Thesis No: CSER-M-16-03

A MULTI-OBJECTIVE GENETIC ALGORITHM WITH FUZZY RELATIONAL CLUSTERING FOR AUTOMATIC DATA CLUSTERING

By

Animesh Kundu

Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna 9203, Bangladesh
December, 2016

A Multi-Objective Genetic Algorithm with Fuzzy Relational Clustering for Automatic Data Clustering

By

Animesh Kundu
Roll No: 1207505

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in Computer Science and Engineering

Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna 9203, Bangladesh
December, 2016

Declaration

This is to certify that the thesis work entitled "A Multi-Objective Genetic Algorithm with Fuzzy Relational Clustering for Automatic Data Clustering" has been carried out by Animesh Kundu in the Department of Computer Science and Engineering (CSE), Khulna University of Engineering & Technology (KUET), Khulna, Bangladesh. The above thesis work or any part of this work has not been submitted anywhere for the award of any degree or diploma.

Signature of Supervisor

Signature of Candidate

Approval

This is to certify that the thesis work submitted by Animesh Kundu entitled "A Multi-Objective Genetic Algorithm with Fuzzy Relational Clustering for Automatic Data Clustering" has been approved by the board of examiners for the partial fulfillment of the requirements for the degree of Master of Science in Computer Science & Engineering in the Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh in December, 2016.

BOARD OF EXAMINERS

1.	Dr. Pintu Chandra Shill Associate Professor Department of Computer Science and Engineering(CSE) Khulna University of Engineering & Technology(KUET)	Chairman (Supervisor)
2.	Head Department of Computer Science and Engineering(CSE) Khulna University of Engineering & Technology(KUET)	Member
3.	Dr. M. M. A. Hashem Professor Department of Computer Science and Engineering(CSE) Khulna University of Engineering & Technology(KUET)	Member
5.	Dr. Rameswar Debnath Professor Department of Computer Science and Engineering(CSE) Khulna University(KU)	Member (External)

Acknowledgment

All the praise to the almighty God, whose blessing helped me to successfully complete this thesis work. I show significant and indescribable gratefulness to my supervisor Dr. Pintu Chandra Shill, Associate Professor, Department of Computer Science and Engineering, Khulna University of Engineering & Technology for his outstanding helpful contribution in giving suggestion and encouragement. I acknowledge his constant co-operation and proper guidance throughout the development process. He has been a great source of effective and feasible ideas, profound knowledge and all time feedback for me.

I thank all the teachers of the Department of Computer Science and Engineering who helped me providing guidelines to perform the work. I would also like to thank my friends and family for their cordial support.

Author

Abstract

A Fuzzy relational clustering algorithm (FRC) based on multi-objective non-dominated sorting genetic algorithm (NSGA-II) called FRC-NSGA-II is proposed for automatic data clustering. A given data set is spontaneously promoted into an optimal number of groups in a precise fuzzy partition through the fuzzy relational clustering algorithm, FRC. FRC operates on a similarity square matrix which is generated by comparing the pair wise similarities between data points. Multi-objective NSGA-II is employed to search for appropriate number of partitions for different cluster shapes. Moreover, two well-known cluster validity indices, compactness and separation, are optimized concurrently through multi-objective NSGA-II where compactness indicates variation between data within a cluster and separation means quantifying the separation between different clusters. Real encoding schema is used for variable length NSGA-II chromosomes representing the variable number of clusters. The simulation result on benchmark data sets exhibits that the proposed method gives promising results in the complex, overlapped, high-dimensional non-gene and gene expression data sets and it has better capability of determining well-separated, hyperspherical and overlapping clusters compared with other existing clustering algorithms.

Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

1.1 Background

Fuzzy relational clustering techniques [1,2] are mostly unsupervised data analysis methods that can be used to organize data into groups based on similarities among the individual data items. In order to obtain appropriate partitioning of complex data sets consisting of different shaped clusters, the chosen objective functions should treat compactness of clusters along with the ability to deal with overlapping clusters. In this case, multi-objective geneic algorithms provide better result since they provide flexibility in problem solving by allowing hybridization.

1.2 Problem Statement

Fuzzy clustering generally fuzzy C-means [3] requires a-priori knowledge of number of clusters. Traditional fuzzy clustering methods use only centroid information for clustering. That is why they cannot differentiate the geometric structures of clusters due to the compactness and separation measures of fuzzy partition. Moreover, an optimization with single objective may not be feasible for different cluster shapes. On the other-hand, fuzzy relational clustering problems are increasing in a number of different applications.

1.3 Motivation

Find a technique to estimate appropriate number of fuzzy clusters without prior knowledge on the number of clusters. The overlap-separation measure using an aggregation operation of fuzzy membership degrees can be accomplished to effectively deal with this limitation. A multi-objective optimization with fuzzy

relational clustering is therefore appropriate to search for fuzzy partitions in this situation.

1.4 Objectives

The objective of this thesis is to develop a fuzzy relational clustering algorithm based on multi-objective non-dominated genetic algorithm (NSGA-II) called FRC-NSGA-II to obtain appropriate partitioning of complex data sets consisting of different geometric shaped clusters.

This general objective can be divided into the following specific ones:

- Development of a new multi-objective evolutionary approach to evolve optimal data clustering without requiring prediction of the number of clusters.
- Optimization of two objectives: fuzzy J_m index and overlap-separation measure simultaneously.
- Integration of multi-objective genetic algorithm, NSGA-II [4] with fuzzy relational clustering so that NSGA-II can be used as multi-objective optimization tool and FRC algorithm [27] can be used to generate fuzzy clusters from data points.
- Application of this approach on non-gene and gene expression data sets.

1.5 Methodology

Fuzzy relational clustering algorithm (FRC) based on multi-objective non-dominated sorting genetic algorithm (NSGA-II) called FRC-NSGA-II is proposed for automatic data clustering. The FRC-NSGA-II method integrates the multi-objective optimization, compactness and separation in an optimization process to automatically estimate the number of clusters, and then partitions the whole given data set to produce the most natural clustering.

A given data set is spontaneously promoted into an optimal number of groups in a precise fuzzy partition through the fuzzy relational clustering algorithm, FRC. This FRC operates on a similarity square matrix which is generated by comparing

the pairwise similarities between data points. In the fuzzy relational clustering process, the degree of membership is assigned into data points by FRC algorithm and the data points contain more information than the hard clustering process. Moreover, fuzzy relational clustering algorithm operates on Expectation-Maximization framework.

Multi-objective NSGA-II [4] is employed to search for appropriate number of partitions for different cluster shapes. Moreover, two well-known cluster validity indices, compactness and separation, are optimized concurrently through multi-objective NSGA-II [4] where compactness indicates variation between data within a cluster and separation means quantifying the separation between different clusters. Real encoding schema is used for variable length NSGA-II [4] chromosomes representing the variable number of clusters. Therefore, the multi-objective evolutionary approach has been developed to evolve optimal data clustering without requiring prediction of the number of clusters.

The proposed approach has been tested on real-life data sets having complex, overlapped, high-dimensional non-gene and gene expression. The simulation results exhibit that the intended method gives promising results and it has better capability of determining well-separated, hyperspherical and overlapping clusters compared with other existing clustering algorithms.

1.6 Scope of the Thesis

This study focuses on optimization of two cluster validity measures: compactness and overlap-separation. This study includes Minkowski score metric to evaluate the quality of clustering solution.

1.7 Contribution of the Thesis

This thesis focuses on construction of a fuzzy relational clustering approach based on fast elitist non-dominated sorting multi-objective genetic algorithm (NSGA-II) for discovery of appropriate number of cluster where it smartly creates a finer trade-off between fuzzy compactness and fuzzy separation of the clusters for

high-dimensional complex data sets. Unlike conventional clustering approaches the proposed technique doesn't require number of clusters to be predefined. This work optimizes multiple objectives while partitioning the data set using fuzzy relational clustering method.

1.8 Organization of the Thesis

- **Chapter II** presents some of the existing prominent multi-objective data clustering schemes while focusing on their limitations. It also includes alluring benefits of data clustering by the proposed method.
- **Chapter III** presents the relevant theoretical materials. It first describes the non-dominated sorting genetic algorithm (NSGA-II) [4] along with its essential features such as fast non-dominated sorting, diversity mechanism and crowding distance assignment. It then describes the fuzzy relational clustering algorithm in detail. It also describes the objective functions used in the proposed approach. This chapter also includes detailed description of the involved genetic operators such as binary tournament selection, SBX operator and polynomial mutation. It also describes the evaluation function used for evaluating the non-dominated solutions.
- **Chapter IV** presents the proposed system and describes its working procedure in detail.
- **Chapter V** shows the experimental results of the proposed approach on both gene and non-gene data sets. Here it demonstrates the improved performance of FRC-NSGA-II in comparison with other single and multi-objective clustering approaches.
- **Chapter VI** lists the results gathered from experiments. It also includes what future researches are needed to explore for more desirable data clustering.

CHAPTER II

Literature Review

2.1 Introduction

This chapter describes some related works based on clustering along with their limitations. This chapter also presents how the limitations of the existing methods are overcomed by the proposed method.

2.2 Related Work

There are a good number of methods [5-26] have been developed for data clustering based on multi-objective evolutionary algorithm. Emin Erkan Korkmaz et. *al*. [5] proposed a clustering method on the basis of multi-objective genetic algorithm. This method uses a novel encoding scheme that uses links to identify clusters in a partition. In a single GA run it can obtain optimal partitioning by optimizing two measures: Total within cluster variation (TWCV) and the number of clusters in a partition. This method does not need to specify number of clusters beforehand. This method is able to encode the solution space in fixed-length chromosomes and explores the search space in a suitable manner. This method however fails to evolve optimal clusters where the cluster borders are not clear.

Dipankar and Paramartha [6] proposed a real coded multi-objective genetic algorithm based K-clustering which suffers from the limitation of clustering categorical features and prediction of the number of cluster.

