathrm{i}\,\det(\mathrm{F}_\mathrm{i})]^{1/\mathrm{p}}\,\mathrm{F}_\mathrm{i}^{-1} \qquad 4.10(e)$$

Step8

Compute the distances between compound Z_k and cluster centre V_i

$$D^2_{ik} = (Z_k - V_i^{(l)})^T A_i (Z_k - V_i^{(l)}) \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \qquad 4.10(f)$$

Step9

Update the partition matrix

If $D_{ik} > 0$

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{c} (D_{ik} / D_{jk})^{2/(m-1)}} \quad \begin{matrix} 1 \le i \le c \\ 1 \le k \le n \end{matrix} \qquad 4.10(g)$$

else

$$\mu_{ik} = 0 \qquad 4.10(h)$$

Step10

if $\left\| U^{(l)} - U^{(l-1)} \right\| \le \varepsilon$ **Stop** 4.10(i)

else go to Step 2

4.5 Experiments and Results

The experiments conducted on the evaluation of fuzzy logic based techniques are given here in this section. Basically different experiments conducted on two different datasets are described. These fuzzy clustering algorithms are quite different than conventional clustering algorithms like Wards and k-means. In the previous sections the difference between hard k-means and fuzzy c-means has been highlighted. The fuzzy based algorithms allow the compounds to be part of more than one cluster.

4.5.1 Experiment 1

This experiment is about the application of fuzzy clustering methods fuzzy c-means and fuzzy Gustafson-Kessel for the clustering of chemical structures. The results are compared with the industry standard Wards and Group Average methods.

Data and Descriptors:

The dataset used in this work consists of 1388 molecules collected from the MDL's MDDR database. The topological indices have been used as descriptors. The details about the dataset and the descriptors used are given in previous section 3.4.1.

Results and Discussion:

As already, we have seen in the previous sections 4.4.3 and 4.4.4 that there are a number of parameters in the fuzzy c-means and fuzzy Gustafson-Kessel algorithms. For example, the fuzzification parameter *m* is of paramount importance, which adjusts the fuzzification of the input data space. The best value for *m* is the one that results in the best partition of the dataset into a number of clusters. First the experiments were carried out to find the best value for *m*. The experiments were repeated for the values 1.1 to 3.9. The default value of *m* being 2.0 which have been used by a number of researches [123].

The analysis of the results was performed using the active cluster subset method which shows the separation of actives from inactives and combination of actives with actives and inactives with inactives. The experiments were repeated for 10, 20,…, 100 clusters with a sampling rate of 10 clusters. The clusters that contain at least one active compound were merged together to form the active cluster subset. The proportion of actives in this active cluster subset was determined. This kind of proportion was determined for each activity in the dataset and an average was computed which is used in all the comparisons.
Figure 4.7. Shows proportion of actives in active cluster subset for various values of *m*. The number of clusters were kept constant at c=20. The best results were obtained for *m=1.2* for both the fuzzy methods. The next best value was found to be 1.7. So, the fuzzification parameter have been selected to be 1.2, and all the rest of computations we used this value. This value can also be confirmed from similar studies on chemical compounds. Although a number of studies such as Feher et al [18] and Lin et al [19] uses the default value of 2.0 but there are numerous other researches that uses the optimal value. For example, a study by Linusson et al that involve the fuzzy cluster analysis of a group of alcohols found the optimal value of *m* to be 1.2 [128].

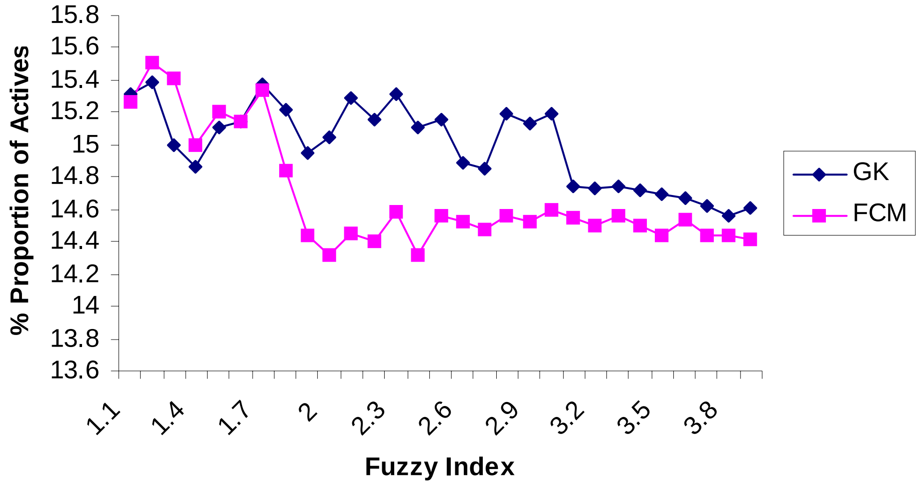


Figure 4.7. Finding the best value for the fuzzification parameter *m*

Similarly in another study that looked at clustering of molecules according to their ecotoxicological properties used the value m = 1.3 [129]. Thus, it appears for chemical applications that where an optimal value is sought a lower fuzziness index is favored, which also approves our results.

Another experiment we have conducted to answer the question whether the results obtained with different initialization of the membership matrix are similar or not. In fuzzy clustering algorithms initially one of the two matrices, the membership matrix or the prototype matrix need to be initialized. We initialized the membership matrix randomly. It has been observed that this random initialization do affects the results. For this purpose, we repeated the experiment 10 times each time the membership matrix was initialized with different memberships randomly and then normalized to 1. Figure 4.8 show the results obtained for 50 clusters and $m = 1.2$.

The results of FCM are more sensitive to the variation in membership matrix random initialization than that of the fuzzy GK algorithm. The experiment has been repeated again for 10 clusters and then the mean of the two experiments were also compared. It has been observed that two mean values were almost similar. For example, the two mean values for FCM were 18.81394 and 18.81515 for 10 repetitions of the experiment with the fuzzification parameter and number of clusters being fixed, which does not show any significant change. The mean values for GK algorithm were even more similar.

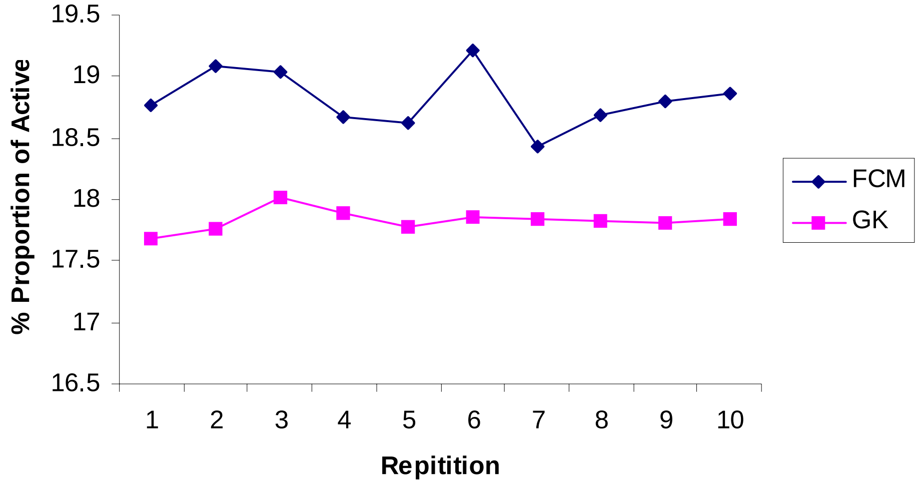


Figure 4.8. FCM and GK results for varying initialization of Membership Matrices

The results have been compared with that of the Wards and Group Average methods. Each experiment for the fuzzy methods was repeated 10 times and then an average performance measure was computed and compared. These results are shown in figure 4.9.

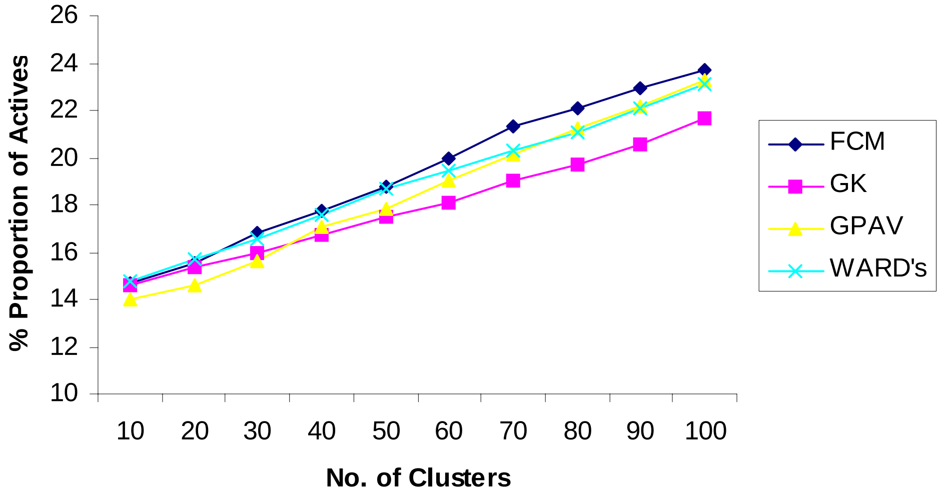


Figure 4.9. Comparison of the performance of FCM and GK Algorithms with that of Wards and Group Average

The results of the FCM and GK algorithms are almost similar to that of Wards and Group Average. The results of FCM is slightly better than the rest of the methods when the number of clusters are increased, where as the performance of GK algorithm shows slight degradation with the increase of number of clusters.

In these results were obtained for the maximum membership of each and every compound. For example if the membership of compound *j* is 0.02 0.5 0.2 0.189 0.1 in five clusters, then the compound *j* will go to cluster 2 of the five clusters as its membership value is the highest.

The convergence of the Gustafson Kessel algorithm is very sensitive to the initialization of membership matrix. Although, the results in figure 4.8 shows very little, but we have observed that for some of the clusters the algorithm did not converged to the minimum error after going through 1000 iterations. In normal cases the algorithm did converge to RMS error as smaller as 0.0001 in less than 400 iterations. It shows that the algorithm is more likely to stick in local minima. This is the cause of its slightly inferior performance than FCM.

4.5.2 Experiment 2

This experiment gives our results for the fuzzy c-means cluster analysis based on the fingerprints dataset collected from the NCI Aids database. The experiment compares the results of the fuzzy c-means initialized randomly; fuzzy c-means initialized with the Wards clusters as seed and the Wards clustering.

Dataset and Descriptors:

The dataset used in this study have already been described in section 3.4.2 and 3.4.3. This dataset comprise only two types of compounds, the compounds that show high activity against the AIDS and the compounds that do not show any activity and so termed as inactive or non-drug.

The descriptors used were the BCI 1052 dictionary based fingerprints that are also discussed in section 3.4.2.

Results and Discussion:

For the AIDS dataset used in this study, first of all experiments were carried out to find the optimal value for the fuzzification parameter *m*. The results show that the best results were obtained when *m = 2.0* as is shown in Figure 4.10

A very interesting result is that as the number of clusters is increased, the proportion of actives in active cluster subset decreases. In other words the increase in number of clusters reduces the combination of actives with actives and inactives with inactives.

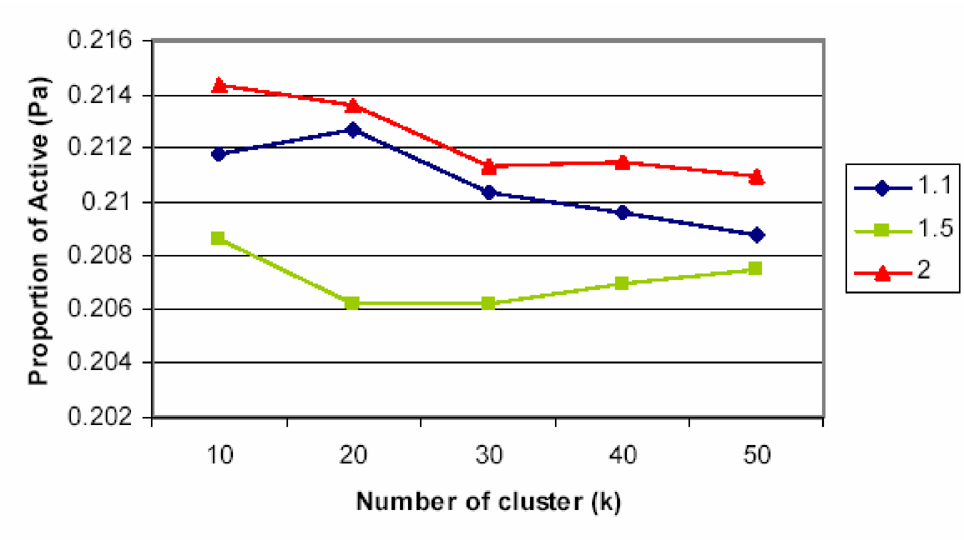


Figure 4.10. In search of the Best fuzzification parameter *m* using the AIDS binary class data

In this work another performance measure have also been utilized. We give this performance measure the name Mean Inter-Molecular Dissimilarity or MIMD for short. The MIMD is defined as given in equation 4.11.

$$\text{MIMD} = \frac{1}{c}\sum_{i=1}^{c}\frac{1}{n}\sum_{j=1}^{n}(1-S_{i,j}) \qquad 4.11$$

Where $S_{i,\,j}$ is the similarity between a cluster centre *i* and any other molecule *j* inside the cluster. There are a large number of similarities coefficients available but here the tonimoto coefficient is preferred as it gives good results when used with bits string descriptors. It gives more weight to the common structures found between atoms [69, 130]. The tonimoto similarity coefficient can be given as:

$$S_{A,B} = \frac{c}{a + b - c} \qquad 4.12$$

where

a	is the number of unique fragments in molecule A
b	is the number of unique fragments in molecule B
c	is the number of common fragments between molecule A and B

Here in calculation of MIMD, one of the molecules is always considered as the central virtual molecule, as in fuzzy clustering the centre of a cluster is necessary to represent a molecule in the dataset.

The clustering is considered good when the value of MIMD is smaller, as the main objective of the clustering being to reduce the dissimilarity within clusters and increase the similarities.

The results obtained using the MIMD performance measure is shown in figure 4.11, for a three value of *m* and for a number of clusters.

The dissimilarity was found to be minimum for *m=2.0* only when the number of clusters were lesser that is 10. As the number of clusters increased the dissimilarity increased for every value of *m*. For large number of clusters the minimum dissimilarity within clusters was obtained for the *m=1.5*. But the dissimilarity for *m=1.5* was also found to be the next closest to the minimum.

After finding the best fuzzification parameter, the rest of the experiments concentrate on the performance of the fuzzy c-means, Wards, and Wards seeded fuzzy c-means algorithms based on the binary class AIDS dataset.

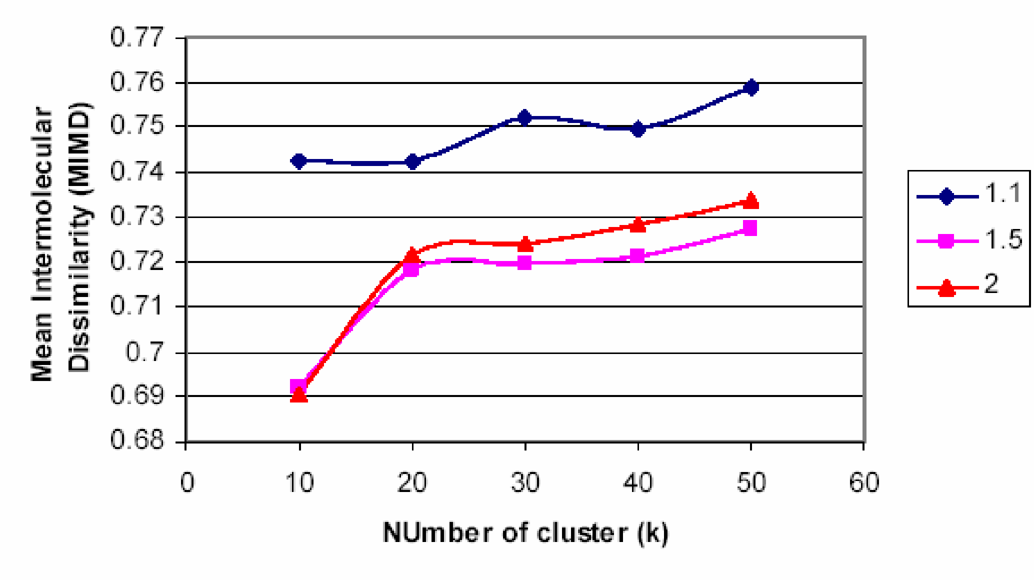


Figure 4.11. Dissimilarities within clusters for various values of *m*

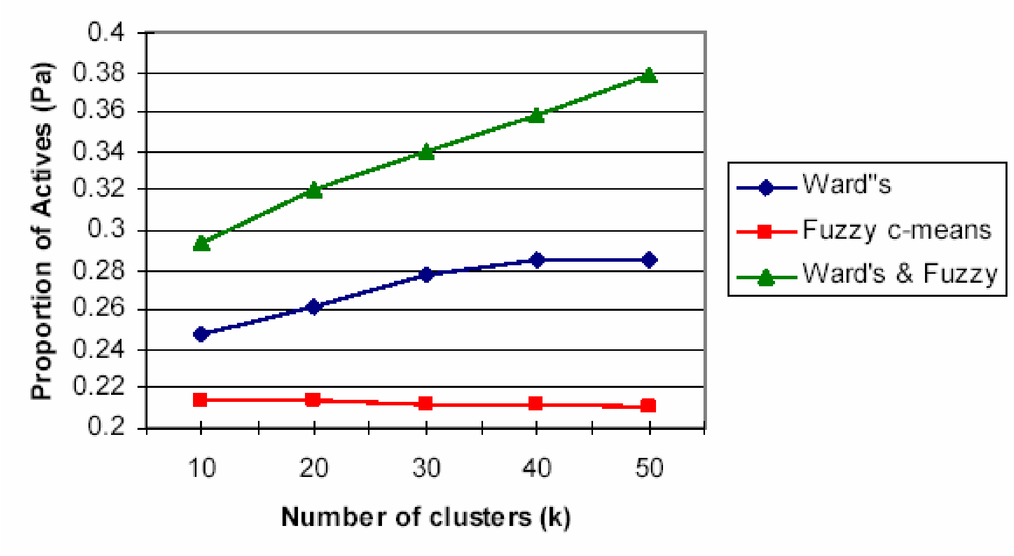


Figure 4.12. Performance based on Actives / In-actives Separation

The performance of the Wards seeded fuzzy c-means algorithm show better results than that of the other two methods based on the proportion of actives in active cluster subset. This result is shown in figure 4.12

The fuzzy c-means alone do not give any good results than the Wards but when accompanied by the initialization from the Wards clustering, the results become better than that of Wards. It means fuzzy c-mean is highly dependent on the initialization. If the prototype matrix is initialized properly, the local minima can be avoided.

The dissimilarity analysis of all three methods show that it is very high for the fuzzy c-mean and show a steady increase as the number of clusters in increased. For the rest of the two methods it shows an altogether different trend. It is minimum for low number of clusters and increases as the number of clusters are increased but then again start decreasing with the increase in number of clusters. This is shown in figure 4.13.

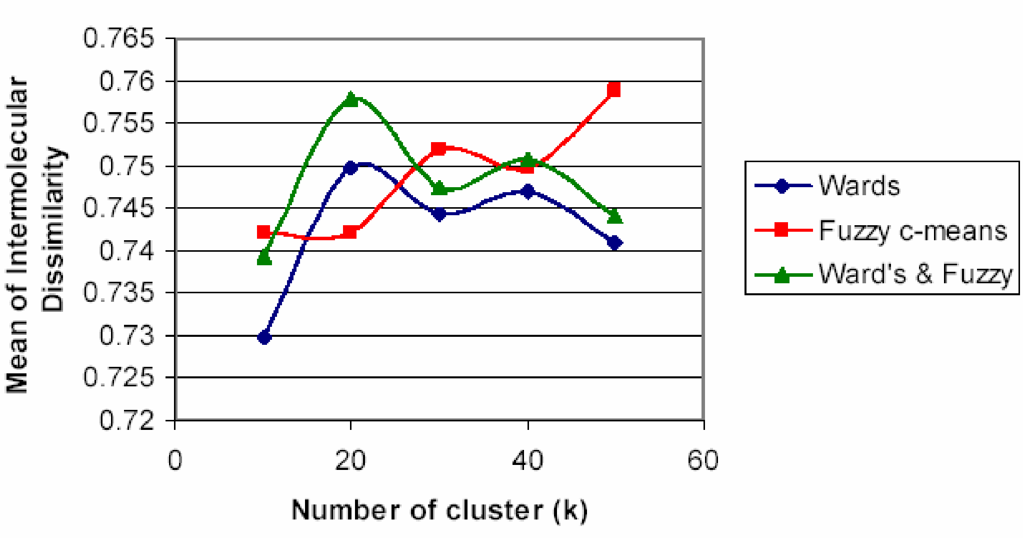


Figure 4.13. Dissimilarity within clusters as computed by MIMD

4.6. Summary

This chapter was dedicated to the application of fuzzy based clustering techniques in clustering the chemical structures. It should be noted that this report concern the clustering of chemical structures which are biologically active. So, in this chapter the two important fuzzy methods have been utilized. The importance of fuzzy clustering over the crisp (or hard) clustering is highlighted and the advantages that the fuzzy logic possesses have been enumerated. The algorithms for fuzzy c-means and fuzzy Gustafson Kessel algorithms have been dealt with in detail.

The results of the fuzzy clustering have been compared with the Wards and Group Average methods using the active / inactive separation based performance measure. In the last experiment another performance measure called the mean intermolecular dissimilarity (MIMD) have also been used to validate the resultant clusters quality. The performance of the algorithms has been evaluated using binary as well as multi class datasets.

The results of the fuzzy c-means show that its performance is better than the rest of methods in combining the actives compounds with the actives and inactives with the inactives for multi class dataset where topological indices have been used. In case of fragments based bits string the performance of fuzzy c-means show high dependence on the initialization. When initialized with clusters centres obtained with the help of Wards clustering, the performance is better than the Wards in terms of separation of actives from inactives in binary classification.

Generally, these results show that fuzzy methods are as good as the conventional methods of clustering, and so can be used in clustering of chemical structures for drug discovery.

Chapter 5

Hybrid Techniques and Compound Clustering

The combination of individual methods can perform better when their good traits are exploited in connection with the requirements of the problem at hand. The chemical databases are large and multidimensional, which needs robust and efficient methods to be analyzed. In the previous chapters a number of clustering techniques based on conventional, fuzzy and neural approaches have been discussed. The results obtained with the help of these methods suggest some directions to improve their performance by utilizing some benefits from the other methods.

This chapter discusses mainly two hybrid techniques developed, and applied to the clustering of chemical structures on their biological activities and properties. The chapter also discusses a genetic algorithm based clustering scheme and its evaluation on a two class dataset.

5.1. Fuzzy Kohonen Self-Organizing Feature Map

Most of the unsupervised neural networks such as kohonen neural network or neural gas network as discussed in chapter 3 are dependent on heuristics for the selection of parameters' values and so the termination is not based on the optimization of any model or process [131]. But it is not the only problem with the unsupervised neural networks. There are two more issues with the training of such networks. First the final weight vectors of the networks are dependent on the input sequence and the results vary with variation in initial conditions.

Second, several parameters of theses networks such as the learning rate and the neighborhood size (of the winner neuron) selection, and the way to change them during the course of learning, need to be varied from one dataset to another dataset.

According to Huntsberger and Ajjimarangsee [132] in order to make the Kohonen algorithm an optimizing process, the fuzzy c-means membership functions can be utilized to select good fit values for the learning rate and neighborhood size. The Kohonen update rule has been modified as follows:

$$W_{i,t} = W_{i,t-1} + \mu_{ik,t}\,(Z_k - W_{i,t-1}) \qquad 5.1$$

where μ_{ik} is the fuzzy membership value of the compound Z_k in cluster i , and $W_{i,\ t}$ is the weight vector for cluster I in time iteration t.

It is obvious from the update rule of the fuzzy Kohonen network (equation 5.1) that the learning rate and neighborhood size are now replaced with the membership values which are continuously changing with time, the compound and the cluster. However, the learning rule still have the room for improvements as the order of input samples can still affect the output and the results will still vary with initial conditions.

Later on Bezdek et al [131] have developed termination criteria and developed an improved algorithm by incorporating the following functions based on the membership function:

$$W_{i,t} = W_{i,t-1} + [\sum_{k=1}^{n} \alpha_{ik,t}\,(Z_k - W_{i,t-1})]/\sum_{k=1}^{n} \alpha_{ik}, t \qquad 5.2$$

where the new parameter α is given as

$$\alpha_{ik,t} = (\mu_{ik,t})^{mt}; \qquad m_t = (m_0 - 1)/\,t_{max} \qquad 5.3$$

The fuzzification exponent m is a function of time and so parameter which is called the learning rate parameter is not only dependent on the current value of the membership function but also on its fuzzification parameter. The initial value (m_0) of the fuzzification parameter can have any value greater than 1 and final value (m_∞) should not be less than 1.1. Usually the final value goes to 1. Bezdek et al [131] has further explained the membership functions and how they affect the winner and non winner neurons in the output layer of the network.

The learning rates for a fixed number of clusters c , and fixed m_t can be written in terms of the distance between the input sample Z_k and weight vector W_{ik} as follows:

$$(\mu_{\text{ik, t}})^{\text{mt}} = (\text{a}/\text{d}_{\text{ik, t}})^{(2\text{mt}/(\text{mt-1}))} \qquad 5.4$$

where a is any positive constant. It is obvious that the distance $d_{ik,\ t}$ will be smaller for the winner neuron and so its learning rate will become higher than the neurons far away from the input Z_k . Thus the membership functions are the natural responsible for the learning rate and the neighborhood size. Moreover, the memberships also update themselves with each iteration using the fuzzy c-mean algorithm.

5.1.1. Fuzzy Kohonen Network (FKN) Algorithm

The fuzzy kohonen neural network is now an optimization process, and the steps involved can be summarized as follows:

Step 1

Initialize the number of clusters c, the weight matrix $W_{ik,\ 0}$, the tolerance ε , initial value of the fuzzification m_0 , the maximum number of iterations t_{max} , also compute the change in fuzzification parameter

$$\Delta\text{m} = (\text{m}_0\ \text{- } 1.1)\ /\ \text{t}_{\text{max}} \qquad 5.5(\text{a})$$

Step 2

Compute the fuzzification parameter

$$\text{m}_{\text{t}} = \text{m}_0\ \text{- }\ \Delta\text{m t} \qquad 5.5(\text{b})$$

Step3

Compute the distances

$$\text{D}_{\text{ik}} = (\text{Z}_{\text{k}} - \text{W}_{\text{i}})^{\text{T}}(\text{Z}_{\text{k}} - \text{W}_{\text{i}}) \qquad 5.5(\text{c})$$

Step 4

Compute the memberships and learning rates

$$U_{ik} = \frac{1}{\sum_{j=1}^{c}(D_{ik}/D_{jk})^{2/(m-1)}}, \quad \begin{array}{l} 1 \le k \le n \\ 1 \le i \le c \end{array} \qquad 5.5(d)$$

$$\alpha_{ik,t} = (U_{ik,t})^{m(t)} \qquad 5.5(e)$$

Step 5

Compute update weights

$$W_{i,t} = W_{i,t-1} + [\sum_{k=1}^{n} \alpha_{ik,t} (Z_k - W_{i,t-1})]/\sum_{k=1}^{n} \alpha_{ik}, t \qquad 5.5(f)$$

Step 6

Compute the update change

$$E_t = \max \|W_t - W_{t-1}\| \qquad 5.5(g)$$

Step 7

If $E_t <= \varepsilon$ OR $t == t_{max}$ Then ***STOP***

Else ***GoTo Step 2***

5.2. A Robust Fuzzy Self – Organizing Network

The enhanced neural gas algorithm, when used with datasets containing weak outliers, the data elements that are not very far away from the main groups found in the dataset, results in coincident clusters. Although, the algorithm is of great advantage when there are strong outliers in the datasets, but at the same time highly expensive when used with large datasets. Moreover, the algorithm has two important parameters the learning rate and neighborhood size that needs to fix in advance, and so will vary from one dataset to another dataset. Keeping these problems in mind, a new algorithm comparatively efficient, with almost no parameter settings, insensitive to the presence of outliers and noise is developed. The algorithm implies the fuzzy memberships to get a strength value for the update of weight vector corresponding to individual output neurons in response of a particular input sample. Thus the update strength value is a function of the position of the output neuron and the input sample and is directly independent of the iterations level as compared to that of the ENG network where the update strength value is highly dependent on the iteration level. The weight update formula for this algorithm is given below:

$$W_i(t) = \quad W_i(t-1) + \frac{\sum_{k=1}^{n} \alpha_{ik}(t).\exp(-{}^{\|Z_k - W_i(t-1)\|}\!\big/_{\beta d_i(t-1)}).\sigma_{ik}(t).\frac{(Z_k - W_i(t-1))}{\|Z_k - W_i(t-1)\|}}{\sum_{k=1}^{n} \alpha_{ik}(t)}, \quad i = 1,2,...,c \qquad 5.6$$

Where *σ_{ik}* is the factor highly dependent on the position of the data element Z_k from the cluster center W_i. When the data element Z_k is far away from the center its Euclidean distance will be equal or greater than the mean historical distance and in such a case the harmonic distance is averaged with the historical distance in order to reduce the effect on the new center. The historical distance is computed as follows:

$$\sigma_{ik}(t) = \begin{cases} d_{ik}(t) & \mathit{if} \quad \|Z_k - W_i\| \geq d_{ik}(t-1) \\ \|Z_k - W_i\| & \mathit{if} \quad \|Z_k - W_i\| < d_{ik}(t-1) \end{cases} \qquad 5.7$$

where

$$d_{ik}(t) = \begin{cases} \{1/2[1/d_{ki}(t-1) + 1/\|Z_k - W_i\|]\}^{-1} & \mathit{if} \quad \|Z_k - W_i\| \geq d_{ik}(t-1) \\ 1/2[d_{ik}(t-1) + \|Z_k - W_i\|] & \mathit{if} \quad \|Z_k - W_i\| < d_{ik}(t-1) \end{cases} \qquad 5.8$$

and

$$d_i(t) = \left[\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\|Z_k - W_k\|}\right]^{-1} \qquad 5.9$$

Similarly, the learning parameters α_{ik} are computed using the fuzzy membership functions μ_{ik} which are computed using equations 5.5(a - e).