J. Handl and J. Knowles proposed a clustering method based on the genetic algorithm [7] called MOCK. The algorithm works based on PESA-II with locus based chromosome encoding. In their method they have shown that the clustering algorithm outperforms the single-objective clustering algorithms and ensemble

techniques. However, it performs well for hyperspherical shaped or well-separated clusters but provides low performance on overlapping clusters [8]. Another disadvantage on the locus based encoding is the length of the string, which increases with the size of the data set. This imposes expensive computation when a large data set is analyzed.

VAMOSA, developed in [8], is based on multi-objective simulated annealing with center-based encoding and the newly developed point symmetry based distance from [9]. The approach has been compared with MOCK and other algorithms in artificial and real-life data sets of varying shapes, sizes or convexity. It successfully determines the appropriate number of clusters and provides overall performance better than other algorithms. However, it fails to detect a cluster having non-symmetrical shapes.

J. J. T. Valenzuela [10] proposed a multi-objective optimization environment method based on Grouping Genetic Algorithms (GGA) introduced by Emmanuel Falkenaeur [11]. This method is able to correctly identify clusters of proteins which have a functional interrelationship. This has slow convergence process due to greater computational complexity of crossover operation. Moreover it requires a greater number of generations to reach convergence.

David G. Bethelmy [12] proposed multi-objective genetic clustering algorithms MOCK and AAMOCK for aspect mining. Due to heuristic nature of MOCK, it can only approximate the true Pareto front and so cannot guarantee the overall best solution.

Rafael et. *al*. proposed a meta-heuristic method for multi-objective clustering problem [13] which is based on the tabu and scatter search methodologies. They employed cluster centers in their work and they performed experiments using [13] Bi-Heur published by Brusco and Cradit(2005).

A multi-objective clustering algorithm that are defined by exclusively extrinsic properties was developed in [14] for clustering data items. These approaches optimized two objectives simultaneously: compactness and connectivity where the length of chromosome encoding is equal to the number of data points which

makes the convergence slower due to larger search space. This technique managed complication by introducing a special mutation operator which maintained a list of L nearest neighbors for each data point.

M. H. Law et.*al*. [15] developed a multi-objective data clustering approach. The algorithm consists of a two-step process where it includes detection of clusters with diverse shapes and sizes by a set of candidate objective functions as well as their integration into the target partition. This method is based on the principle of data space partitioning, where different learning algorithms are applied to different parts of the data space because clusters in different regions can be of different shapes. The problem here is that it uses a wide range of values for the parameters involved. It also fails to perform multi-objective clustering when the "effective" regions of different clustering objectives overlap significantly.

Peter peng, et. *al*., [16] developed a multi-objective K-means genetic algorithm (MOKGA) for data clustering. They integrated the two optimization algorithms, Fast Genetic K-means Algorithm (FGKA) and the Niched Pareto Genetic Algorithm. This clustering approach was developed to deliver a pareto optimal clustering solution set for microarray and other data sets. But this method supports crisp clustering only and is not capable of identifying outliers.

Dipankar and Paramartha [17] proposed hybrid elitist real coded multi-objective genetic algorithm based K-clustering method for fuzzy clustering of categorical data where K represents the number of clusters known apriori. It can work only with categorical features and cannot decide optimal value of k.

K.P.Malarkodi and S.Punithavathy [18] proposed a method named Fuzzy based Evolutionary Multi objective Clustering for Overlapping Clusters (FEMCOC) for identification of overlapping clusters on complex data sets. Hence, Genetic Algorithm with variable length chromosome and local search, and a Fuzzy GA with variable length chromosome and local search are coupled with the existing Evolutionary Multi objective Clustering approach were used. But this approach does not work on well defined fitness functions.

Kaushik Suresh et. *al*. [19] applied the differential evolution (DE) algorithm as the optimization algorithm to the task of automatic fuzzy clustering in a Multi-objective Optimization (MO) framework. A real-coded representation of the search variables, accommodating variable number of cluster centers, is used for DE. It produces final clustering solutions satisfying multiple objective functions and simulation results on various data sets proves its better performance as compared to NSGA-II and MOCK. This method cannot handle discrete chromosome representation and hence has the restriction to use cluster centroids.

M. Anusha and Sathiaseelan [20] used Evolutionary Clustering Multi-objective Optimization Algorithm (ECMO) which was the extended work of NL-MOGA [21] for analyzing diabetes disease data sets. ECMO generates pareto optimal solutions for selected objectives with higher accuracy in short time. However, ECMO is not suitable for variety of complex data sets.

Kartick et. *al*. [22] proposed a new multi-objective approach MOSCFRA for simultaneous clustering and gene ranking. This approach uses a new encoding technique and uses two performance measures, CP Index and R Index for this purpose. This approach still needs the number of clusters to be predefined and this approach fails to detect various cluster shapes.

Anirban, Ujjwal and Sanghamitra [23] proposed a multi-objective genetic algorithm-based approach for fuzzy clustering of categorical data that encodes the cluster modes. This approach simultaneously optimizes fuzzy compactness and fuzzy separation of the clusters. They also proved a novel method for obtaining the final clustering solution from the set of resultant pareto-optimal solutions. This method also suffers from trouble of specifying the number of cluster.

Jun Du, Emin, Reda and Ken [24] presented a linked-list based encoding scheme for multiple objectives based genetic algorithm (GA) to identify clusters in a partition. A new scheme is proposed for encoding clustering solutions into chromosomes. The proposed representation forms a linked-list structure for objects in the same cluster. This method efficiently explores the solution space but possesses the problem of early mentioning of optimal number of clusters.

Sanghamitra, Anirban and Ujjwal [25] proposed a two-stage clustering algorithm (SiMM-TS) for clustering gene expression data using the idea of points having significant membership to multiple classes (SiMM points). A VGA-based clustering scheme and a MOGA-based clustering technique is utilized in the process. This method automatically determines the number of cluster but cannot perform well for overlapped clusters.

In [26] Tansel Özyer and Reda performed the clustering task of high dimensional data by multi-objective genetic algorithm integrated with divide and conquer strategy. Their work partitions a large data set into subsets of manageable sizes and then clusters the partitions separately in parallel. This suits well for interactive online clustering and facilitates for incremental clustering because chunks of instances are clustered as stand alone sets, and then the results are merged with existing clusters. This approach is capable to automatically estimate the number of clusters for large and high dimensional data sets but this work makes heavy use of cluster centroid and is not well tested for overlapping clusters.

The proposed work in this thesis is efficient enough to address the shortcomings of most of the aforementioned works. This method is able to automatically determine optimal number of partitions with relatively faster convergence to near true pareto optimal front while discovering clusters of various shapes for various high dimensional and overlapping data sets without any prior knowledge of the number of clusters. This method is reliable enough to reach convergence in fewer numbers of generations. Moreover, differences between the previous approaches lie mainly in the type of multi-objective GA coding and how objective functions are optimized simultaneously. Simulation results demonstrate its supreme capability to produce significantly better clustering result in comparison with other single and multi-objective approaches.

CHAPTER III

Theoretical Considerations

3.1 Introduction

This chapter describes the theoretical considerations of the proposed methodology. The multi-objective NSGA-II is first discussed along with its several features. The Fuzzy Relational Clustering (FRC) used in the proposed method is then described in details. The objective functions employed in the proposed approach are then presented. The clustering solution evaluation function is also given.

3.2 NSGA-II

The NSGA-II [4] is an exoteric non-domination based genetic algorithm for optimizing multi-objectives. In a sphere of multi-objective optimization, Pareto-optimal solutions are discovered by examining the solution domain using NSGA-II. It uses the better elitism and diversity. It has better non-domination sorting and no extra niching parameter (such as sharing parameter needed in the NSGA) is required. The non-dominated solution among the parent get favor from the feature elitism and here the good solution will never be lost until a better fitted solution is found. The different solutions are found from the near Pareto-optimal solution set of the final generation.

In the process of NSGA-II, it discovers the non-dominated solutions and finds the Pareto front from the Pareto-optimal solutions. It maintains a diversity rank to entire individuals containing identical non-dominated front by crowding comparison process. The individuals with lower rank or larger crowding distance are selected as parents by binary tournament selection. Through higher rank, the individuals are dispended for each non-dominated front. Based on crowding

distance on the last front and on rank from the current population, it selects only best N individuals as parents for the next generation and sorts the generated offspring according to non-domination.

The following are some features of NSGA-II.

3.2.1 Fast Non-dominated Sorting

In NSGA-II, the individuals are sorted based on non-domination in each front. The first front is one which is not dominated by any other individuals and the second front is one which is only dominated by individuals whose front number are 1 and so on. Based on front to which the individuals belong, the fitness value of the individuals are assigned. Hence, two entities are calculated for each solution: a) domination count n_p, where p solution is dominated by n_p number of solutions b) S_p, solution set dominated by solution p. In the fast non-dominated sorting process, If p dominates q, add q to the set S_p; else increment the domination counter n_p. If n_p = 0, p belongs to the first front and also initialize the front counter and find all other fronts for each element of S_p.

```
for each p ∈ P
    S_p = ∅
    n_p = 0
    for each q ∈ P
        If (p ≺ q) then                    If p dominates q
            S_p = S_p ∪ {q}                Add q to the set of solutions
dominated by p
        else if (q ≺ p) then
            n_p = n_p + 1                  Increment the domination
counter of p
    if n_p = 0 then                        p belongs to the first front
        p_rank = 1
        F_1 = F_1 ∪ {p}

i = 1                                      Initialize the front counter
while F_i ≠ ∅
    Q = ∅                                  used to store the members of
the next front

     for each p ∈ F_i
        for each q ∈ S_p
```

$n_q = n_q - 1$
if $n_q = 0$ then — q belongs to the next front
$q_{rank} = i + 1$
$Q = Q \cup \{q\}$
$i = i + 1$
$\mathcal{F}_i = Q$

Algorithm 3.1: NSGA-II non-dominated sorting procedure.

3.2.2 Diversity Mechanism

In the original NSGA, it contains two difficulties with sharing function system those are choosing appropriate σ_{share} value and comparing each solution to each other which creates high complexity. In the NSGA-II, it replaces the sharing function with a crowded-comparison process. Here, it uses a density estimation and crowed-comparison operator based on sorting.

3.2.3 Crowding Distance Assignment

The crowding-distance computation requires sorting the population according to each objective function value in ascending order of magnitude. Thereafter, for each objective function, the boundary solutions (solutions with smallest and largest function values) are assigned an infinite distance value. All other intermediate solutions are assigned a distance value equal to the absolute normalized difference in the function values of two adjacent solutions. This calculation is continued with other objective functions. The overall crowding-distance value is calculated as the sum of individual distance values corresponding to each objective. Each objective function is normalized before calculating the crowding distance. The algorithm as shown below outlines the crowding-distance computation procedure of all solutions in an non-dominated set *I*.

l = | I |
for each i, set $I[i]_{\text{distance}} = 0$
for each objective m
  I = sort(I, m)
  $I[1]_{\text{distance}} = I[l]_{\text{distance}} = \infty$
  for i = 2 to (l - 1)
    $I[i]_{\text{distance}} = I[i]_{\text{distance}} + \frac{(I[i+1]\cdot m - I[i-1]\cdot m)}{(f_m^{\max} - f_m^{\min})}$

Algorithm 3.2: Crowding distance assignment in NSGA-II.

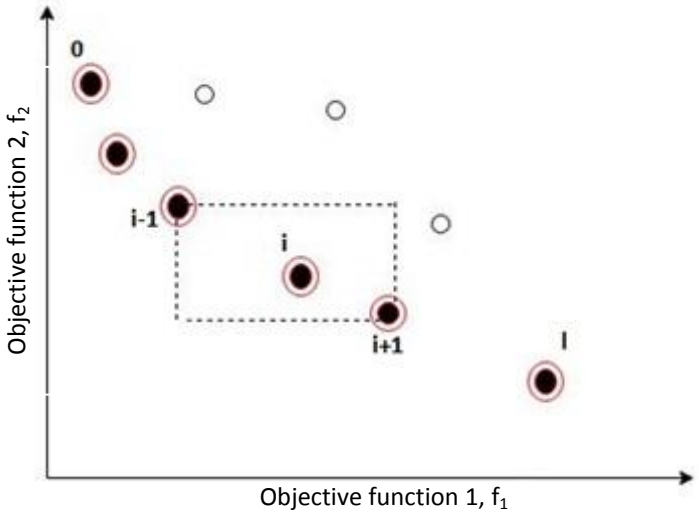


Figure 3.1: Crowding distance calculation for i^{th} solution.

Here, $I[i].m$ refers to the m^{th} objective function value of the i^{th} individual in the set I and the parameters f_m^{max} and f_m^{min} are the maximum and minimum values of the m^{th} objective function. The complexity of this procedure is governed by the sorting algorithm. Since M independent sortings of at most N solutions (when all population members are in one front I) are involved, the above algorithm has $O(MN \log N)$ computational complexity.

After all population members in the set I are assigned a distance metric, two solutions can be compared for their extent of proximity with other solutions. A solution with a smaller value of this distance measure is, in some sense, more crowded by other solutions. This is exactly what is compared in the proposed crowded-comparison operator. Although Figure 3.1 illustrates the crowding-

distance computation for two objectives, the procedure is applicable to more than two objectives as well.

3.3 Fuzzy Relational Clustering Algorithm (FRC)

The fuzzy relational clustering algorithm called FRC [27] has been proposed in this study for clustering of complex relational data set.

FRC uses graph representation in which nodes represent objects, and weighted edges represent the similarity between objects. This algorithm operates in an Expectation Maximization framework in which the graph centrality of an object in the graph is interpreted as likelihood.

DataRank [27] can be used within the Expectation-Maximization [27] algorithm to optimize the parameter values and to formulate the clusters. The modified DataRank algorithm [27] to deal with weighted undirected edges of a graph representing objects with nodes V.

$$DR(V_i) = (1-d) + d \times \sum_{j=1}^{N} \left(w_{ji} \quad \frac{DR(V_j)}{\sum_{k=1}^{N} w_{jk}} \right) \tag{3.1}$$

where w_{ji} is the similarity between V_j and V_i, and it is assumed that these weights are stored in a matrix W = $\{w_{ij}\}$, which is referred to as the *"affinity matrix"*.

The algorithm involves the following steps:

Initialization: Firstly, cluster membership values are initialized randomly, and normalized in this that cluster membership for an object sums to unity over all clusters. Mixing coefficients are initialized such that priors for all clusters are equal.

Expectation: Calculates the DataRank value for each object in each cluster. Calculation of DataRank values requires affinity matrix weights w_{ij} obtained by scaling the similarities by their cluster membership values [27]; i.e.,

$$w_{ij}^{m} = s_{ij} \times p_i^m \times p_j^m \tag{3.2}$$

where w_{ij}^m is the weight between objects *i* and *j* in cluster *m*, s_{ij} is the similarity between objects *i* and *j*, and p_i^m and p_j^m are the respective membership values of objects *i* and *j* to cluster *m*. The intuition behind this scaling is that an object's entitlement to contribute to the centrality score of some other object depends not only on its similarity to that other object, but also on its degree of membership to the cluster. Likewise, an object's entitlement to receive a contribution depends on its membership to the cluster. Once DataRank scores have been determined, these are treated as likelihoods and used to calculate cluster membership values.

Maximization: Updating the mixing coefficients based on membership values calculated in the Expectation Step.

The FRC algorithm [27] is described as follows where w_{ij}^m, s_{ij}, p_i^m and p_j^m are defined as above, π_m is the mixing coefficient for cluster *m*, DR_i^m is the DataRank score of object *i* in cluster *m*, and l_i^m is the likelihood of object *i* in cluster *m*.

The *FRC* Algorithm:

Input: Pairwise similarity values $S = \{ s_{ij} \mid i = 1,....,N, j = 1,....,N \}$ where s_{ij} is the similarity between objects *i* and *j*. Number of clusters, *C*.

Output: Cluster membership values $\{ p_i^m \mid i = 1,....,N, m = 1,....,C \}$

1. *// INITIALIZATION*
2. *// initialize and normalize membership values*
3. **for** *i* = 1 to *N*
4. **for** *m* = 1 to *C*
5. p_i^m = rnd *// random number on [0, 1]*
6. **end for**
7. **for** *m* = 1 to *C*
8. $p_i^m = p_i^m / \sum_{j=1}^{C} p_i^j$ *// normalize*
9. **end for**
10. **end for**
11. **for** *m* = 1 to *C*
12. $\pi_m = 1/C$ *// equal priors*
13. **end for**
14. **repeat until convergence**

15. // *EXPECTATION STEP*
16. **for** *m* = 1 to *C*
17. // *create weighted affinity matrix for cluster m*
18. **for** *i* = 1 to *N*
19. **for** *j* = 1 to *N*
20. $w_{ij}^{m} = s_{ij} \times p_i^m \times p_j^m$
21. **end for**
22. **end for**
23. *// calculate DataRank scores for cluster m*
24. **repeat until convergence**
25. $DR_i^m = (1 - d) + d \times \sum_{j=1}^{N} w_{ji}^m \left(DR_j^m \Big/ \sum_{k=1}^{N} w_{jk}^m \right)$
26. **end repeat**
27. *// assign DataRank scores to likelihoods*
28. $l_i^m = DR_i^m$
29. **end for**
30. *// calculate new cluster membership values*
31. **for** *i* = 1 to *N*
32. **for** *m* = 1 to *C*
33. $p_i^m = (\pi_m \times l_i^m) / \sum_{j=1}^{C} (\pi_j \times l_i^j)$
34. **end for**
35. **end for**
36. *// MAXIMIZATION STEP*
37. *// Update mixing coefficients*
38. **for** *m* = 1 to *C*
39. $\pi_m = \frac{1}{N} \sum_{i=1}^{N} p_i^m$
40. **end for**
41. **end repeat**

3.4 Objective Functions

In this study, two objective functions have been considered which are based on two indexes: compactness and separation. The compactness indicates variation between data within a cluster or between data and cluster centroids, and it must be kept small. The separation measures the isolation of clusters, which is preferred to be large.

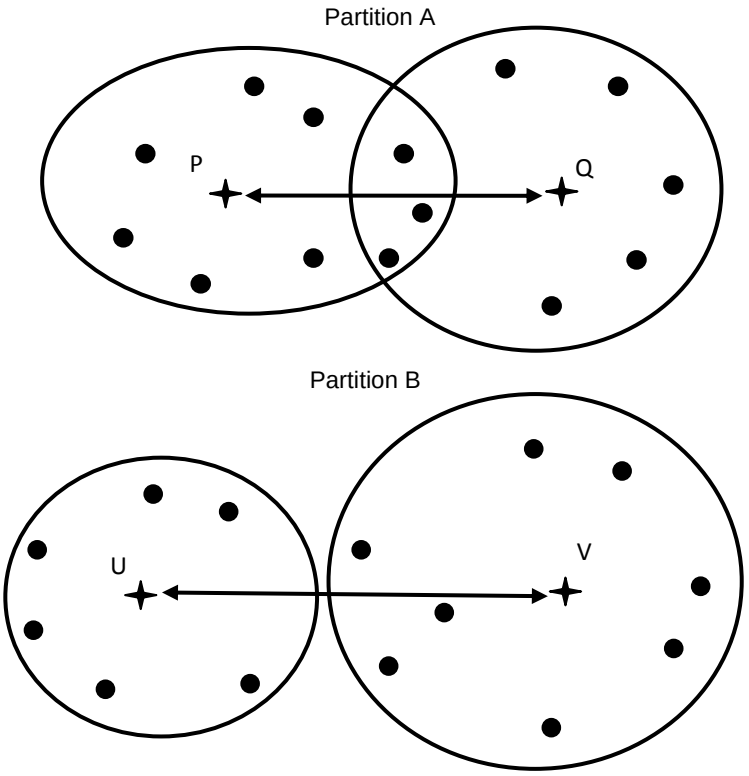


Figure 3.2: Two partitions having equal distance between centroids.