5.2. A Hierarchical Fuzzy Algorithm

Fuzzy clustering is the intrinsic solution to the problem of overlapping data, where the data elements can be member of more than one cluster as detailed in the previous chapter. The traditional clustering methods do not allow this by restricting the data elements to belong to only one of the many clusters exclusively. There can be almost three types of partitioning concepts, the traditional hard or crisp one where a compound can belong to only one cluster and so the membership degree of the compound is said to be 1 in any one cluster and zero in the rest of the clusters. Another approach is provided by the fuzzy logic where the membership degree of a compound can be [0, 1] and so the compound can belong with varying degree to more than one cluster [9, 112]. In both of these partitioning scenarios the memberships $\mu(i,k)$ or (μ_{ik}) follow a few conditions such as the sum of the membership values over a range of clusters *c* is always equal to one:

$$\sum_{i=1}^{c} \mu_{ik} = 1 \quad 1 \leq k \leq n \qquad 5.10$$

$$0 < \sum_{k=1}^{n} \mu_{ik} < n \quad 1 \leq i \leq c \qquad 5.11$$

Where n is the number of compounds in the dataset, c is the number of clusters and I and k are the indexes for the clusters and data elements respectively.

In the possibilistic partitioning [114], this constrained is also relaxed and the sum of the membership degrees is not required to be equal to one, however, clustering algorithms based on this theory are out of the scope of this work.

The hierarchical fuzzy algorithm is a recursive procedure of fuzzy clustering, where each cluster formed is further re-clustered. The number of child clusters in each recursive call can be 2, 3 or any other number greater than 1. However, here in every recursive call the value of *c* is kept at 2 to obtain a binary tree like order on the structures, a fashion more suitable and historical to the chemical structures based on their biological activities. The inputs are a *n X m* data matrix *Z* composed of *n* (number of structures in each recursive call) rows of compounds $Z_k \in \Re^m$ and m columns of features. The output of each recursive call is a c *X n* membership matrix *U*. The two child clusters are formed using the membership matrix U, where a structure Z_k can be a member of either one of the two clusters if the membership of one is greater than the other to some extent, or can be part of both the clusters if their membership degrees do not show much difference. Once a cluster is partitioned into its child clusters, the membership matrix is discarded but the algorithm keeps the necessary global information in the constituent clusters by adding the structures which are closely related to both the clusters. Thus in each recursive call a new membership matrix is generated and optimized based on the local information of the cluster.

This recursive process of clustering continues until every cluster is a singleton (a cluster containing only one structure) or when an optimal partition is obtained. For this purpose the partition validity measure suggested by Bocker et al [133] is adopted. The clustering process is repeated for a number of threshold and at the end of each repetition, the number of singletons, the number of non singleton clusters, and a distance measure D_{max} (Equation 5.12) are calculated and plotted against the thresholds to find the optimal threshold. Once the optimal threshold is obtained from the graph (one shown in Fig 2), the clustering process is repeated once again.

$$D_{\max} = \sum_{i} \max[d(Z_k, V_i)], \quad 1 \le k \le n \qquad 5.12$$

where d is the Euclidean distance, between the structure $Z_k \in Z$ and the prototype of the cluster V_i , and n is the number of structures in each cluster. The value of D_{max} represents the maximum deviation of the clusters from their prototypes.

The algorithm is a divisive binary tree like procedure, where each cluster is divided into 2 clusters by means of the fuzzy c-means algorithms as described stepwise in section 4.4.3. The clustering process continues until the maximum diameter of each cluster become equal to or less than a predefined threshold. The ultimate and optimal threshold value is selected heuristically by the user from a range of thresholds. The main steps of the algorithms are ordered below:

Run1: For finding the optimal threshold

(i) A threshold is selected from a range of thresholds

(ii) The value c is initialized which is 2 for binary trees, the membership matrix U is initialized

(iii) The dataset is clustered using the fuzzy –c-means algorithm

(iv) Each of the cluster is checked if the deviation is greater than the Threshold selected, then go to step (ii) for sub-clustering the resultant cluster

(v) Plot the number of clusters, singletons and the metric D_{max} against the range of threshold

Run2: (i) Select the optimal threshold through visual inspection of the graph resulted in step (v) of Run 1

(ii) Repeat the algorithm for the last time using the optimal threshold.

In clustering a good method is supposed to combine highly similar activity structures together, so high number of singletons is not considered a good gesture. Thus, an appropriate point for a good clustering will be a threshold for which the number of singletons is a minimum.

5.3. Validity of Clustering for Fuzzy Hierarchical Algorithm

In most of the clustering methods, the number of cluster has to be provided a priori to the clustering process. Thus the system has two main issues, one how many clusters are there and second how to find this number automatically. In the literature, we find a large number of validity measures for the evaluation of clustering methods, but each and very validity measure succeeds in finding a certain type of clusters and fails in finding other types of clusters. Most of these measures tries to find a particular geometric shape (lines, circles, spheres, ellipsoids, curves etc) clusters, because of the different distance measure used in the clustering process [134]. Most of the cluster validity measures try to find more compact clusters increases the separation among the clusters and decreases the volume of the clusters.

In this work we have developed a new validity measure that tries to find more compact, highly separated clusters, and minimize the formation of singletons. The last point the formation of singletons is of high importance in the clustering of chemical clusters especially with hierarchical clustering methods. The formation of singletons is not considered a good trait of clustering chemical structures into biological activities as the activity of a singleton cluster can not be used for the prediction of the activity of another structure [15]. The fuzzy hyper volume and Xe Benie indices have been used in many researches due to their accurate prediction of the number of clusters in a datasets. Here these indices are combined into one index which we call the overall index (OAI) of cluster validity is given as:

$$OAI = \frac{\textbf{Non - Singleton Clusters} + \textbf{RelClusDist}}{\textbf{Singletons} + \textbf{Volume} + \textbf{XeBenie}} \qquad 5.13$$

The variables used in the above formula such as Non-singleton clusters, the RelClusDist, Volume etc are all normalized values over a range of thresholds of hierarchical fuzzy clustering algorithm discussed in the previous section. The Non singleton clusters represent the number of clusters less the number of singletons. The RelClusDist is a distance measure which is the sum of all the mean distances of all the clusters in a partition given as:

$$\mathrm{Re}\,lClusDist = \sum_{i=1}^{c} \frac{\sum_{k=1}^{n_i} d(i,k)}{n_i} \qquad 5.14$$

where $d(i,k)$ is the Euclidean distance between a molecule *k* of the cluster *i* and its center, n_i is the total number of molecules in cluster *i* and *c* is the total number of non singleton clusters.
The volume of the partition is defined by a covariance matrix as follows:

$$Volume = \sum_{i=1}^{c} \left[\det(F_i)\right]^{1/2} \qquad 5.15$$

where

$$F_i = \sum_{k=1}^{n_i} (Z_k - V_i)^T (Z_k - V_i) \quad 1 \le i \le c \qquad 5.16$$

The Xe Benie index gives a smaller value for the most compact and well separated partition of a dataset, so its addition to the denominator will increase our new index OAI when the most compact and well separated partition is obtained. The XeBenie index is given as:

$$XeBenie = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n_i} \left\| Z_k - V_i \right\|^2}{n \min_{i \neq j} \left\| V_i - V_j \right\|^2} \qquad 5.17$$

The results obtained with this new index are very encouraging and are discussed here in section 5.5.3 with the help of the Golub's Lekuemia and fisher's Iris benchmarked datasets.

5.4. Genetic Algorithm and Clustering

The theory of Genetic Algorithm was proposed by Holland [135], who was inspired by the theory of evolution. Genetic Algorithms have been applied successfully to a large number of problems in automatic design, pattern recognition, robotic control, and speech and image processing. The output of the gradual adaptation of natural chromosomes has led to the production of fittest individuals who possess higher survival capabilities in an ever-changing environment. This evolving process of natural genetics is artificially simulated in Genetic Algorithms to find the optimum solution to an optimization problem. Genetic Algorithms apply some of natural evolution mechanisms such as crossover, mutation, and survival of the fittest to optimization and machine learning. Genetic Algorithms provide an efficient search method working on population and have been applied to many problems of optimization and classification [136]. In the conventional Genetic Algorithms, each solution is presented by fixed length string. A general Genetic Algorithms process consists of the following steps [137]:

(1) Initialize the population of chromosomes,

(2) Calculate the fitness for each individual in the population,

(3) Reproduce the individuals selected to form a new population according to each individual’s fitness,

(4) Perform crossover and mutation on the population,

(5) Repeat Steps (2) through (4) until some condition is satisfied.

Reproduction makes a copy of a solution. The operator is an artificial version of natural selection. Using reproduction, individual chromosomes are selected according to their fitness, which is evaluated by using an objective function. The result of reproduction means that each individual chromosome with a higher fitness value will have a higher probability to contribute one or more chromosomes in the next generation.

Crossover operation swaps some part of genetic bit string within parents. It emulates just as crossover of genes (chromosomes) in real world that descendants have inherited characteristics from both parents. There are two simple ways in terms of the number of crossover points that are used to crossover operation: single-point and two-point crossover. Mutation operation inverts some bits from whole bit string at very low rate. In real world we can see that some mutants come out. Each individual in the population evolves to getting higher fitness level from generation to generation.

For a more detailed description of Generic Algorithms Goldberg [136], Adeli and Hung [138], and Mitchell [139] can be consulted.

5.5. Experiments and Results

5.5.1. Experiment1

The experiments based on the application of fuzzy kohonen self organizing neural networks are discussed in this section. The results obtained with the help of fuzzy SOM network are given comparison with the Wards and Group Average and the other neural network discussed in chapter 3, such as Kohonen and neural gas networks.

Data and Descriptors:

The same dataset as described in section 3.4.1 is used in this work. The topological descriptors have been used as features and PCA [67] have been used to reduce the dimensionality.

Results and Discussion:

For the evaluation of the performance, the active cluster subset method has been used. In this method the dataset is clustered for a number of sets of clusters and then merged. The individual clusters containing at least one active compound and which are not singletons are merged to gather to form the active cluster subset. The proportion of actives in this cluster subset is determined and plotted against the number of clusters.

The fuzzy Kohonen neural network has only one parameter (initial fuzzification m_0) that need to be initialized in advance and the rest of the parameters are selected with the help fuzzy membership functions which are automatically determined using the fuzzy objective function optimization. The initial fuzzification parameter is large and then continuously decreased as the learning process succeeds, until it reaches its final minimum value of 1.1.

The results obtained show that the large values of the fuzzification parameter does not give good results. The results obtained with our dataset is the best of *m=2.0*, for all other values larger than 2 the results are poor. These results are shown in figure 5.1.

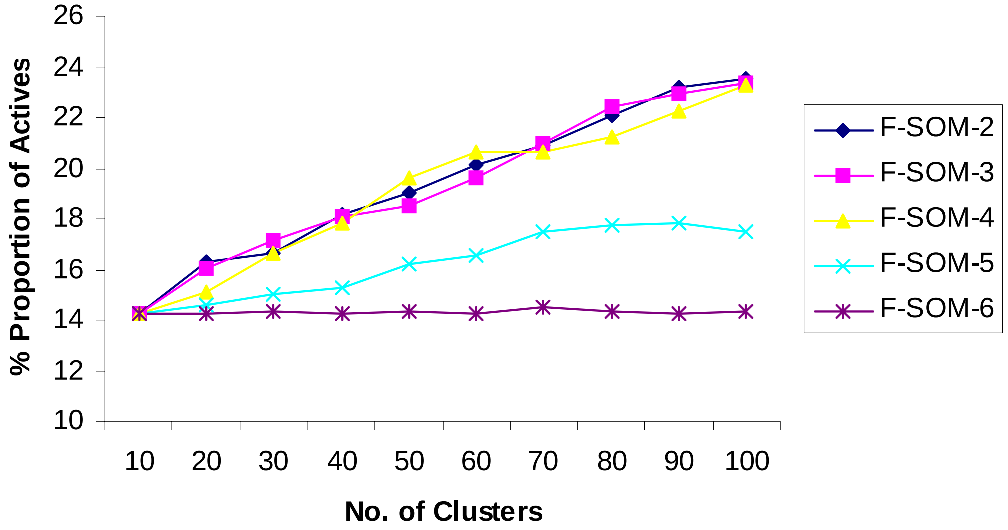


Figure 5.1. Performance of Fuzzy Kohonen SOM Network with different initial value for the fuzzification parameter.
F-SOM-2, F-SOM-3 stands for fuzzy SOM initialized with fuzzification parameter of 2 and 3 respectively.

It has been observed that performance of the algorithm becomes better and better as we decrease the fuzzification parameter.

Once the best value for the fuzzification parameter has been obtained, the results of the algorithm have been compared with Ward's, Group Average and neural networks. These results are shown in figure 5.2.