If only compactness or separation is considered in a single-objective optimization of clustering, some limitations may degrade the performance of the optimization problem. The drawback of compactness is well known that it suffers from a monotonic decrease with increasing number of clusters [28-29]. For the traditional separation, its disadvantage has been raised in [28]. Separation considering inter-distance measurement between cluster centroids cannot correctly detect geometric structures. As demonstrated in Figure 3.2, the distance from centroids *P* to *Q* in the partition *A* compared to the distance between *U* and *V* in the partition *B* are equal. In terms of the traditional measure of inter-cluster distance, these two partitions have the same property of separation but intuitively partition *B* is shown to have more separation. It can be seen that the measurement of centroid distance only can misjudge the separation of clusters because the overall shape is not considered. This leads to limited information about cluster structures.

In order to solve the aforementioned problems, two objectives are optimized simultaneously in the multi-objective optimization of FRC-NSGA-II. These objectives are based on compactness together with overlap and separation measure.

Let $X = \{x_1, x_2, \ldots, x_k\}$ be data points of M patterns, where each pattern x_k is a vector of features in R^n (n-dimensional space). C is the number of clusters.

The compactness is formulated by the objective function J_m proposed by Bezdek [30] as shown in Eq. (3.3)[56].

$$J_m(U,\ V) = \sum_{k=1}^{M} \sum_{i=1}^{C} (u_{ik})^m\ d_{ik}^2 \tag{3.3}$$

where d_{ik}^2 denotes the squared Euclidean distance calculated in n-dimensional space. Squared Euclidean distance calculates a distance from a data point x_k to a cluster center v_i as follows [56]:

$$d_{ik}^2 = \sum_{j=1}^{n} \left(x_{kj} - v_{ij}\right)^2,\ 1 \leq k \leq M, 1 \leq i \leq C$$

The distance is used in the calculation of membership degree in Eq. (3.4)[56]

$$u_{ik} = \frac{1}{\sum_{j=1}^{C} \left(\frac{d_{ik}}{d_{jk}}\right)^{2/(\mathrm{m}-1)}},\ 1 \leq k \leq M, 1 \leq i \leq C \tag{3.4}$$

u_{ik} denotes a degree of membership of x_k in the i^{th} cluster. m>1 is a parameter which controls a degree of fuzziness. This means that each data pattern has a degree of membership in every cluster.

The sum of the squared error is measured by the squared Euclidean distance from a pattern to each centroid with the weight $(u_{ik})^m$. The aim is to minimize J_m to optimize compactness taking into account distance and degree of membership.

Overlap and separation measure (overlap-separation for short) is the second objective that has a functionality to tackle the overlapping problem and to solve the problem of centroid separation as shown in Figure 3.2. This measure is based on an aggregation operation of fuzzy membership degrees. There are two parts for

this index. The first part is the overlap measure proposed by [31,29], which computes an inter-cluster overlap using fuzzy degrees as shown in Eq. (3.5).

$$O\perp(u_k(x_k), C) = \underset{l=2,C}{\overset{1}{\perp}} \left(\underset{i=1,C}{\overset{l}{\perp}} u_{ik} \right) \tag{3.5}$$

A pattern x_k has membership vector $u_k(x_k) = (u_{1k}, \ldots., u_{ck})$. The ambiguity measurement of several membership values requires an aggregation operator (AO). The AO applied here is based on triangular norms (t-norms) for *l*-order ambiguity measurement. The standard t-norm and t-conorm are used in this aggregation. The standard t-norm has the basic property that $a \top b = \min(a, b)$ and for the standard t-conorm $a \perp b = \max(a, b)$. With a single value $\overset{l}{\underset{\perp}{}} u_k \in [0, 1]$, the *l*-order fuzzy-OR operator (fOR-*l*) which is the combination of a dual couple ($\top\perp$) [32] is associated with u_k, defined by

$$\underset{i=1,C}{\overset{l}{\perp}} u_{ik} = \underset{A\epsilon\mathcal{P}_{l-1}}{\top} \left(\underset{j\in C\setminus A}{\overset{\perp}{}} u_j \right) \tag{3.6}$$

$\mathcal{P}$ denotes the power set of C= {1,2, ….., *c*} and $\mathcal{P}_l = \{ A \epsilon \mathcal{P} : |A| = l \}$,where $|A|$ denotes the cardinality of the subset *A*. From Eq. (3.6), the sorting in decreasing order ($u_1 \geq \ldots \geq u_C$) is obtained, and then, the l^{th} highest value is chosen. We apply *l*=2 (i.e., ambiguity measures between two classes) so that the second largest element of u_k can be used.

For the separation measure [28], the maximum degree of $\overset{\max}{\underset{i=1,C}{}} u_{ik}$ is taken into account. Therefore, the overall overlap-separation (OS) measure for *M* patterns is defined as follows:

$$OS = \frac{1}{M} \sum_{k=1}^{M} \frac{O\perp(u_k(x_k), C)}{max_{\,i=1\,,c} u_{ik}} \tag{3.7}$$

3.5 Genetic Operators

In the FRC-NSGA-II clustering method, solutions have to be optimized in continuous search space according to the real-coded string representation of cluster centers. The stochastic search for the real-coded string is made possible by the binary tournament selection [33], the simulated binary crossover (SBX) operator [33] and the polynomial mutation operator in [34].

3.5.1 Binary Tournament Selection

In NSGA-II, the binary tournament selection is used to select parents to create the new generation. Two individuals are randomly chosen to play a tournament and a winner is chosen by the crowded comparison operator ($\prec_n$). This operator considers two attributes which are non-domination rank (i_{rank}) and crowding distance (i_{dist}). Let two individuals be *i* and *j*, the crowded comparison operator $\prec_n$ is defined as:

$$i \prec_n j \text{ if } (i_{rank} < j_{rank})$$

$$\text{or } ((i_{rank} = j_{rank})$$

$$\text{and } (i_{distance} > j_{distance}))$$

The lower rank is preferred if two individuals are in different ranks. If both individuals are in the same front (same rank), the solution with lesser crowded region is chosen.

3.5.2 SBX Operator

The SBX operator [33] performs similarly to the search power of a single-point crossover on binary strings and maintains the interval schemata processing in continuous variables instead of discrete variables. To control how children are different from their parents, a spread factor β is defined under the probability distribution function:

$$C(\beta) = \begin{matrix} 0.5(\eta_c+1)\beta^{\eta_c}, & \text{if } \beta \leq 1 \\ 0.5(\eta_c+1)\frac{1}{\beta^{\eta_c+2}}, & \text{otherwise} \end{matrix} \quad (3.8)$$

The value of the distribution index η_c which is any non negative real number has an impact on the spread of child solutions from parent solutions. A large value of η_c gives a high probability of obtaining children solutions near to parent solutions whereas a small value of η_c allows children far from their parents. From the probability distribution in Eq. (3.8), $\overline{\beta}$ is a random variable which makes the area under the probability curve equal to a uniform random number $u(0, 1)$, as follows:

$$\overline{\beta} = \begin{matrix} (2u)^{\frac{1}{\eta_c+1}}, & if\ u \leq 0.5 \\ \left(\frac{1}{2(1-u)}\right)^{\frac{1}{\eta_c+1}}, & otherwise \end{matrix} \tag{3.9}$$

To obtain children solutions, two parent solutions P and Q are selected from a mating pool by the binary tournament selection. Thereafter, a random number u is generated and $\overline{\beta}$ is calculated from Eq. (3.9). Consider $P = (p_1, \cdots, p_n)$ and $Q = (q_1, \cdots, q_n)$, where n is the length of strings. Two children c_i^1 and c_i^2 are calculated in Eqs. (3.10) and (3.11).

$$c_i^1 = 0.5\left[\left(1+\overline{\beta}\right)p_i + \left(1-\overline{\beta}\right)q_i\right] \tag{3.10}$$

$$c_i^2 = 0.5\left[\left(1-\overline{\beta}\right)p_i + \left(1+\overline{\beta}\right)q_i\right] \tag{3.11}$$

Since each chromosome in the population has different lengths, the chromosomes of two parents may have equal or different lengths. In case their lengths are equal, the crossover operation can be illustrated in Figure 3.3. Let two parents $P = \{p_1, p_2, p_3, p_4\}$ and $Q = \{q_1, q_2, q_3, q_4\}$ denote parent solutions with four cluster centers, where each p_i and q_i is a vector of features. Two children A and B are created. In uniform crossover, the decision to perform crossover on each pair of centers from parents is with a probability 0.5 [33]. In this example, the crossover operation does not perform on position 3, the value of p_3 and q_3 are directly copied to position 3 of strings A and B, respectively. Positions 1, 2 and 4 have new values a_1, a_2, a_4 and b_1, b_2, b_4 on A and B, respectively.

The new values of centers for two children are calculated from Eqs. (3.10) and (3.11) by the SBX procedure.

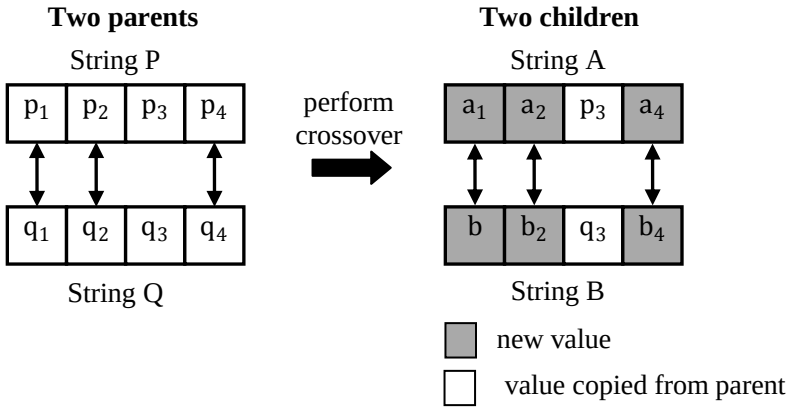


Figure 3.3: Crossover operation for two chromosomes with same length.

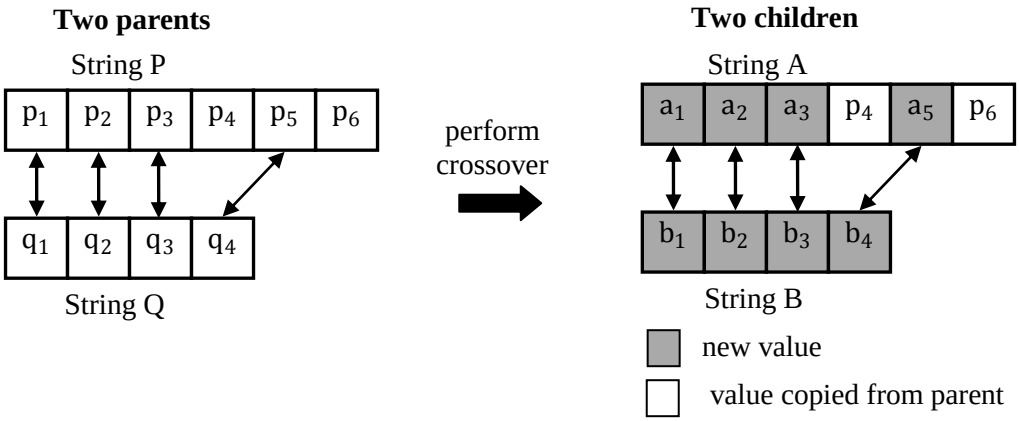


Figure 3.4: Crossover operation for two chromosomes with different length.

3.5.3 Mutation Operator

The polynomial mutation operator defined in [34, 35] is applied with a low probability to perturb a solution for the new population. The result of mutation is controlled by a probability distribution:

$$P(\delta) = 0.5(\eta_m + 1)(1 - |\delta|)^{\eta_m} \tag{3.12}$$

From the above distribution, the perturbation factor $\overline{\delta}$ can be calculated according to a random number r_i in the range $(0,1)$ and the distribution index η_m, as follows:

$$\overline{\delta}_i = \begin{cases} (2r_i)^{1/(\eta_m+1)} - 1, & \text{if } r_i < 0.5 \\ 1 - [2(1-r_i)]^{1/(\eta_m+1)}, & \text{if } r_i \geq 0.5 \end{cases} \tag{3.13}$$

One parent is chosen as x_i which is the value of the i^{th} cluster. x_i^L and x_i^U are the lower and upper bound of x_i, respectively. The mutated value y_i is therefore calculated by

$$y_i = x_i + \left(x_i^U - x_i^L\right)\overline{\delta}_i \tag{3.14}$$

3.6 Selection and Evaluation of the Solution Set

In the final generation, the FRC-NSGA-II algorithm produces a set of non-dominated solutions whose number varies according to the population size. All the solutions are considered to be equal in terms of fitness values compromised by the two objectives. In most real world problems, a single solution must be chosen out of this set.

Here the selection mechanism presented in [8] is deployed where a semi-supervised method has been used. The class label of 10% of the whole data set is assumed to be known. The remaining 90% of the sample has no class label information provided and FRC-NSGA-II executes on these unknown label samples called *test patterns.* After the clustering procedure in the multi-objective optimization has finished, patterns are grouped in their corresponding clusters but class labels of clusters have not been defined. The class labels are later assigned by the following procedure. In [8], the assignment of class labels is based on the nearest center criterion. In FRC-NSGA-II, the fuzzy degree of membership is combined with the center-based criterion in the class label assignment. First, each known label sample is mapped to a cluster corresponding to the maximum degree of membership. Second, a frequency table is built by the frequency of the samples

that fall in their mapped clusters. In the final step, the label of the cluster is chosen from the known label of the samples having the maximum frequency. If frequencies between the groups are equal, the cluster label is assigned by the label of the sample that has the maximum fuzzy degree with the corresponding cluster.

After the label assignment procedure, the *Minkowski score* (MS) according to [8] is computed to measure the amount of misclassification as shown in Eq. (3.15). Consider the *true* solution set T and the solution set to be measured S. The measure is defined by the number of point-pairs assignments of data items between T and S. n_{11} denotes the number of point-pairs that are in the same class in both S and T. n_{01} and n_{10} represents the mismatched number of point-pairs between the two sets. n_{01} denotes the number of point-pairs that are in S only, and n_{10} denotes the number of point-pairs that belong to T only. A lower score indicates a better solution. After the scores of all non-dominated solutions have been calculated, the solution with the minimum score is chosen as the best solution.

$$MS(T,S) = \sqrt{\frac{n_{01}+n_{10}}{n_{11}+n_{10}}} \tag{3.15}$$

3.7 Summary

This chapter presents the working procedure of multi-objective NSGA-II followed by representation of the fuzzy relational clustering, FRC along with its pseudocode. The objective functions used in the proposed approach were then described. The involved genetic operators named binary tournament selection, simulated binary crossover and polynomial mutation are also discussed.

CHAPTER IV

Proposed Method

4.1 Introduction

This chapter describes the concept, working principle of proposed fuzzy relational clustering algorithm based on multi-objective genetic algorithm called FRC-NSGA-II in detail.

4.2 Proposed Method: Fuzzy Relational Clustering

In this study, we propose an evolutionary fuzzy relational clustering (FRC) algorithm called FRC–NSGA-II by combining FRC with NSGA-II algorithm in which it sustains the goodness of both FRC and NSGA-II algorithms. The overall block diagram of the proposed system is shown in the Figure 4.1. The details description of our proposed algorithm is as follows:

Data in n-dimensional Space: The n-dimensional data set (non-gene/gene expression) without the prior knowledge of number of cluster is used as the input of the proposed fuzzy relational clustering as shown in Figure 4.1.

Non-dominated Sorting GA (NSGA-II): NSGA-II has been applied as an optimization framework for multi-objective optimization of appropriate fuzzy relational clusters. NSGA-II smartly creates a finer trade-off between two objectives named compactness and overlapping/separation. It assigns the appropriate number of cluster and optimizes multiple clustering validity measures simultaneously for yielding robust clustering solutions. It seeks a great coverage of solutions and faster convergence close to the near true Pareto-Optimal front.

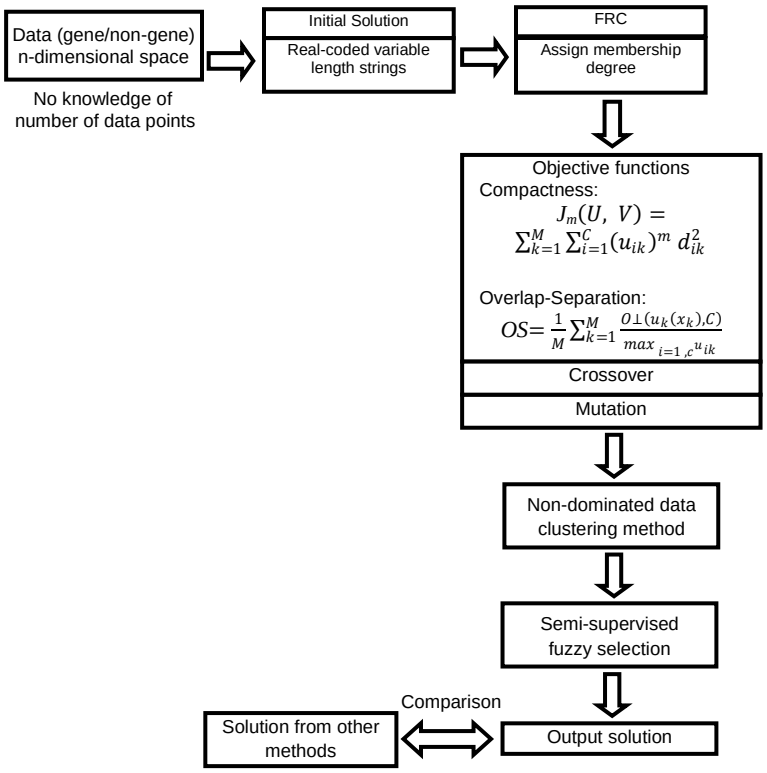


Figure 4.1: Overall block diagram of the proposed system.

Fuzzy Relational Clustering (FRC): FRC method assigns membership values to each data point. It takes a relational input data which is in the form of a similarity's square matrix between data objects. The Data-Rank score of an object within some cluster is interpreted as likelihood and then the Expectation-Maximization (EM) framework is used to determine cluster membership values and mixing coefficients.

In the expectation-maximization steps, likelihood is the graph centrality of an object in the graph. In Expectation Step, it calculates the value of DataRank for each object in each cluster using the affinity matrix weights. In Maximization step, based on the Expectation step's calculated membership values of cluster, mixing coefficients are updated.

Objective functions: The objective functions for two well-known cluster validity indices, compactness and separation, are optimized concurrently through multi-objective NSGA-II where compactness indicates variation between data within a cluster and separation means quantifying the separation between different clusters. These objective functions enable to obtain expected results in case of well-separated, hyperspherical and overlapping clusters.

Non-dominated data clustering method: The non-dominated solutions are generated in this phase. In a solution domain a solution is Pareto-optimal if it denies domination from other solutions.

Semi-supervised fuzzy selection: This phase performs fuzzy clustering of non-dominated solutions.

Output solution: The Minkowski score is assessed for all the non-dominated solutions of the final generation. The non-dominated solution with lowest Minkowski score is then chosen as the single best solution.

Comparison: The solution obtained from the proposed method is compared with those from other existing methods.

The overall functional block diagram in detail of the proposed algorithm is presented in Figure 4.2.

The initial numbers of clusters are generated from a uniform distribution over the range 2 to $\sqrt{M}$ which is recommended by [8] where M is the number of data points. The values of fitness are then computed according to the two objectives detailed in chapter 3. Thereafter the fast non-dominated sorting and crowding distance assignment of NSGA-II are applied.