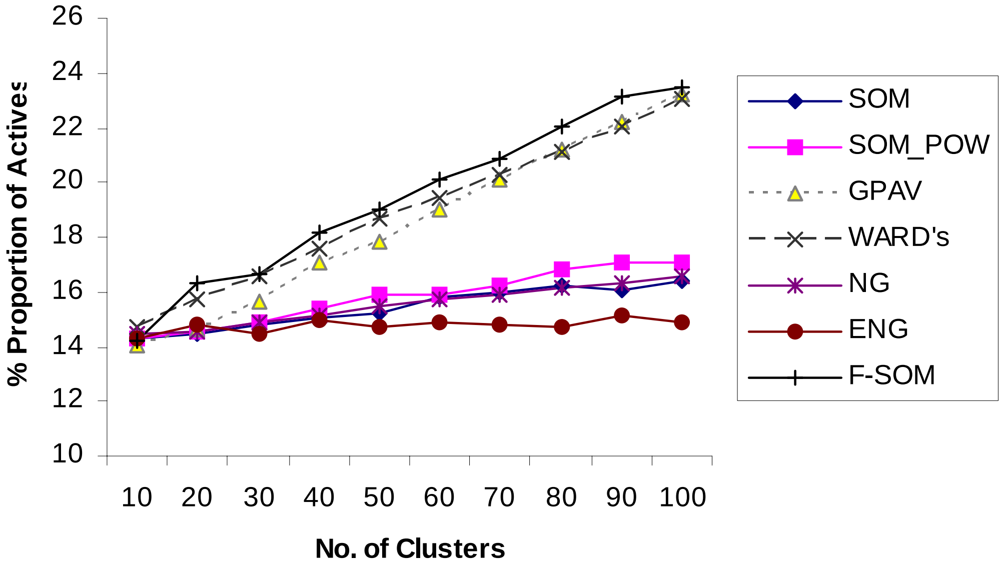


Figure 5.2. Performance of the Fuzzy SOM Neural Network is compared with other clustering methods.
GPAV - Group Average Method, WARDS - Ward's Clustering Method, SOM – Kohonen Self-Organizing NN with linear learning rate, SOM_POW – Kohonen Self-Organizing NN with exponential learning rate, NG – Neural Gas Network, ENG – Enhanced Neural Gas Network, and F-SOM – Fuzzy Kohonen Self Organizing NN.

The results of the fuzzy SOM are very excellent than that of other neural networks. The Kohonen neural network which is the best and mostly used self organinzing network can not perform well like the fuzzy Kohonen due to its dependence on heuristics about the parameters and their updating. The Fuzzy kohonen algorithm converges excellently and it has been found that there are no unlabelled cells in the self-organizing map obtained. This is shown in table 5.1.

It has been observed that as the number of clusters is increased the number of unlabelled cells of the self- organizing maps increases in neural methods. The enhanced neural gas performance is the worst due to its behavior to reduce the number of clusters. This algorithm tries to avoid the noise and outliers to be accommodated in clusters of their own and tries to confine them to the inside of the clusters closer to them, which results in less number of clusters.

Table 5.1
number of Unlabelled Cells

No. of Clusters	Methods					
	Fuzzy SOM	Kohonen SOM	Neural Gas	Enhanced Neural Gas	Wards	Group Average
10	0	2	2	2	0	0
20	0	0	5	3	0	0
30	0	1	8	14	0	0
40	0	2	9	16	0	0
50	0	4	11	32	0	0
60	0	8	16	34	0	0
70	0	8	21	38	0	0
80	0	12	26	49	0	0
90	0	15	29	47	0	0
100	0	20	34	62	0	0

The fuzzy SOM has distributed the compounds equally well according to their natural cluster occurring and there is no empty cluster like the other molecular clustering methods like Wards and Group Average. It should also be noted that the unlabelled cells can not be found in agglomerative clustering due to their merging rather than divisive nature. However, when the number of clusters was increased too much some empty clusters were found. For example, for 400 clusters there were 6 unlabelled cells and for 1000 clusters their number was 106.

5.5.2. Experiment 2

This experiment is about the application of the fuzzy hierarchical algorithm developed here in section 5.2. especially for the clustering of overlapping clusters. The dataset used in this experiment contain overlapping biological activities.

Dataset and Descriptors:

In this work three datasets been utilized to evaluate the performance of the proposed algorithm: two benchmark datasets known as the Fisher's Iris dataset [140] and Golub's Lekuemia datasets [141] and one drug dataset composed of bioactive molecules exhibiting overlapping as well as non-overlapping activities collected from the MDDR database. The MDDR database contain 132000 biologically relevant compounds taken from patent literature, scientific journals, and meeting reports [93]. Each entry of the database contain a 2D molecular structure field, an activity class and an activity class index fields besides many other fields like biological investigation phase, chemical abstract service (CAS) [34] compound identity , and patent information fields. The activity index is a five digit code used to organize the compounds' structures based on biological activity; for example the left most one or two digits describe a major activity group and the next three digits describes sub activities inside the bigger activity. For example, the activity index 31000 shows a large activity of antihypertensive agents and the activity indexes 31250, 31251 show Adrenergic (beta) Blocker, Adrenergic (beta1) Blocker respectively. The dataset used here comprised of 12 major activities where each group can further be divided into a few sub categories. Initially 55000 compounds have been extracted from the database using a number of filtering strategies (as described below in equations 5.18-5.19). The number of compounds in dataset1 (DS1) was 29843 and dataset2 (DS2) 6626. The DS1 contain exactly non-overlapping structures where each compound in the dataset can exhibit only and only one activity among the list of activities selected for this work. And the dataset DS2 contain bioactive (compounds exhibiting only two activities) such that each compound exhibit two activities only. Let ***A*** be the set of activities selected and ***l*** be the length of activities in this set, then the DS1 is a superset of sets ***D$_i$*** , a set of compounds exhibiting activity $i \in A$. If $z \in DS1$is an arbitrary compound then,

$$DS1 = \quad \{z \in D_i \mid \quad i \in A \wedge i \notin A - \{i\}, \quad i = 1,2,...,l\} \qquad 5.18$$

Similarly, DS2 is a superset of sets ***D$_{ij}$***, where compound exhibit two activities $i, j \in A$ and So,

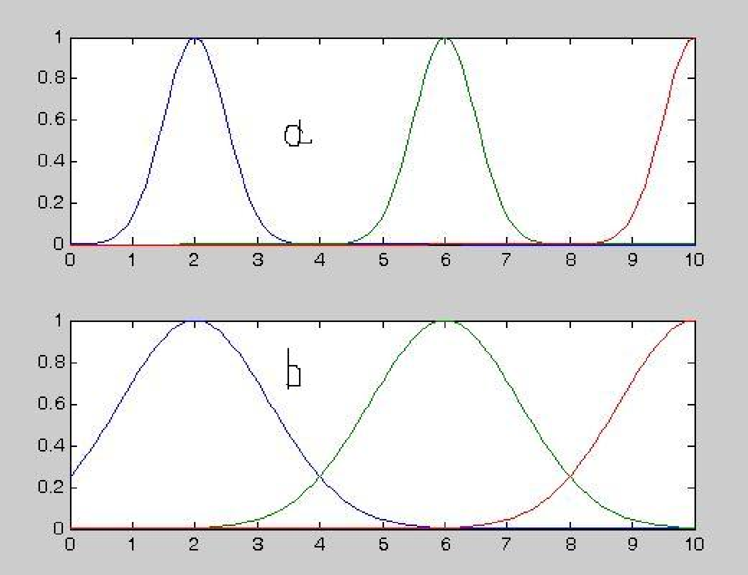


Figure 5.3. (a) Describes the non-overlapping clusters whereas (b) describes the overlapping clusters
The vertical axis plots the activity of the compound structure and horizontal axis plots the number of activities

$$DS2 = \left\{ z \in D_{ij} \mid \quad i, j \in A \wedge i, j \notin A - \{i, j\} \quad \begin{array}{l} i = 1,2,\ldots,l-1 \\ j = i+1,\ldots,l \\ \end{array} \right\} \qquad 5.19$$

By combining these two datasets, another dataset DS3 has been organized in the same way as depicted in figure 5.3. It contains single activity compounds from DS1 and in between any two activity groups there are bi-activity molecules from DS2 which belong to both the groups on its right and left.

The descriptors generation or features extraction is an important step in computational clustering of molecular structures and other problems such as classification and quantitative property/activity relationship modeling. A number of modeling tools are available that can be used to generate structural descriptors. In this work, we use the Dragon software [33] to generate around 99 topological indices for the molecules. Topological indices are a set of features that characterize the arrangement and composition of the vertices, edges and their interconnections in a molecular bonding topology. These indices are calculated from the matrix information of the molecular structure using some mathematical formula. These are real numbers and possess high discriminative power and so are able to distinguish slight variations in molecular structure. This software can generate more than 1600 descriptors which include connectivity indices, topological indices, RDF (radial distribution function) descriptors, 3D-MORSE descriptors and many more.

Table 5.2
Selected Topological Indices

TI	Description
Gnar	Narumi geometric topological index
Hnar	Narumi harmonic topological index
Xt	Total structure connectivity index
MSD	Mean square distance index (Balaban)
STN	Spanning tree number (log)
PW2	path/walk 2 – Randic shape index
PW3	path/walk 3 – Randic shape index
PW4	path/walk 4 – Randic shape index
PW5	path/walk 5 – Randic shape index
PJI2	2D Petitjean shape index
CSI	Eccentric connectivity index
Lop	Lopping centric index
ICR	Radial centric information index

Scaling of the variables generated is very important in almost all computational analysis problems. If magnitude of one variable is of larger scale and the other one is of smaller scale then the larger scale variable will dominate all the calculations and effect of the smaller magnitude variables will be marginalized. In this work all the variables used were normalized such that the maximum value for any variable is 1 and the minimum is 0.

In order to reduce the descriptor space and to find the more informative and mutually exclusive descriptors a feature selection method principal component analysis (PCA) [67] was used. PCA was carried out using the MVSP 3.13 [93]. It has been found that 13 components can represent more than 98% of the variance in the dataset. The input to the clustering system is thus a 13 X 28003 data matrix. The 13 selected topological indices are shown in Table 5.2.

Results and Discussion:

Three datasets have been used to evaluate the performance of the clustering process. These include two small size benchmark datasets a leukemia cancer dataset and fisher iris dataset, and chemical dataset DS3 described earlier in detail. Leukemia dataset is a collection of 72 genes expressions belonging to two types of cancer, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Almost 62 of the specimens for this genes expression data were obtained from bone marrow samples of the acute leukemia patients while the rest had been collected from the peripheral blood samples. The fisher's Iris dataset consists of 150 random samples of flowers belonging to the Iris species setosa, versicolor, and virginica. For each of the specie, the dataset contain 50 samples and each sample consists of four variables, sepal length, sepal width, petal length and petal width.

The clustering of the iris dataset into three clusters is in line with its principal component analysis. When the two principal components are plotted, there is a very clear boundary between the cluster shown in blue and the rest of the two clusters goes to the other axis and when these two cluster points are plotted on the last two of the three principal components, it is observed that the two yellow and green clusters points go to different quadrants.

Since these two real datasets are almost non-overlapping but when the number of clusters decreased, the accuracy (performance) of the clustering degrades. These two results are for the threshold level 0.25 (iris) and 3.0 (leukemia). As the threshold is decreased, the clustering accuracy increases but resulting in more number of clusters and as we increase threshold the number of clusters decreases and are more heterogeneous. These results are shown in figure 5.4.

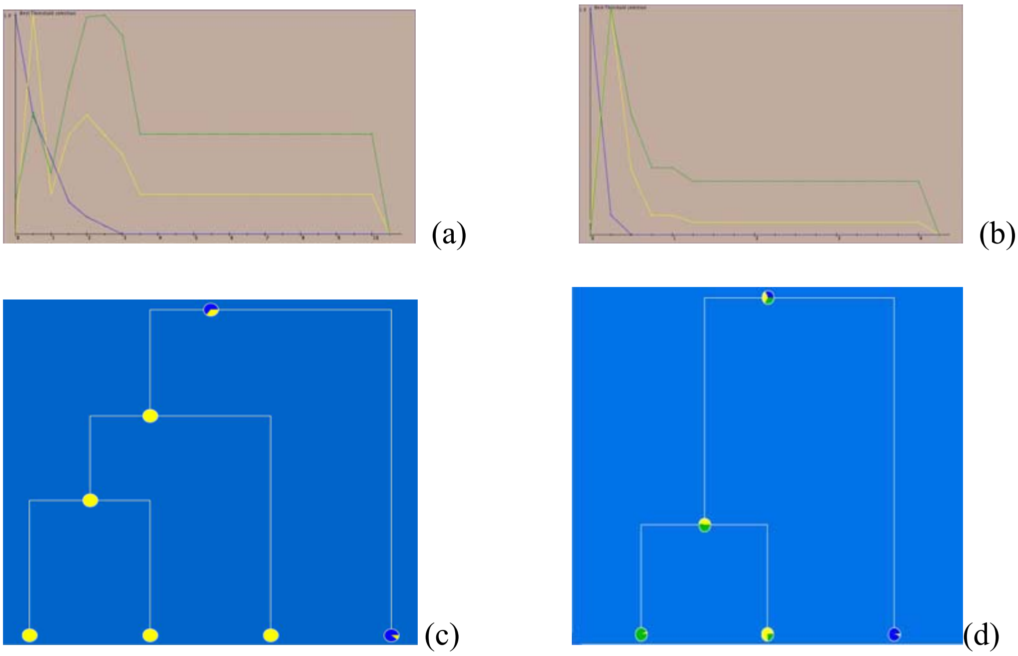


Figure 5.4: Clustering results of (a) lekuemia threshold, (b) iris threshold (c) leukemia clustering tree and (d) iris clustering tree

After confirming the results with the help of benchmark datasets, the methods was applied to the real molecular dataset DS3 developed in section 2. This dataset contain around 12 biologically active and overlapping clusters and the objective of the work is to evaluate the clustering performance of the developed hierarchical fuzzy c-mean algorithm. For this purpose, we use the active cluster subset method [15]. A threshold range of 0.01-0.1 with an increment step of 0.01 was used in this work. For each threshold a number of clusters were obtained. Some of the clusters obtained may be having only actives or inactives structures but many of them will have both. The clusters having at least one active structure are combined to make one super cluster called the active cluster subset. This subset of the dataset used should not contain any of the