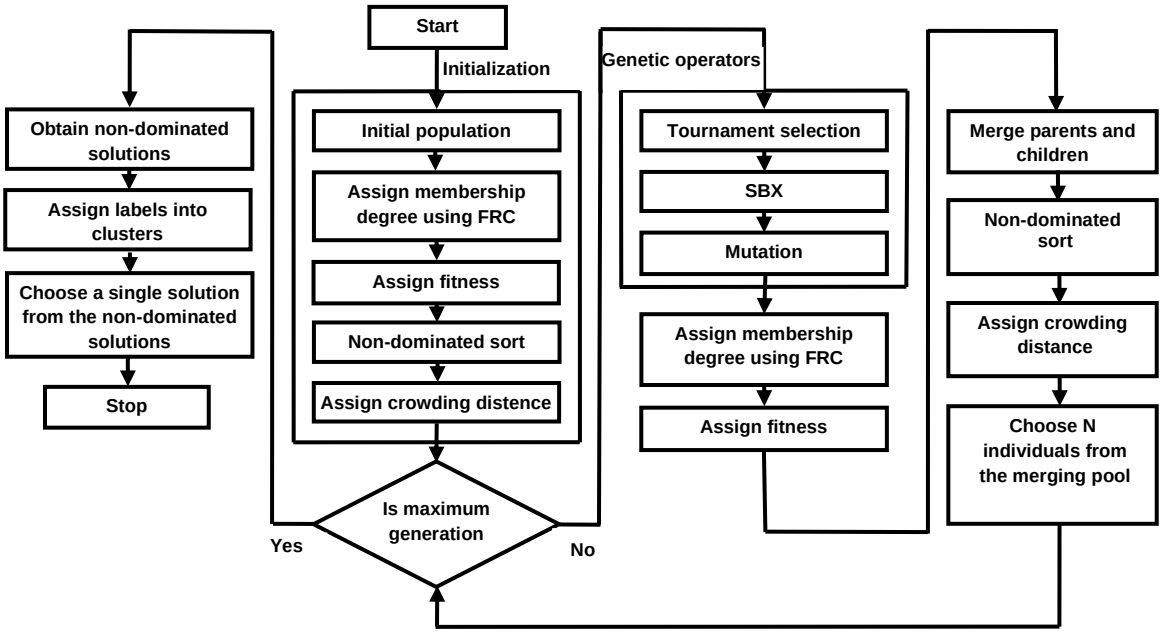


Figure 4.2: Functional block diagram of FRC-NSGA-II.

After the first generation, the NSGA-II operators then produce child solutions (chromosomes) in the evolutionary process through simulated binary crossover (SBX) and mutation operators. The parent and child solutions are then merged to 2*N* solutions but only *N* individuals are chosen after the non-dominated sorting and finish the procedure of crowded comparison. The process runs and searches for non-dominated front solutions until it meets the termination criterion. In this study maximum number of generation is used as the termination criteria. Finally, the non-dominated solutions are obtained. On each solution, each data point is assigned into a cluster corresponding to its maximum degree of membership.

For solving the fuzzy clustering problem, the FRC–NSGA-II algorithm can be explained as follows:

Algorithm 4.1: FRC-NSGA-II for clustering

Input: High dimensional data points, *N*.

Output: *K* clusters while minimizing/maximizing multiple objectives.

1. Initialize parameters.
    - population size
    - number of objective functions
    - number of generations
    - minimum and maximum limit of the encoded variable for cluster number
    - cluster fuzziness
2. Form initial population.
3. Fitness evaluation of solutions.
4. Estimates cluster number and delivers it to FRC.

// FRC execution begins

5. Assign membership values to data points randomly and nomalize them.
6. While until convergence
7. repeat
8. generate affinity matrix.
9. estimate datarank scores.
10. transform datarank scores to likelihoods.
11. end repeat
12. Update cluster membership of datapoints.
13. Update mixing coefficients based on new cluster membership values.
14. End While.

// FRC execution ends with deliverance of membership values

15. Fitness values are calculated using membership values found from FRC for each chromosome in the population.

16. Ranking of the individuals in the population is performed by

    a) Fast non-dominated sorting process

    b) Crowding distance calculation.

17. Produce new population filled with child chromosomes generated by selection, crossover and mutation manipulations on individuals in parent population.
18. Parent population replaced by new population.
19. Repeat step 3 if termination criterion is not met.
20. Non-dominated solution set is returned.

CHAPTER V

Simulation Results

In this chapter, experimental data sets and implementation results are presented in order to interpret the effectiveness of the proposed fuzzy relational clustering algorithm called FRC-NSGA-II based on multi-objective NSGA-II algorithm.

5.1 Data Sets

The proposed approach has been applied on two categories of data sets: non-gene and gene expression data sets.

5.1.1 Difference between Non-gene and Gene Expression Data Sets

A non-gene data set consists of many instances which are comprised of several attributes. The instances are distributed over some classes depending upon certain conditions.

A Gene expression data set contains values collected from DNA microarray gene expression profile of the respective disease. Each gene expression data set has several samples collected from relative disease patients. These samples are assigned to some predefined classes depending on the subtype of that disease.

5.1.2 Non-gene Expression Data Sets

In order to validate the proposed model, it is first examined on eleven well known non-gene expression data sets. Among the eleven non-gene expression experimental data sets, seven data sets are artificial and another four data sets are real. Based on these data sets, implementation results are illustrated and compared with other existing multi-objective and single-objective clustering algorithms.

Real life non-gene expression data sets (Iris, Glass, Wine and Liver-Disorder) are collected from the UCI Machine Learning Repository (Chang et *al.*, 2009) [36] for experimental purpose. The artificial non-gene data sets are AD_5_2 [37], AD_10_2 [38], Square-1, Square-4, Long-1, Sph_5_2 and Sph_6_2. Table 5.1 shows the brief description of non-gene data sets, where D is the number of features, K is the number of clusters, M is sample size and M_i is the number of members in cluster i.

Table 5.1: Summary of non-gene expression data sets

DataSet	D	K	M	M_i
AD_5_2	2	5	250	5×50
AD_10_2	2	10	500	10×50
Square-1	2	4	1000	4×250
Square-4	2	4	1000	4×250
Long-1	2	2	1000	2×500
Sph_5_2	2	5	250	5×50
Sph_6_2	2	6	300	6×50
Glass	9	6	214	70, 76, 17, 13, 9, 29
Wine	13	3	178	59, 71, 48
Iris	4	3	150	50, 50, 50
Liver-Disorders	6	2	341	142, 199

The non-gene data sets shown in the Table 5.1 have the following features:

- AD_5_2: This data set contains five classes with highly overlapped data. The data set has 250 data points with 50 data points in each cluster
- AD_10_2: This data set contains ten classes which are equal in size. The data set has 500 data points each of which has two attributes.
- Square-1: This data set contains four classes which are well separated and equal in size. The data set has 1000 data points each of which has two attributes. It is generated by normal distributions with a standard deviation of two in both dimensions.
- Square-4: The data set contains four classes of equal size. The data set has 1000 data points each of which has two attributes. It is generated by normal distributions with a standard deviation of two in both dimensions.

- Long-1: This data set contains two long shaped clusters with 1000 instances.

- Sph_5_2: This data set comprises 250 instances each of which has two attributes. The data set contains five classes of equal size.

- Sph_6_2: This data set comprises 300 instances which are divided into six classes of equal size.

- Glass: This data set contains 214 data points which are distributed over six classes of unequal size.

- Wine: This data set consists of 178 instances which are divided into three classes. Each data point has thirteen chemical attributes.
- Iris: This data set contains 150 instances which are distributed over three classes of equal size. Each instance is described by four physical attributes which are length and width of sapals and petals.
- Liver-Disorders: This data set comprises 341 data points which are divided into two classes. Each data point has six attributes.

5.1.3 Gene Expression Data Sets

DNA microarray gene expression data enables us to monitor expression patterns of thousands of genes simultaneously. In the field of medical diagnosis, biological research and development, there exists many applications using microarray technology. Clustering of microarray gene expression data is used to group the sets of co-expressed genes with similar expression characteristics. Clustering biomedical data is challenging due to their noise, high dimensionality, limited amount of samples and abundance of redundancy among genes.

There are many benchmark microarray data sets in the field of cancer gene expression analysis, including Leukemia cancer data set [39], Lymphoma cancer data set [40], Prostate Tumor cancer data set [41], Colon cancer data set [55], Breast cancer data set and Ovarian cancer data set. In this research, only 4 gene

expression data sets have been used: Leukemia, Lymphoma, Prostate Tumor, Colon cancer data set.

The gene expression data sets have the following features:

- Leukemia: This data set that has been used in this research is probably the most famous gene expression cancer data set (Golub et *al*., 1999)[39]. In Leukemia data set, it has 7129 genes which are divided into 2 classes: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The data set consists of 72 samples in which 47 are in class ALL and 25 are in class AML.
- Lymphoma: This data set contains three most prevalent adult lymphoid malignancies. It consists of 66 samples from 4026 genes. This is composed of 46, 9 and 11 samples of DLBCL, FL and CLL respectively.
- Prostate Tumor: For this cancer data set, detailed information and data is available in [41]. The data set contains prostate tissue 102 patient's samples from 12,600 genes. Among 102 samples, there are 52 prostate tumor samples and 50 normal samples.
- Colon: This data set contains 62 samples of 2000 genes collected from colon cancer patients. There are 40 tumor biopsies (labelled as "Tumor") and 22 normal (labelled as "Normal").

5.1.4 Effective Gene Selection for Gene Expression Data Sets

In microarray cancer data set, it contains a large number of genes expressions. It is essential to find those genes which are highly correlated with the class distinction to be predicted. Here the Correlation coefficient method has been used for finding the degree of their correlation. Then among a large number of genes only some genes with highest correlation coefficient values have been selected for further clustering process.