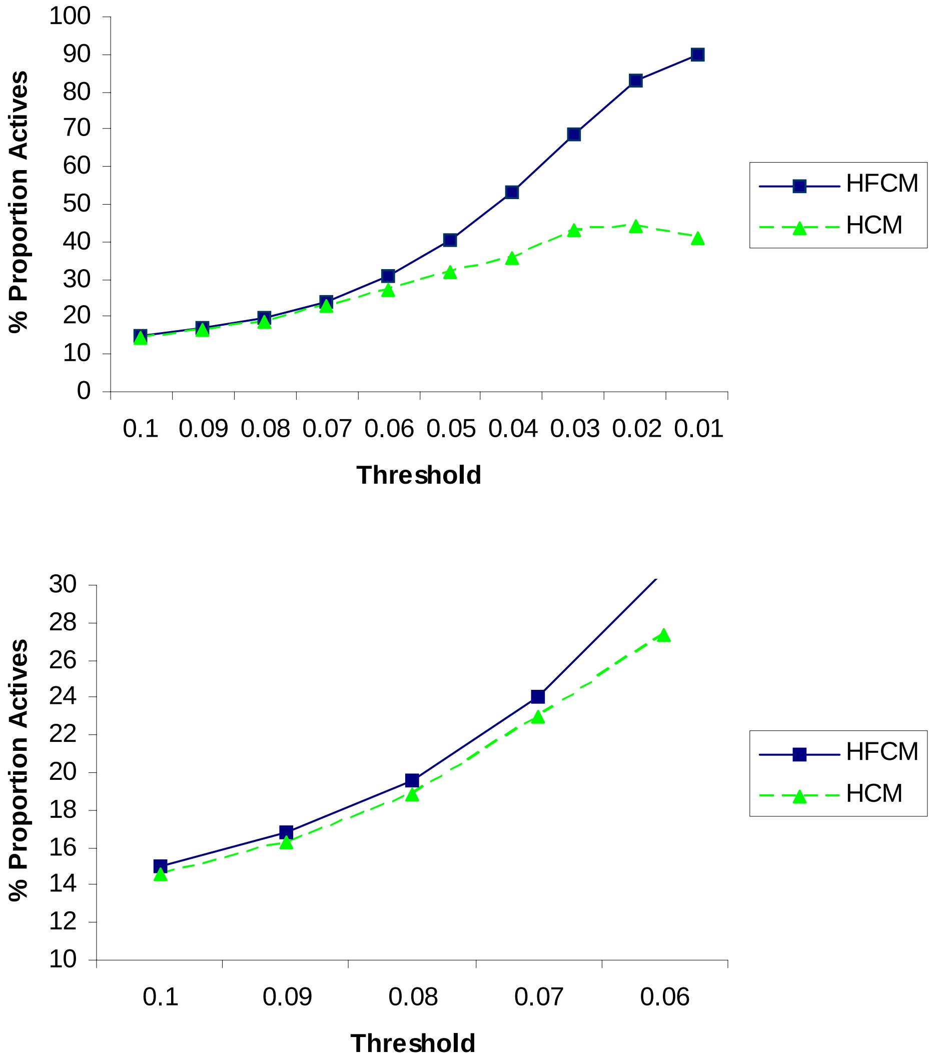


Figure 5.5. Performance of the HFCM and HKM
(a) -Shows the % proportion of actives in active cluster subset for various threshold level (b)-a zoomed view of the graph inside the red rectangle in (a) HFCM-hierarchical fuzzy c-mean, HCM-hierarchical c-means

singletons, as the singletons do not give any clue about the performance of the clustering method. The proportion of actives to inactive structures in the active cluster subset is determined. The clustering method is required to group together the active structures into one group and the inactive structures into a separate group. For each activity group of the dataset, the structures belonging to that activity group were taken as active and the rest of the groups were taken as inactive. The process is repeated for all the 12 bioactivity groups of the dataset for all the clusters obtained for each threshold level and an average proportion was determined.

The fuzzification index of the fuzzy c-mean determines the spread in the dataset [142] whose value can range from 1.1 to any finite number, a smaller value means that the data and natural clusters are spread over wide area (volume). As the value of fuzzification index is increased the data and the clusters within becomes more and more compact. The best value of the fuzzification index for which the best clustering can be obtained depends on the dataset used. So, first a fuzzy c-mean method was used to determine the best value for fuzzification index, and was found to be 1.4. The performance of the hierarchical fuzzy c-means is shown in figure 5.5, for various values of the threshold level. The threshold level is decreased from level 0.1 to 0.01 with a step of 0.01 and is plotted along the x-axis. Since, the threshold represents the radius of the cluster and when it is decreased the number of clusters (leaves) increases. The performance of HFCM is better than that of HCM even for very low number of clusters, and for large number of clusters (when the threshold is small) its performance is excellent.

The above results have been obtained for the maximum membership value. The compounds go to one of the two clusters if its membership for that cluster is greater than the other one. The compound is regarded as overlapping only if it has the same membership in both the clusters. In another experiment the compounds have been restricted to go to one of the clusters only if its membership for the cluster is greater than that for the other cluster by some fixed margin. If the membership of the compound in one of the two clusters is not different than the other by this margin or threshold, the compound is assigned to both the clusters. The results of this process show some improvement of performance over the previous method as evident from figure 5.6.

It has been observed that as the margin is increased the number of clusters increases at a very high rate (may be called exponential increase), which needs very high computational and memory resources so we have to limit the experiment to a maximum margin of 0.3, beyond this margin a system crash is evident.

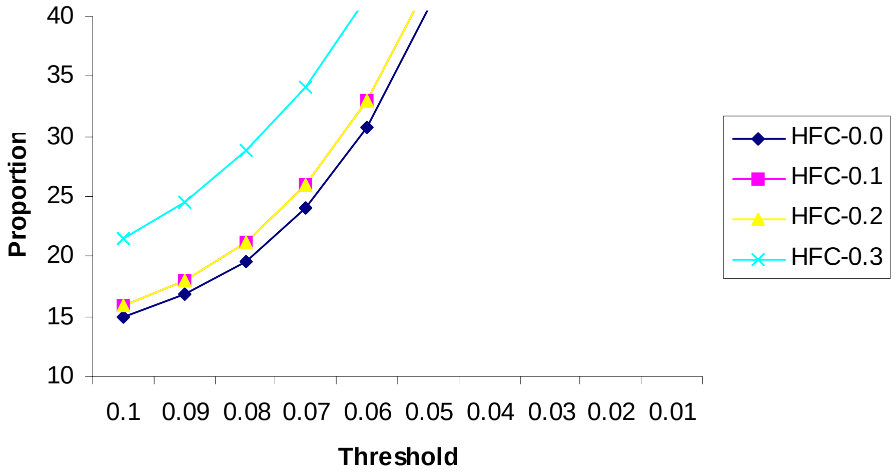


Figure 5.6 Performance of HFCM Algorithm with variations in Membership Margin

5.5.3. Experiment 3

In this experiment, the results of the newly developed validity measure are discussed. In this experiment the Golub's Leukemia and Fisher's Iris datasets have been used to evaluate the validity index. These datasets have been discussed in section 5.4.2 in more detail.

Results and Discussion:

The Leukemia dataset has 72 genes expression for two types of leukemia; the first class contains 41 examples while the second class contains 31 examples. The Sammon mapping [] has been used to analyze the dataset visually and manually. For each of the dataset, the clustering process has been repeated for a number of distance thresholds that limits the clusters sizes. For Leukemia dataset this threshold ranged from 0.5 to 10.0 with a step size of 0.5, whereas the threshold for the Iris dataset was between 0.05 and 2.0 with a step size of 0.05. These thresholds are shown in figure 5.7. It has been found that the best partitioning was obtained for a distance threshold of 2.5 for the leukemia dataset where the value of our new validity index is at the maximum as is shown in figure 5.7(a). Similarly, for the iris dataset the algorithm has been run repeatedly five times and two best thresholds have been obtained instead of only one. These threshold values were 0.20 and 0.25 one of which is shown in figure 5.7(b) where the value of OAI is maximum at 0.20.

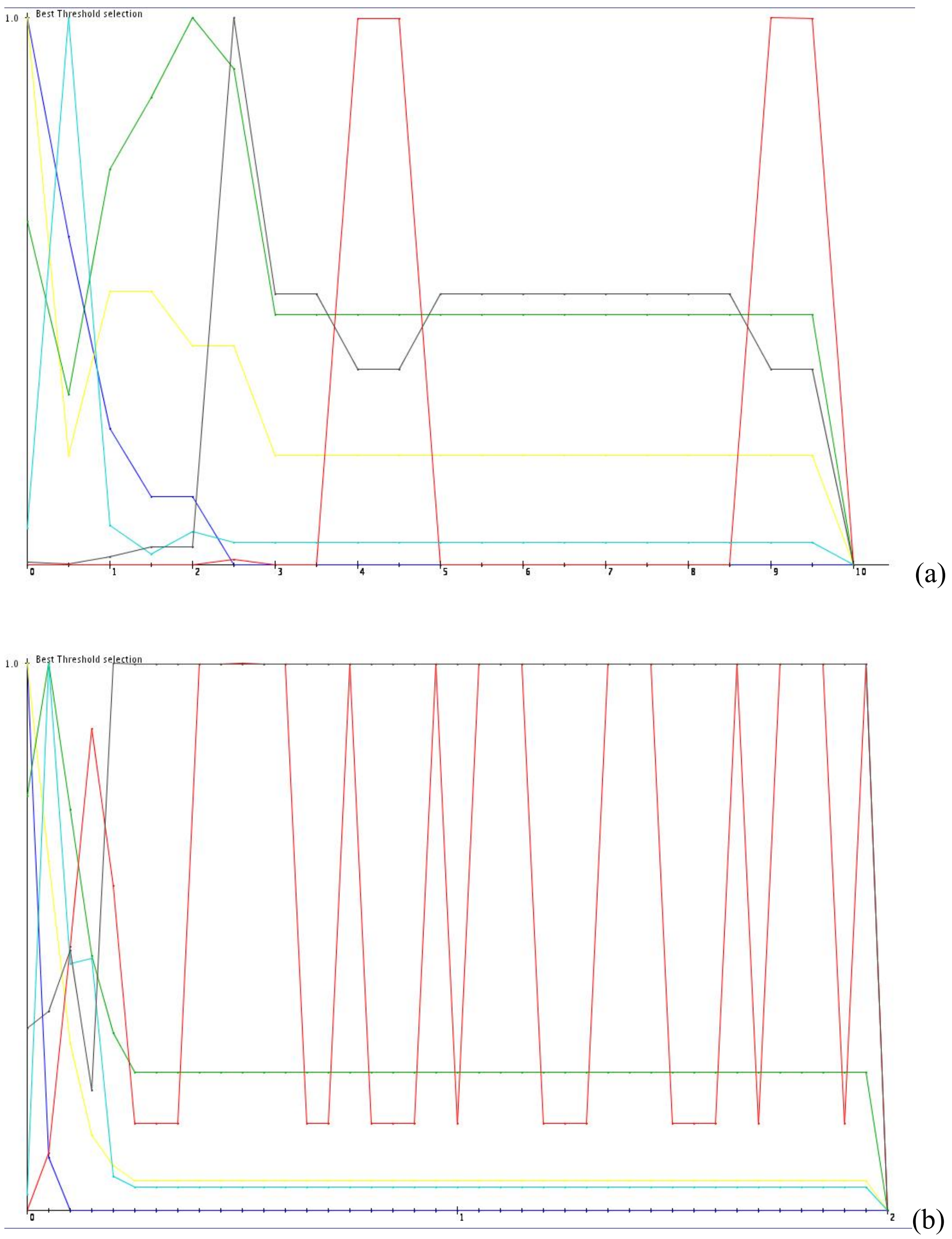


Figure 5.7 Threshold selections for (a) Lekuemia and (b) Iris datasets
Color codes: Blue – No. of singletons, Yello – No. of Non singleton clusters, Green – Average Cluster Distance, Red – Cluster hypervolume, Cyan – Xe Benie Index and Black – our new index OAI

It has been observed that the leukemia dataset has one comparatively compact cluster and the other cluster is highly dispersed, so, in our result for the threshold 2.5, the compact group has been assigned 6 data elements from the dispersed group and most of these 6 data elements lie very close to the compact group boundary. This result is scatter plotted in figure 5.8(a). As already noted the partitioning of iris dataset is more complicated and it is very difficult to partition it into three clusters. For the threshold value 0.25 we obtained three clusters and for

0.20 we obtained 5 clusters. One of the three clusters is highly compact and separated whereas the other two clusters are not properly separated. Among the last two clusters 18 data points shown in cyan in figure 5.8(b) are not clustered accurately.

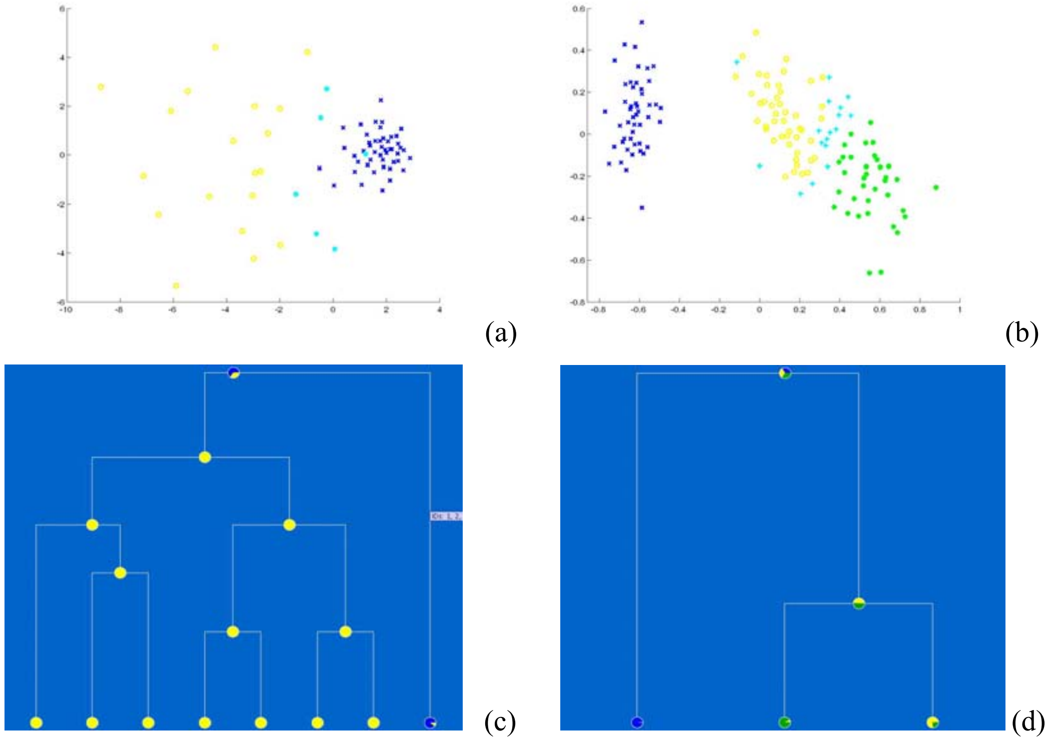


Figure 5.8 Scatter plot of the (a) Lekuemia dataset, (b) Iris dataset and partition dendograms of (c) Lekuemia and (d) Iris datasets using a threshold of 2.5 and 0.25 respectively. The scatter plots have been obtained using Matlab 6.5 with Sammon Mapping implementation of SOM Toolbox [143]. Here blue, green and yellow colors represent the various clusters and the cyan color shows the misclassified data.

It can be found in figure 5.7(b) that our validity index OAI has maximum values for a number of thresholds greater or equal to 2.5. The reason is evident from figure 5.8 (c) where the left side tree is further divided into 7 sub cluster for a threshold of 2.5 of the leukemia dataset and all these seven sub-clusters are actually partitions of the same group. When the threshold is increased these sub-clusters get merged until we left with only two clusters. When it is the situation, it is suggested that the smallest threshold should be selected for which the index value is maximum. This will although contain more clusters than actually there, yet there will be no singletons in the obtained partition.

5.5.4. Experiment 4

In this experiment genetic algorithms have been utilized for the clustering of chemical structures. Where each compound is represented as a chromosome of 1052 bits as each compound is described by the BCI 1052 bit dictionary. The best chromosomes are obtained at the end of the process applying the necessary genetic operations like cross over, mutations and selection of the fit individual.

Methodology:

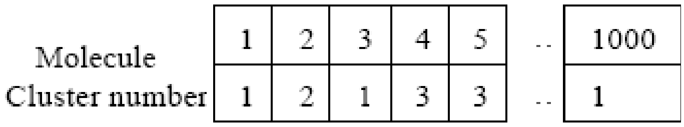
A population with *C* chromosomes was randomly generated with every chromosome representing an *n*-member subsets where each *i* entry of the *n* genes representing the assignment of the cluster *i*-th compound. The chromosomes are scored by the mean inter-cluster molecular dissimilarity (MIMDS). The equation for MIMDS is defined as follows:

$$\mathbf{MIMDS} = \mathbf{1} - \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} \mathbf{T_{ij}}}{\mathbf{n^2}} \qquad 5.20$$

where *Tij* represents the Tanimoto coefficient between cluster centroids and *n* is the number of centroids [144].

One point crossover for every chromosome {*v1, ..., vC*}*,* is applied with probability *Pc*. Another chromosome {*v'1, ..., v'C*} will be chosen randomly by selecting a random number *i* between 1 and *P*. Chromosome *i* will be chosen as the other candidate for crossover. A random integer *J* between 1 and *C* will be generated. Both chromosomes will be partitioned to two parts at position *J* and part {*vJ+1, ..., vC*} will be switched between them. Mutation is done on every element with a small probability *Pm*, a random number between 0 and 1. From *vC*, one from *n* candidates is chosen randomly to replace the element. Chromosomes with the highest fitness function will be copied into the next generation. The process will continue until a fixed number of iterations have been executed.

The GA is initialized with a population of 50 chromosomes that represented 50 different randomly chosen subsets. Each chromosome represents a molecule, whilst each allele represents the cluster number where the respective molecule is allocated. Genes with the same allele means the molecules represented contains in the same cluster. The representation of a chromosome is depicted in Figure 5.9.



Molecule	1	2	3	4	5	..	1000
Cluster number	1	2	1	3	3	..	1

Figure 5.9. Representation of Chromosomes for GA Clustering

The mutation and crossover rate is 0.09 and 0.6 respectively. Mutation involves changing the cluster number of a molecule to a new randomly chosen cluster. One point crossover exchanges sets of molecules between different clusters.

For the optimization of Ward's using GA, the GA is used to improve the clusters produced by Ward's clustering. The initial population will contain a chromosome representing the clusters produced by Ward's, and all chromosomes use the same representation as has been previously explained. The same crossover and mutation rates as before are used.

For the evaluation, the active cluster subset method was used to compare the results of all three methods namely the Wards, GA and optimized GA. The mean inter cluster molecular dissimilarity (MIMD) was also used for comparison.

Results and Discussion:

Figure 5.10 shows comparison graph based on the proportion of actives structures between Ward's clustering, GA-based clustering and the combination of Ward's and GA.
The x-axis represents the number of structure in active clusters while y-axis refers to percentage of active structures in active subsets. Experimental result based on separation active and inactive as depicted in the graph shows that clusters produced by using Ward's clustering gives better result compared to clusters produces by using GA technique and the combination of Ward's and GA. The GA-based clustering gives very poor results compared to all methods. This result also has been agreed by the findings of Brown and Martin [15], and Borosy et al. [145] who proved that Ward's showed the best distribution of active structures for chemical dataset. The reason of Ward's superiority over the GA-based clustering could be that by increase structural dissimilarity between clusters, more actives end up being clustered

together with inactives. Figure 5.11 shows the comparison graph based on the mean inter-cluster molecular dissimilarity (MIMDS) measure using Ward’s, GA and the combination of

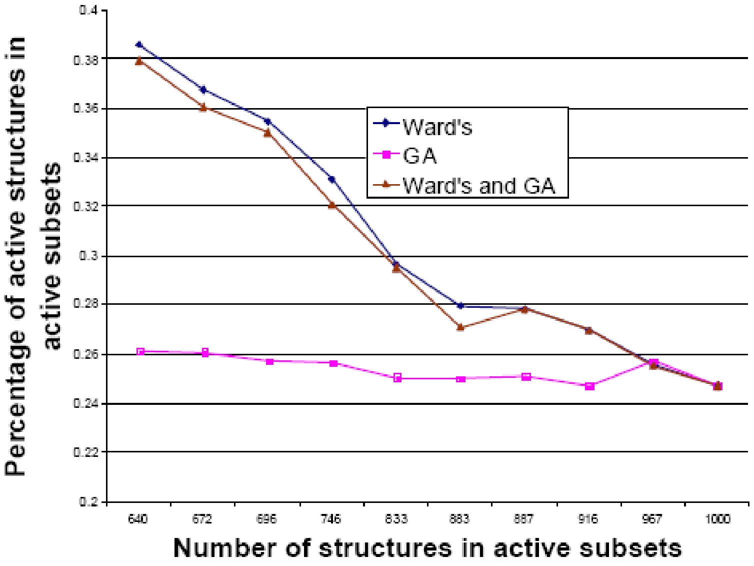


Figure 5.10. Performance of GA, Wards and optimized GA based on Proportion of actives in active cluster subset

Ward’s and GA. The x-axis plots the number of clusters, whilst the y-axis plots the value for mean intra-cluster molecular similarity. The graph is plotted for every 10 clusters to 100 clusters. The result based on MIMDS measure shows that clustering chemical database by combining Ward’s and GA based on MIMDS fitness function produces the best result compared to Ward’s and GA. The improvement in terms of inter-cluster dissimilarity shows the ability of GA technique to optimize clusters that have produced by Ward’s clustering. The GA based clustering again gives the worst result. Although MIMDS is used as the fitness function, reason for the poor performance could be because the solution is trapped in local optima.

Figure 5.12 shows the comparison graph based on the mean intra-cluster molecular similarity measure (MIMS) between Ward’s, GA and the combination of Ward’s and GA. The evaluation is done for 34 clusters determined the cut-off point suggested by Mojena’s stopping rule [146]. The result from this measure shows that clusters produced by using Ward’s clustering have molecules that are slightly more similar to one another compared to the clusters produced by the Ward’s and GA combination approach. Again, the GA-based clustering produces much worse results compared to the other two approaches. The result shows that the Ward’s clustering has higher potential to cluster together structurally similar

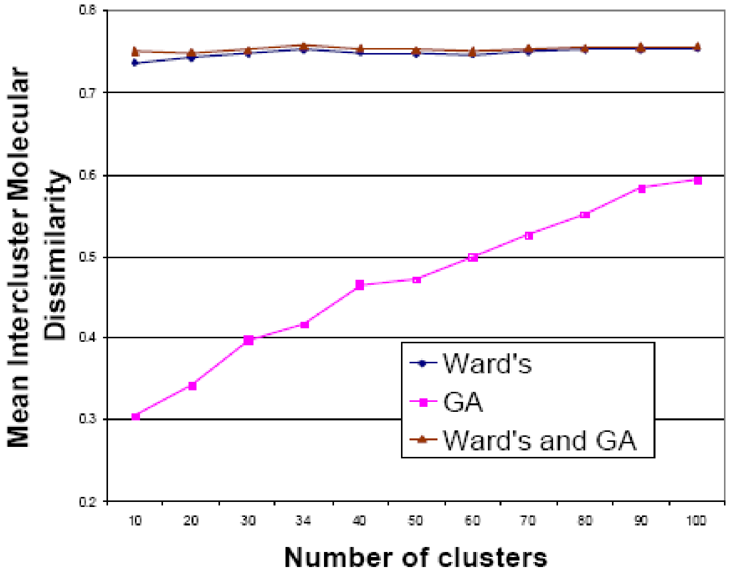


Figure 5.11. Performance of GA, optimized GA and Wards based on MIMD

structures than the other two methods. The results also show that by trying to increase inter-cluster dissimilarity, molecules in the same cluster become dissimilar instead of more similar to one another.

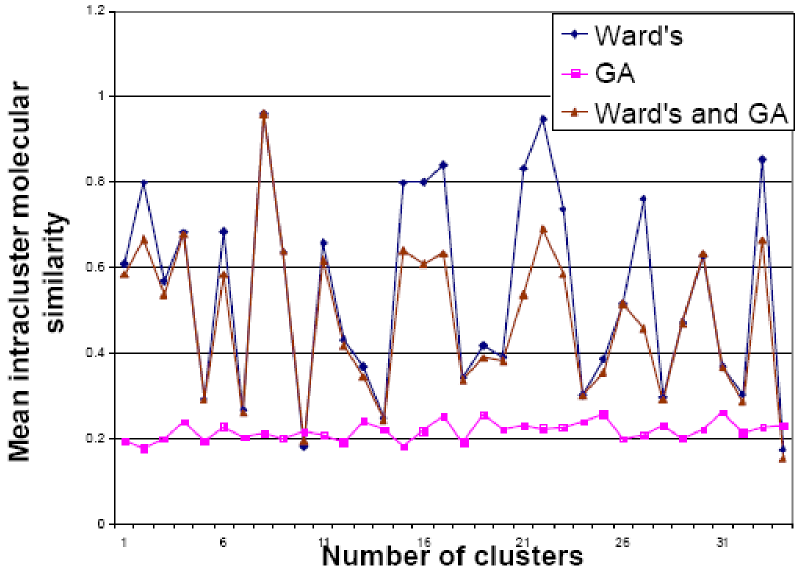


Figure 5.12. Performance of Wards, GA and optimized GA using MIMS

Based on all the three analysis methods used in this experiment confirm the weakness of GA clustering as compared to the Ward's clustering results. The optimized Genetic based clustering which takes advantage of the Ward's clustering is however very close to the Wards clustering it.

5.6. Summary

The objective of the research discussed in this chapter was to develop hybrid clustering methods for the clustering of biologically active structures utilizing the benefits of different clustering approaches. The applications of fuzzy SOM neural network and a fuzzy hierarchical approach produced excellent results in comparison to the already existing methods. It has been shown that on one hand the fuzzy SOM method is more efficient than the standard hierarchical methods and on the other hand the fuzzy hierarchical method has an inherent and simple way to accommodate the overlapping activities in multiple clusters. A new fuzzy neural network has been introduced here and its effectiveness for small datasets has been shown. The algorithm is not sensitive to the noise and presence of outliers in the data. An improved and composite validity measure for the fuzzy hierarchical clustering have also been introduced which has the ability to find the best hierarchy automatically without any need of the visual interaction. The effectiveness of the genetic algorithms could not been established due to its poor performance for the clustering of chemical structures.

Chapter 6

Discussions and Conclusions

This report discusses the research conducted on the analysis of chemical structure databases for the mining of useful knowledge about the biological activities that can be used as drugs, using fuzzy, neural and combination of these and other classical approaches like hierarchical clustering and classification. The main objective of the research was to find or develop performance wise the most appropriate methods for the analysis of drug like databases into their activities using 2-D descriptors such as topological indices and 2-D screens.

6.1. Research Objectives and their Achievements

The discipline of chemoinformatics which deals with the application of computers in chemistry in general is a new area of research. The computational clustering and classification are important processes and are used in all branches of social and scientific knowledge and thus the chemoinformatics is no exception where it is used for the extraction of chemical knowledge from large chemical databases containing millions of compounds. The present work focuses on the clustering of chemical structures that are important in medicinal chemistry such as drug design and discovery. The traditional clustering and classification methods are in use in virtual screening, lead discovery and lead optimization, molecular structure searching and similarity based molecular sub-set selection.

The performance of these classical methods is not very good and there is a room for research to evaluate the recently developed methods for chemical structures, and to develop and integrate new and existing methods such that the performance of analysis is enhanced. In this work neural network, fuzzy based clustering, genetic algorithms besides traditional methods have been evaluated and based on their merits and demerits, a few strategies have been developed that posses the promise for better performance.

6.2. Neural Networks and Chemical Structures

A number of neural networks such as the Kohonen Self-organizing Maps, Neural Gas and Enhanced Neural Gas have been evaluated for the clustering of chemical structures. Although, a number of other works show the effectiveness of Kohonen SOM for binary clustering into drug and non drug classes [25, 27], here their performance was found to be very weak as compared to traditional methods already in use in the industry. Many of the nodes in the resultant map have found to be empty and many a time as we increased the number of nodes (i.e. clusters) the number of dead nodes increased and thus no performance improvement even with increasing number of clusters. The results of the neural gas and enhanced neural gas are even poorer than the kohonen SOM despite their ability to work fairly well even in presence of noise and outliers as reported by other researchers [90, 147] in application like image segmentation. The performance of these neural networks has also been evaluated on the BCI fingerprints. For this purpose, the fingerprints dimensionality has been reduced to 105 variables using a grouping strategy of 8, 10 and 12 bits as compared to its original dimension of 1052 highly correlated variables. But the performance of the algorithms could not show any improvement all together.

In this work four supervised methods feed forward neural network, radial basis functions network, support vector machines and rough sets methods have been evaluated for the classification and recognition of a number of inhibitors as well as for the binary classification of NCI Aids dataset into AIDs active and inactive. A number of previous works show the effectiveness of support vector machines in classification of compound structures into drug and non drug structures [26, 75], and our results for the binary classification of AIDs dataset show the same trend. The support vector machines show good results with more than 98% of accuracy even when the smaller training set is used for the AIDs dataset, but was unable to show the same trend for the inhibitors dataset. In contrast the feed forward neural network shows comparable performance for the classification of AIDs dataset. The performance of rough set classifier was very poor only 45% at the most. Both the neural networks, the feed forward and radial basis functions were good in classifying the inhibitor datasets. The feed forward neural network show a recognition accuracy of 85% when trained with 50% of the dataset and similarly the radial basis functions show an accuracy of 84% when trained with lesser number of training examples, but the performance of SVMs was less than 75%. Thus it can be argued that neural networks are classifiers for the chemical structures than the SVMs and rough set classifier.