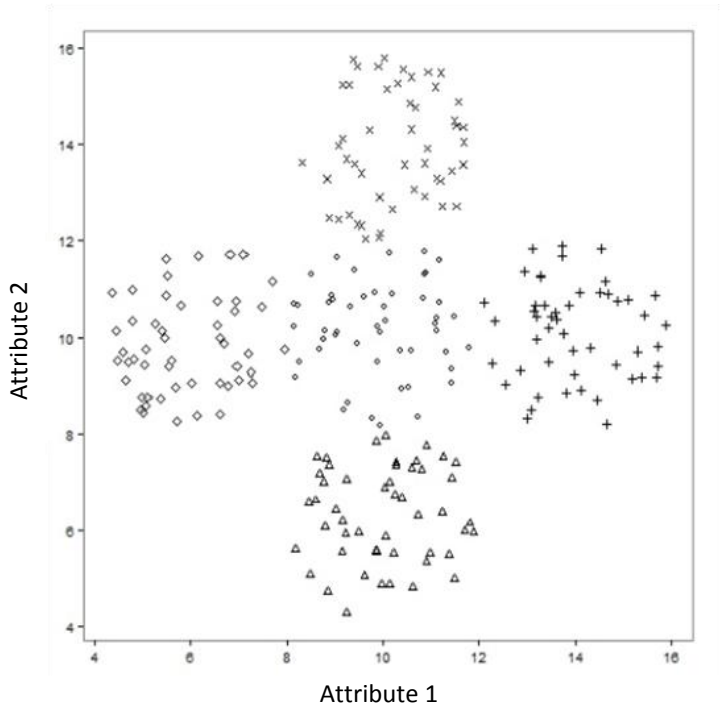
5.2 Implementation Results of Data Sets

This section shows the implementation results found for both non-gene and gene expression data sets.

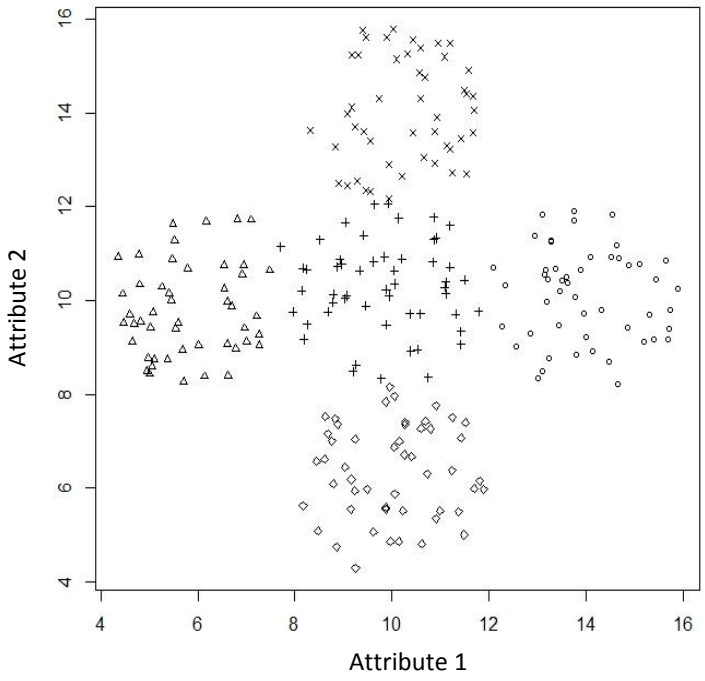
5.2.1 Non-gene Expression Data Sets

- **Non-gene Data Set: AD_5_2**

Figure 5.1(a) shows the original/true solution of clustering the data set AD_5_2 where number of cluster is 5 and Figure 5.1(b) shows the clustering performed by the proposed fuzzy relational clustering algorithm, FRC-NSGA-II. In this case, number of cluster is 5 and the Minkowski Score(MS) is 0.28.



(a)

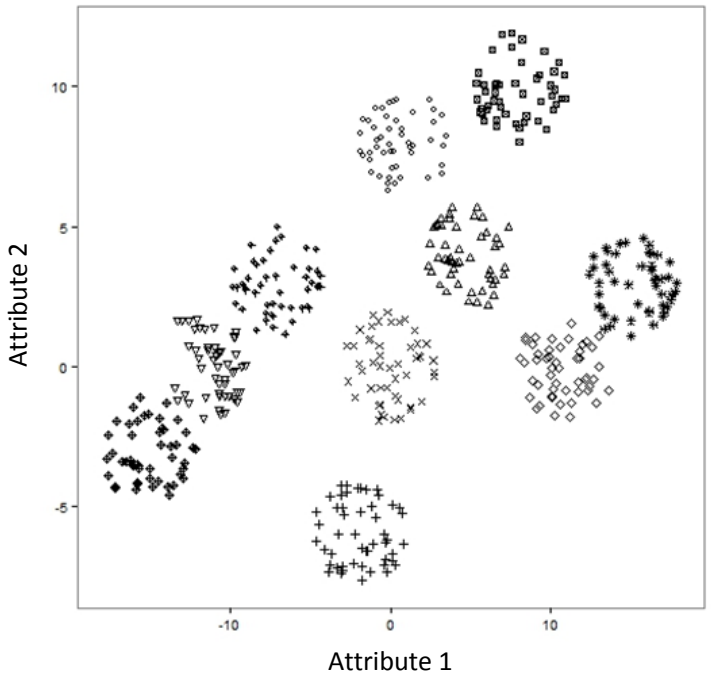


(b)

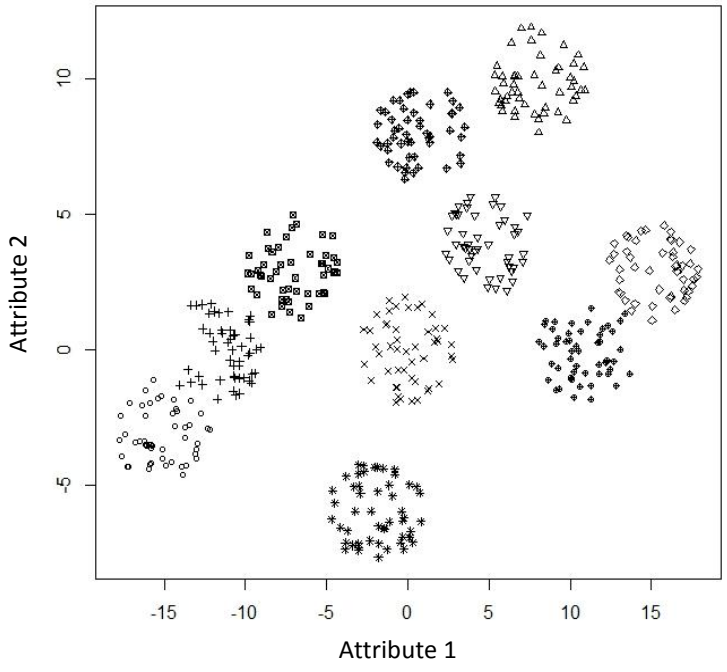
Figure 5.1: (a) Original/True solution and (b) Clustering using FRC-NSGA-II (MS = 0.28) for the data set AD_5_2 (K = 5).

- **Non-gene Data Set: AD_10_2**

Figure 5.2(a) shows the original/true solution of clustering the data set AD_10_2 where number of cluster is 10 and Figure 5.2(b) shows the clustering performed by the proposed fuzzy relational clustering algorithm, FRC-NSGA-II. In this case, number of cluster is 10 and the MS is 0.12.



(a)

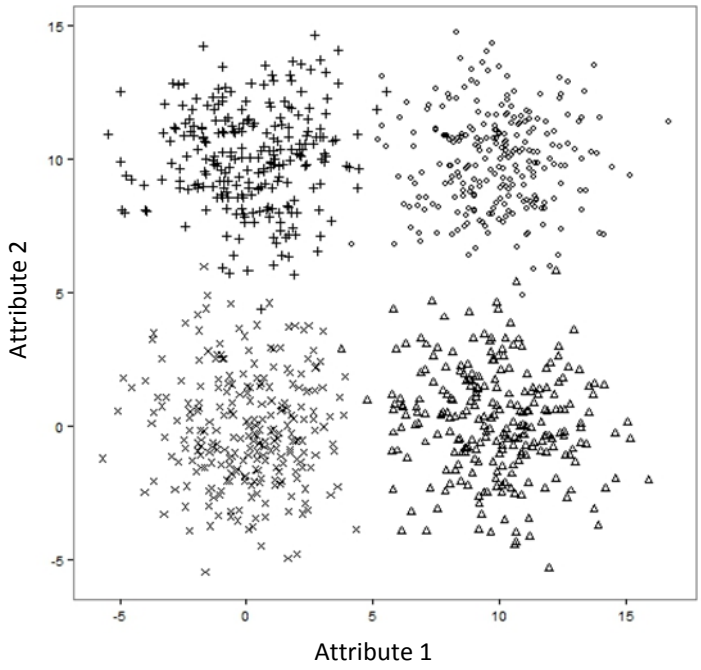


(b)

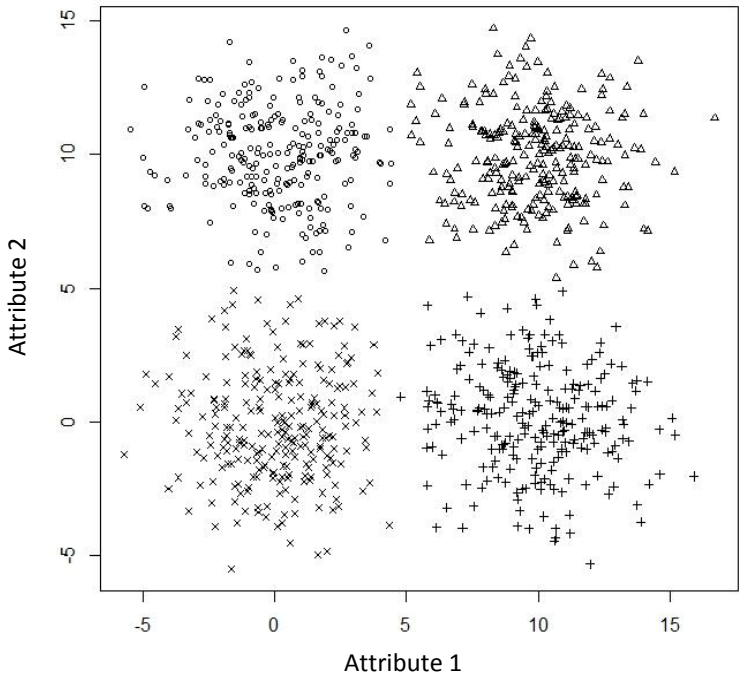
Figure 5.2: (a) Original/True solution and (b) Clustering using FRC-NSGA-II (MS = 0.12) for the data set AD_10_2 (K = 10).