6.3. Fuzzy Clustering

In this work two clustering methods based on fuzzy logic known as the fuzzy c-means and fuzzy Gustafson – Kessel (GK) clustering algorithms have been used to cluster a small dataset composed of multiple activities. The results have been compared with the existing hierarchical methods like Ward's and group average. The results of both the clustering methods especially the fuzzy c-means are slightly better than the existing methods. The results of GK method were not as good as that of the fuzzy c-means. This can be attributed to the methods tendency to form ellipsoidal clusters instead of spherical clusters as the dataset used contained only spherical clusters. Another problem with the GK method is its dependence on the fuzzy covariance matrix, and some time the covariance matrix can not be computed exactly, in such a case an estimator is used and thus the results become dependant on how good the covariance matrix estimator is.

6.4. Hybrid Clustering Methods

It has been found that the hybrid methods, the combination of fuzzy and neural, and combination of fuzzy and hierarchical posses stronger promise. In this part of the work almost three hybrid methods have been developed for the clustering of compound structures. The fuzzy kohonen neural network have been shown to have better performance than the simple neural networks like kohonen self organizing maps, neural gas and enhanced neural gas. The fuzzy kohonen network was evaluated here on the same dataset of multiple activities and the results show and a good performance in terms of accuracy and there were fewer dead nodes in the map even with the increasing number of nodes. Another fuzzy based neural network have been developed which we call here as the robust fuzzy neural network due its insensitivity to the presence of noise and outliers in the dataset. The algorithm is independent of any heuristics based parameter selection and employees the fuzzy membership functions of the fuzzy c – means method to find the appropriate parametric values. A number of artificial datasets have been used to evaluate the performance the algorithm and it has been found that the algorithm results in good clusters even in presence of noise and outliers.

It has been found that most of the molecular structures show overlapping activities that is they can be used as drugs for more than one target disease. So, it is very unrealistic to group such compounds under only one category of biological activities in a clustering process. Most of the classical clustering methods are based on the assumption that the structures (objects) in the datasets exhibit only one activity and in such a case the clustering process is not valid. In order to prove this hypothesis a new fuzzy hierarchical algorithm has been developed and a new strategy has been applied for the clustering of the chemical structures. For this purpose a large dataset have been designed containing structures exhibiting one or dual biological activities and comprising a total of 12 biological activities. The results show that if the compounds are allowed to go into more than one clusters the clustering performance improve. The structures have been allowed to go into both of the two clusters in a binary tree like clustering if their membership functions are very similar for both of the two clusters.

A new validity measure for the fuzzy hierarchical clustering method has been developed in this work. The results of the validity measure are better in predicting the best hierarchy level than some of the standard methods.

6.5. Conclusion

The main focus of this research was the application of soft computing methods for the clustering of biologically active chemical structures into their biological activities. This work also evaluated a number of soft computing methods for the purpose of classification. The report has shown that the standard neural network methods are not very effective; however, the fuzzy and fuzzy neural methods are as powerful as the traditional chemical structure clustering methods. This research has also looked into the issue of multiply active molecular structures and has proved that if the multiple activity structures are allowed to belong to more than one cluster, the clustering results are better. A new fuzzy hierarchical clustering method has been developed and a validity measure for the evaluation of the suitable hierarchy is also developed.

References

[1] "Pharmaceutical Industry Profile Report 2006," Pharmaceutical Research and Manufacturers of America, Washington, DC 2006.

[2] P. Willett, "Similarity-based virtual screening using 2D fingerprints," *Drug Discovery Today*, vol. 11, pp. 1046-1053, 2006.

[3] G. M. Maggiora, R. W. Johnson, and M. A. Dower, *Concepts and applications of molecular similarity*. New York: John Willey and sons, 1990.

[4] T. W. Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, pp. 1857 – 1874, 2005.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: a review," *ACM computing surveys*, vol. 31, 1999.

[6] J. Xu and A. Hagler, "Review: Chemoinformatics and Drug Discovery," *Molecules*, vol. 7, pp. 566-600, 2002.

[7] W. Fisanick, K. P. Cross, and A. Rusinko, "A similarity search on CAS Registry Substances. 1. Global molecular property and generic atom triangle geometric searching," *Journal of Chemical Information and Computer Science*, vol. 32, pp. 664-674, 1992.

[8] R. C. Tryon, *Cluster Analysis*: MI: Edwards Brothers, 1939.

[9] F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis: methods for classification, data analysis, and image recognition*. New York: John Wiley, 1999.

[10] G. M. Downs and J. M. Barnard, "Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures," *Journal of chemical information and computer science*, vol. 32, pp. 644-649, 1992.

[11] G. M. Downs and J. M. Barnard, "Clustering methods and their uses in computational Chemistry," in *Reviews in Computational Chemistry*, vol. 18, K. B. Lipkowitz and D. B. Boyd, Eds.: John Wiley, 2002.

[12] P. Willett, "An Evaluation of Relocation Clustering Algorithms for the Automatic Classification of Chemical Structures," *Journal of chemical information and computer science*, vol. 24, pp. 29-33, 1984.

[13] P. Willett, *Similarity and Clustering in Chemical Information Systems*. Letchworth: Research Studies Press, 1987.

[14] P. Willett, V. Winterman, and D. Bawden, "Implementation of Non-Hierarchic cluster Analysis Methods in Chemical Information Systems: Selection of Compounds

for Biological Testing and Clustering of Substructure Search Output," *Journal of chemical information and computer science*, vol. 26, pp. 109-118, 1986.

[15] R. D. Brown and Y. C. Martin, "Use of structure- Activity data to compare structure based clustering methods and descriptors for use in compound selection," *Journal of chemical Information and computer science*, vol. 36, pp. 572-584, 1996.

[16] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of American Statistical Association*, vol. 58, pp. 236-244, 1963.

[17] R. A. Jarvis and E. A. Patrick, "Clustering using a similarity measure based on shared nearest neighbors," *IEEE Transaction on Computers*, vol. C-22, pp. 1025-1034, 1973.

[18] M. Feher and J. M. Schmidt, "Fuzzy Clustering as a Means of Selecting Representative Conformers and Molecular Alignments " *Journal of chemical Information and computer science*, vol. 43, pp. 810-818, 2003.

[19] T.-H. Lin, G.-M. Wang, and Y.-H. Hsu, "Classification of Some Active HIV-1 Protease Inhibitors and Their Inactive Analogues Using Some Uncorrelated Three-Dimensional Molecular Descriptors and a Fuzzy c-Means Algorithm," *Journal of chemical Information and computer science*, vol. 42 pp. 1490-1504, 2002.

[20] J. R. Mansfield, M. G. Sowa, G. B. Scarth, R. L. Somorjai, and H. H. Mantsch, "Analysis of Spectroscopic Imaging Data by Fuzzy C-Means Clustering," *Analytical Chemistry*, vol. 69, pp. 3370-3374, 1997.

[21] M. Misra, A. Banerjee, R. N. Dave, and C. A. Venanzi, "Novel Feature Extraction Technique for Fuzzy Relational Clustering of a Flexible Dopamine Reuptake Inhibitor," *Journal of chemical information and computer Sciences*, vol. 45, pp. 610-623, 2005.

[22] H. F. Pop and C. Sarbu, "The Fuzzy Hierarchical Cross-Clustering Algorithm. Improvements and Comparative Study," *Journal of chemical Information and computer science*, vol. 37, pp. 510-516, 1997.

[23] J. D. Holliday, S. L. Rodgers, and P. Willet, "Clustering Files of chemical Structures Using the Fuzzy k-means Clustering Method," *Journal of chemical Information and computer science*, vol. 44, pp. 894-902, 2004.

[24] D. Plewczynski, S. A. H. Spieser, and U. Koch, "Assessing Different Classification Methods for Virtual Screening," *Journal of chemical Information and computer science*, vol. 46, pp. 1098-1106, 2006.

[25] P. Bernard, A. Golbraikh, D. Kireev, J. R. Chretien, and N. Rozhkova, "Comparison of chemical databases: Analysis of molecular diversity with self-organizing maps (SOM)," *Analusis, EDB Sciences, Wiley-VCH* vol. 26, pp. 333-341, 1998.

[26] V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V. Pletnev, "Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions," *Journal of chemical Information and computer science*, vol. 43, pp. 2048-2056, 2003.

[27] Z. R. Yang and K.-C. Chou, "Mining Biological Data Using Self-Organizing Map," *Journal of chemical Information and computer science*, vol. 43, pp. 1748-1753, 2003.

[28] "Barnard MAKEBITS version 7.0 Chemical Information Fingerprint Software Documentation," Barnard Chemical Information Ltd., 1997, pp. 1-5.

[29] "MDL's Drug Data Report," Elsevier MDL. http://www.mdli.com/products/knowledge/drug_data_report/index.jsp

[30] "AIDs Database," National Cancer Institute. http://dtp.nci.nih.gov

[31] "Investigational Drug Database (IDDb)," Thomson Scientific. http://scientific.thomson.com/products/iddb

[32] A. B. Richon, "A Scrolling History of Computational Chemistry," Network Science. http://www.netsci.org/Science/Compchem/feature17b.html

[33] P. Willet, "Nomenclature processing and the interconversion of chemical structure representations," presented at the CNA (UK) seminar on chemical Structure Searching of the published literature, chemical notation association (UK), London, 1980.

[34] V. Stouw, P. M. Elliott, and A. C. Isenberg, "Automated conversion of chemical substance names to atom-bond connection tables," *Journal of Chemical documentation*, vol. 14, pp. 185-193, 1974.

[35] P. N. Craig and H. M. Ebert, "Elevan years of structure searching using the SKF (Smith, Kline,and French) fragment codes," *Journal of chemical documentation*, vol. 9, pp. 141-146, 1969.

[36] W. J. Wiswesser, *A line formula chemical Notation*. New York: Crowell co., 1954.

[37] J. M. Barnard, C. J. Jochum, and S. M. Welford, "ROSDAL: A universal structure/substructure representation for PC-host communication, in chemical structure information systems: Interface communication and standards," presented at ACS Symposium series No. 400, Washington, DC, 1989.

[38] S. Ash, M. A. Cline, R. W. Homer, T. Hurst, and G. B. J. Smith, "SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation," *Journal of chemical Information and computer science*, vol. 37, pp. 71-79, 1997.

[39] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and Encoding Rules," *journal of chemical information and computer Sciences*, vol. 28, pp. 31-36, 1988.

[40] "Daylight." http://www.daylight.com

[41] "CTfile Formats." CA: MDL Information Systems. http://www.mdli.com/downloads

[42] G. W. Adamson and J. A. Bush, "A comparison of some similarity and dissimilarity measures in the classification of chemical structures," *Journal of chemical Information and computer science*, vol. 15, pp. 55-58, 1975.

[43] L. Hodes, "Selection of descriptors according to discrimination and redundancy. Applications to chemical structure searching," *Journal of chemical Information and computer science*, vol. 16, pp. 88-93, 1976.

[44] P. Willett, "A screen set generation algorithm," *Journal of chemical Information and computer science*, vol. 19, pp. 159-162, 1979.

[45] P. G. Dittmar, N. A. Farmer, W. Fisanick, R. C. Haines, and J. Mockus, "The CAS online search system. 1. General system design and selection, generation and use of search screens," *Journal of chemical Information and computer science*, vol. 23, pp. 93-102, 1983.

[46] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL keys for use in drug discovery," *Journal of Chemical Information and Computer Science*, vol. 42, pp. 1273-1280, 2002.

[47] H. Wiener, "Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons " *Journal of American Chemical Society*, vol. 69, pp. 2636-2638, 1947.

[48] H. Wiener, "Relation of the Physical Properties of the Isomeric Alkanes to Molecular Structure. Surface Tension, Specific Dispersion, and Critical Solution Temperature in Aniline " *Journal of Physical chemistry*, vol. 52, pp. 1082-1089, 1948.

[49] H. Wiener, "Structural Determination of Paraffin Boiling Points " *Journal of American Chemical Society*, vol. 69, pp. 17-20, 1947.

[50] D. Plavsic, S. Nikolic, N. Trinajstic, and Z. Mihalic, "On the Harary index for the characterization of chemical graphs.," *Journal of Mathematical Chemistry*, vol. 12, pp. 235-250, 1993.

[51] W. R. Mueller, K. Szymanski, J. V. Knop, and N. Trinajstic, "Molecular topological index " *Journal of chemical Information and computer science*, vol. 30, pp. 160-163, 1990.

[52] H. P. Schultz, "Topological organic chemistry. 1. Graph theory and topological indices of alkanes " *journal of chemical Information and computer science*, vol. 29, pp. 227-228, 1989.

[53] A. T. Balaban, "Highly Discriminating Distance-Based Topological Index," *Chemical Physics Letters*, vol. 89, pp. 399-404, 1982.

[54] M. Randic, "Characterization of molecular branching " *Journal of American Chemical Society*, vol. 97, pp. 6609-6615, 1975.

[55] I. Gutman, B. Ruscic, N. Trinajstic, and J. C. F. Wilcox, "Graph theory and molecular orbitals. XII. Acyclic polyenes," *Journal of Chemical Physics*, vol. 62, pp. 3399-3405, 1975.

[56] I. Gutman and N. Trinajstic, "Graph theory and molecular orbitals. Total ¼-electron energy of alternant hydrocarbons," *Chemical Physics Letters*, vol. 17, pp. 535-538, 1972.

[57] H. Yuan and C. Cao, "Topological Indices Based on Vertex, Edge, Ring, and Distance: Application to Various Physicochemical Properties of Diverse Hydrocarbons," *Journal of chemical Information and computer science*, vol. 43, pp. 501-512, 2003.

[58] M. Randic, A. T. Balaban, and S. C. Basak, "On Structural Interpretation of Several Distance Related Topological Indices," *Journal of chemical Information and computer science*, vol. 41, pp. 593-601, 2001.

[59] A. Mercader, E. A. Castro, and A. A. Toropov, "Maximum Topological Distances Based Indices as Molecular Descriptors for QSPR. 4. Modeling the Enthalpy of Formation of Hydrocarbons from Elements," *International Journal of Molecular Science*, vol. 2, pp. 121-132, 2001.

[60] A. Thakur, M. Thakur, N. Kakani, A. Joshi, S. Thakur, and A. Gupta, "Application of topological and physicochemical descriptors: QSAR study of phenylamino-acridine derivatives," *ARKIVOC*, vol. xiv, 2004.

[61] E. V. Konstantinova and M. V. Vidyuk, "Discriminating Tests of Information and Topological Indices. Animals and Trees," *Journal of chemical Information and computer science*, vol. 43, pp. 1860-1871, 2003.

[62] "Chem-X Software, Oxford Molecular Group, Medawar Centre, Oxford Science Park, Oxford, OX4 4GA, U.K.."

[63] H. Matter and T. Potter, "Comparing 3D Pharmacophore Triplets and 2D Fingerprints for selecting diverse compound subsets," *Journal of chemical Information and computer science*, vol. 39, pp. 1211-1225, 1999.

[64] B. S. Everitt, *Cluster Analysis*. London: Edward Arnold, 1993.

[65] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, 3rd ed. San Francisco: WH Freeman, 1973.

[66] A. D. Gordon, *Classification*, 2nd ed. London: Chapman and Hall, 1999.

[67] I. Jolife, *Principal component analysis*. New York: Springer-Verlag, 1986.

[68] J. Holliday, N. Salim, and P. Willet, "Analysis and display of the size dependence of chemical similarity coefficients," *Journal of chemical Information and computer science*, vol. 43, pp. 819-828, 2003.

[69] N. Salim, "Analysis and Comparison of Molecular Similarity Measures," in *Information Studies*, vol. Ph.D. thesis. university of Sheffield: Sheffield, 2002.

[70] H. T. Clifford and W. Stephenson, *An Introduction to Numerical classification*. New York: Academic Press, 1975.

[71] B. S. Duran and P. L. Odell, *Cluster analysis: A survey*. Berlin: Springer Verlag, 1974.

[72] K. Florek, J. Lukaszweiz, J. Perkal, H. steinhas, and S. Zubrzchi, "Sur la liason et la division des points d'un ensemble fini," *Colloquium Mathematicum*, vol. 2, pp. 282-285, 1951.

[73] V. Schnecke and J. Bostrom, "computational chemistry driven decision making in lead generation," in *Drug Discovery Today*, vol. 11, 2006.

[74] J. W. Godden, L. Xue, and J. Bajorath, "classification of biologically active compounds by median partitioning," *Journal of chemical Information and computer science*, vol. 42, pp. 1263-1269, 2002.

[75] E. Byvatove, U. Fechner, J. Sadowski, and G. Schneider, "comparison of support vector machines and artificial neural network systems for the drug/nondrug

classification," *Journal of chemical Information and computer science*, vol. 43, pp. 1882-1889, 2003.

[76] Y. H. Wang, Y. Li, S. L. Yang, and L. Yang, "classification of Substrates and Inhibitors of P-Glycoprotein Using Unsupervised Machine Learning Approach," *Journal of chemical Information and computer science*, vol. 45, pp. 750-757, 2005.

[77] L. H. Tsoukalas and R. E. Uhrig, *Fuzzy and Neural Approaches in Engineering*: John Wiley, 1997.

[78] J. S. R. Jang, C. T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A computational approach to learning and machine intelligence*: Prentice Hall, 1997.

[79] T. Kohonen, "The self organizing map," *IEEE Proc.*, vol. 78, 1990.

[80] G. Carpenter and S. Grossberg, "ART3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures," *Neural Networks*, vol. 3, pp. 129 –152, 1990.

[81] T. Eltof and R. J. P. deFigueiredo, "A new neural network for cluster detection and labeling," *IEEE Trans. on Neural Networks*, vol. 9, 1998.

[82] T. T. Martinez, S. G. Berkovich, and K. J. Schulten, "Neural gas network for vector quantization and its application to time series prediction," *IEEE Trans. On Neural Networks*, vol. 4, 1993.

[83] M. Fontana, N. A. Borghese, and S. Ferrari, "Image reconstruction using improved neural-gas," presented at Workshop Italiano Reti Neurali, 1995.

[84] T. Villmann, B. Hammer, and M. Strickert, "Supervised neural gas for learning vector quantization," presented at Fifth German Workshop on Artificial Life, 2002.

[85] F. Camastra and A. Vinciarelli, "combining neural gas and vector quantization for cursive character recognition," *Journal of Neurocomputing*, vol. 51, pp. 147-159, 2003.

[86] B. L. Zhang, M. Y. Fu, and H. Yan, "Hand written character verification based on neural gas based vector quantization," *IEEE proc.*, 1998.

[87] Z. Cselenyi, "Mapping the dimensionality, density and topology of data: the growing adaptive neural gas," *computer methods and programs in biomedicine*, vol. 78, pp. 141-156, 2005.

[88] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: Fuzzy c-means algorithm," *Computers and Geoscience*, 1984.

[89] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," presented at IEEE CDC, San Diego, CA, USA, 1978.

[90] A. K. Qin and P. N. Suganthan, "Enhanced neural gas network for prototype-based clustering," *Pattern recognition*, vol. 38, pp. 1275-1288, 2005.

[91] D. W. Ruck, S. K. Rogers, K. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to an optimal Bayes estimator," *IEEE Transactions on Neural Networks*, vol. 1, pp. 296--298, 1990.

[92] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Data Processing*, vol. 1, D. Rumelhart and J. McClelland, Eds. Cambridge: The M.I.T. Press, 1986, pp. 318--362.

[93] J. Z. Shah and N. Salim, "FCM and G-K clustering of chemical dataset using topological indices," presented at First International Symposium on Bio-Inspired Computing, Johor Bahru, Malaysia, 2005.

[94] P. Arabshahi, J. J. Choi, R. J. Marks, and T. P. Caudell, "Fuzzy parameter adaptation in optimization: some neural net training examples," *IEEE Computational Science and Engineering*, vol. 3, pp. 57 - 65, 1996.

[95] J. A. Leonard and M. A. Kramer, "Radial basis function networks for classifying process faults," *IEEE Control Systems Magazine*, vol. 11, pp. 31 - 38, 1991.

[96] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," *Advances in Neural Information Processing Systems*, vol. 9, pp. 281–287, 1996.

[97] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[98] A. Ohrn, "ROSETTA Technical Reference Manual," A. Ohrn, Ed. Trondheim, Norway. : Norwegian University of Science and Technology, 2001.

[99] Z. Pawlak, "Why Rough Set?," presented at Fifth IEEE International Conference on Fuzzy Systems, 1996.

[100] "Dragon," melano chemoinformatics. http://www.talete.mi.it

[101] "ACE Inhibitors." http://www.healthyhearts.com/medications.htm

[102] E. Amsallem, C. Kasparian, G. Haddour, J. Boissel, and P. Nony, "Phosphodiesterase III inhibitors for heart failure," *The Cochrane Database of Systematic Reviews*, 2006.

[103] A. Triaa, O. Hiortb, and G. H. G. Sinneckera, "Steroid 5-Reductase 1 Polymorphisms and Testosterone/Dihydrotestosterone Ratio in Male Patients with Hypospadias," *Hormone Research*, vol. 61, pp. 180-183, 2004.

[104] L. A. Zadeh, "Fuzzy Sets," *Information Control*, vol. 8, pp. 338-353, 1965.

[105] B. Kosko, *Fuzzy Thinking: The New Science of Fuzzy Logic*. London: HarperCollins, 1993.

[106] M. Dummett, "Wang's Paradox," *Journal of Synthese*, vol. 30, pp. 301-324, 1975.

[107] J. C. Bezdek and S. K. Pal, *The Fuzzy Models For Pattern Recognition*. New York: IEEE Inc., 1992.

[108] H. J. Zimmerman, *Fuzzy Set Theory - and Its Applications*. Boston: Kluwer Academic Publishers, 1996.

[109] R. Krishnapurum, O. Nasraoui, and H. Frigui, "New Fuzzy shell clustering algorithms for boundary detection and pattern recognition," presented at Intelligent Robots and Computer Vision X, 1991.

[110] R. N. Dave, "Fuzzy shell-clustering and applications to circle detection in digital images," *International Journal of General Systems*, vol. 16, pp. 343-355, 1990.

[111] R. N. Dave and K. Bhaswan, "Adaptive Fuzzy c-shells Clustering and Detection of Ellipses," *IEEE Transaction on Neural Networks*, vol. 3, pp. 643-662, 1992.

[112] R. Krishnapurum, O. Nasraoui, and H. Frigui, "The Fuzzy C-shells algorithm: A new approach," *IEEE Transaction on Neural Networks*, vol. 3, pp. 663-671, 1992.

[113] J. MacCuish, C. Nicolaou, and N. E. MacCuish, "Ties in Proximity and Clustering Compounds," *Journal of Chemical Information and Computer Sciences*, vol. 41, pp. 134-146, 2001.

[114] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Transaction on Fuzzy Systems*, vol. 1, pp. 98-110, 1993.

[115] M. Barni, V. Cappellini, and A. Mecocci, "Comments on “A possibilistic approach to clustering”," *IEEE Transaction on Fuzzy Systems*, vol. 4, pp. 393-396, 1996.

[116] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A Possibilistic Fuzzy c-Means Clustering Algorithm," *IEEE Transaction on Fuzzy Systems*, vol. 13, pp. 517-530, 2005.

[117] H. Timm, C. Borgelt, C. Doring, and RudolfKruse, "An extension to possibilistic fuzzy cluster analysis," *Journal of Fuzzy sets and Systems*, vol. 147, pp. 3-16, 2004.

[118] J. M. Jolion and A. Rosenfeld, (), . "Cluster Detection in Background Noise," *Pattern Recognition*, vol. 22, pp. 603-607, 1989.

[119] Z. Y. Chi and T. Pham, "Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition," presented at World Scientific, Singapore, 1996.

[120] E. H. Ruspini, "A new approach to clustering," *Information and Control*, vol. 15, pp. 22–32, 1969.

[121] J. C. Dunn, "A fuzzy relative of the Isodata process and its use in detecting compact, well-separated clusters," *Journal of Cybernetics*, vol. 3, pp. 32–57, 1973.

[122] J. C. Bezdek, M. Windham, and R. Ehrlich, "Statistical parameters of fuzzy cluster validity functionals," *International Journal of computer and information science*, vol. 9, pp. 332-336, 1980.

[123] J. C. Neto, G. E. Meyer, and D. D. Jones, "Individual leaf extractions from young canopy images using Gustafson–Kessel clustering and a genetic algorithm," *Computers and Electronics in Agriculture*, vol. 51, pp. 66-85, 2006.

[124] R. L. Cannon, J. V. Dave, and J. C. Bezdek, "Efficient Implementation of the Fuzzy c-Means Clustering Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8 pp. 248-255, 1986.

[125] E. Rignot, R. Chellappa, and P. Dubois, "Unsupervised segmentation of polarimetric SAR data using the covariance matrix," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, pp. 697 - 705, 1992.

[126] M.-S. Yang and K.-L. Wu, "Unsupervised possibilistic clustering," *Pattern Recognition*, vol. 39, pp. 5-21, 2006.

[127] R. Babuška, P. J. Vanderveen, and U. Kaymak, "Improved covariance estimation for Gustafson-Kessel clustering," presented at IEEE International Conference on Fuzzy Systems, Honolulu, Hawaii, 2002.

[128] A. Linusson, S. Wold, and B. Nordén, "Fuzzy Clustering of 627 Alcohols, Guided by a Strategy for Cluster Analysis of Chemical Compounds for Combinatorial Chemistry," *Chemometrics and Intelligent Laboratory Systems*, vol. 44, pp. 213-227, 1998.

[129] M. Friederichs, O. Fränzle, and A. Salski, "Fuzzy Clustering of Existing Chemicals According to their Ecotoxicological properties," *Ecological Modelling*, vol. 85, pp. 27-40, 1996.

[130] X. Chen and C. H. Reynolds, "Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients," *Journal of Chemical Information and Computer Sciences*, vol. 42, pp. 1407-1414, 2002.

[131] J. C. Bezdek, E. C.-K. Tsao, and N. R. Pal, "Fuzzy Kohonen clustering networks," presented at IEEE International Conference on Fuzzy Systems, 1992.

[132] T. Huntsberger and P. Ajjimarangsee, "Parallel Self-organizing feature maps for unsupervised pattern recognition," *International journal of General Systems*, vol. 16, pp. 357-372, 1989.

[133] A. Bocker, S. Derksen, E. Schmidt, A. Teckentrup, and G. Schneider, "A Hierarchical Clustering Approach for Large Compound Libraries," *Journal of chemical Information and Modeling*, vol. 45, pp. 807-815, 2005.

[134] C.-H. Chou, M.-C. Su, and E. Lai, "A New Cluster Validity Measure for Clusters with Different Densities," *Proc. of Intelligent Systems and Control*, vol. 388, 2003.

[135] J. H. Holland, "Genetic Algorithms and the optimal allocation of trials," *SIAM Journal on Computers*, vol. 2, pp. 88-105, 1973.

[136] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, 1st ed. Reading, MA, USA: Addison-Wesley, 1989.

[137] F. Busetti, "Genetic Algorithm Overviews," 2004.
http://www.geocities.com/francorbusetti/gaweb.pdf

[138] H. Adeli and S. L. Hung, *Machine Learning-Neural Networks, Genetic Algorithm, and Fuzzy System*. New York: John Wiley & Sons, 1995.

[139] M. Mitchell, *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1996.

[140] R. A. Fisher, "The use of multiple measurements in axonomic problems," *Annual Eugenics*, vol. 7, pp. 179-188, 1936.

[141] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S.

Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science Magazine*, vol. 285, pp. 531 - 537, 1999.

[142] F. Hopner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*: John Wiley & Sons, 1999.

[143] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "SOM Toolbox for Matlab," Helsinki University of Technology, P.O. Box 5400, FIN-02015 HUT , Finland, 2000.

[144] D. B. Turner, S. M. Tyrell, and P. Willet, "Rapid Quantification of Molecular Diversity for Selective Database Acquisition," *Journal of American Chemical Society*, vol. 37, pp. 18-22, 1997.

[145] A. Borosy, F. Csizmadia, and A. Volford, "Structure Based Clustering of NCI's Anti-HIV Library," presented at First Symposium of the European Society of Combinatorial Science, Budapest, Hungary, 2001.

[146] R. Mojena, "Hierarchical grouping methods and stopping rules: an evaluation," *The Computer Journal*, vol. 20, pp. 359-363, 1977.

[147] A. K. Qin and P. N. Suganthan, "Robust growing neural gas algorithm with application in cluster analysis," *Journal of Neural Networks*, vol. 17, pp. 1135-1148, 2004.