- **Non-gene Data Set: Square-1**

Figure 5.3(a) shows the original/true solution of clustering the data set Square-1 where number of cluster is 4 and Figure 5.3(b) shows the clustering performed by the proposed fuzzy relational clustering algorithm, FRC-NSGA-II. In this case, number of cluster is 4 and the MS is 0.12.



(a)

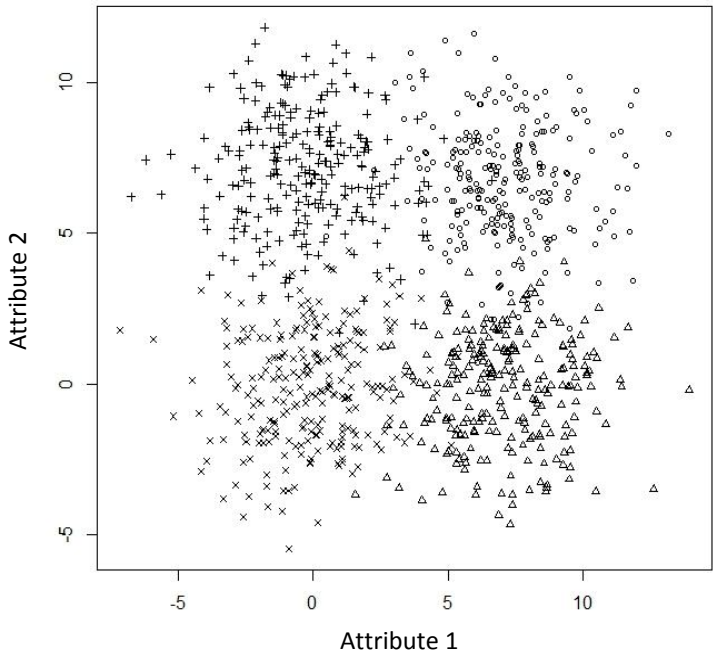


(b)

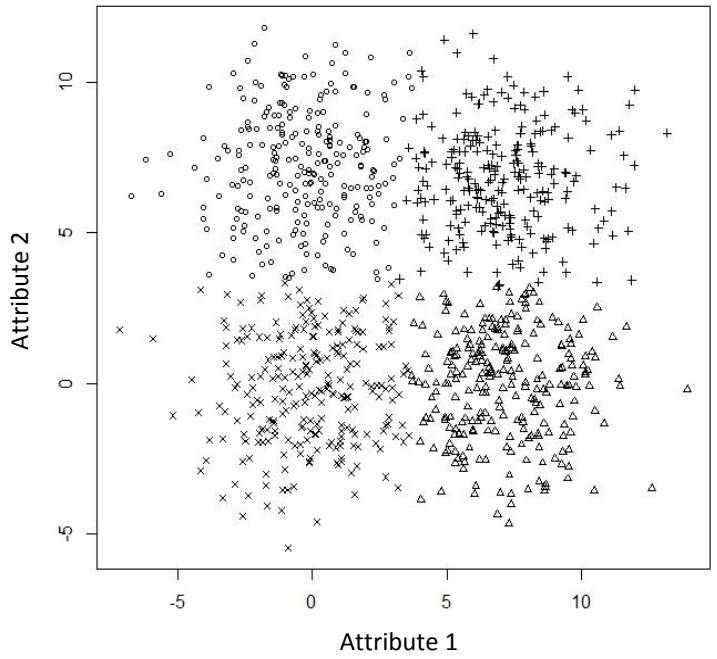
Figure 5.3: (a) Original/True solution and (b) Clustering using FRC-NSGA-II (MS = 0.12) for the data set Square-1 (K = 4).

- **Non-gene Data Set: Square-4**

Figure 5.4(a) shows the original/true solution of clustering the data set Square-4 where number of cluster is 4 and Figure 5.4(b) shows the clustering performed by the proposed fuzzy relational clustering algorithm, FRC-NSGA-II. In this case, number of cluster is 4 and the MS is 0.50.



(a)

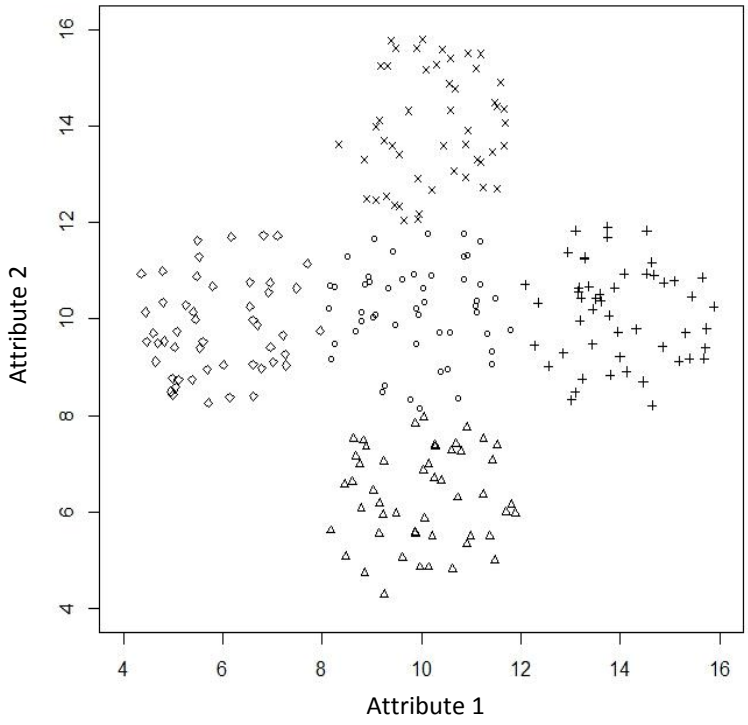


(b)

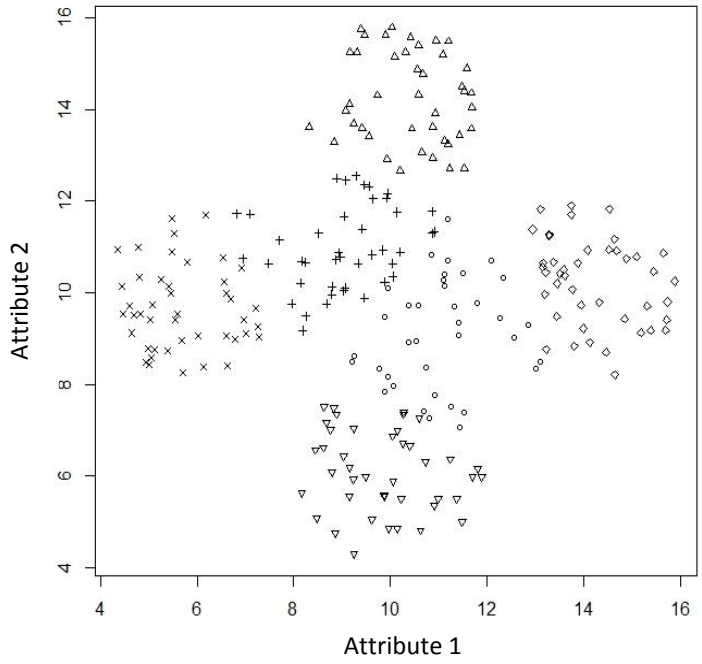
Figure 5.4: (a) Original/True solution and (b) Clustering using FRC-NSGA-II (MS = 0.50) for the data set Square-4 (K = 4).

- **Non-gene Data Set: Sph_5_2**

Figure 5.5(a) shows the original/true solution of clustering the data set Sph_5_2 where number of cluster is 5 and Figure 5.5(b) shows the clustering performed by the proposed fuzzy relational clustering algorithm, FRC-NSGA-II where number of cluster is 6 and the MS is 0.65.



(a)

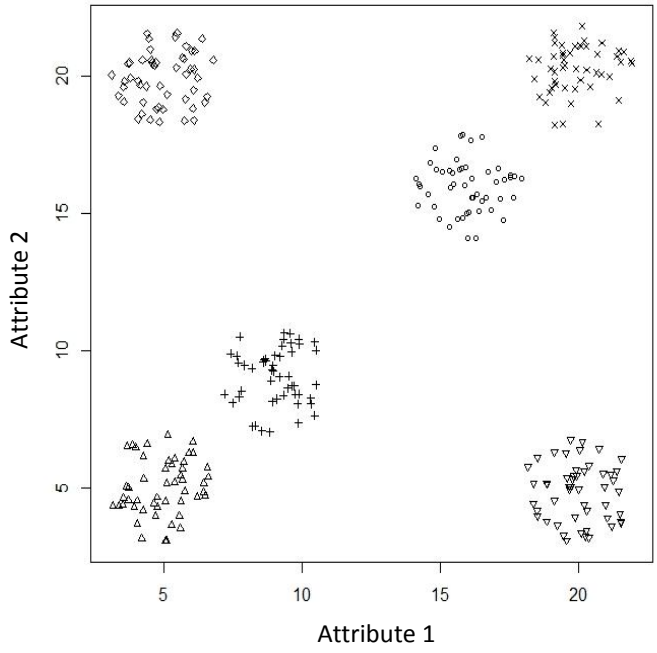


(b)

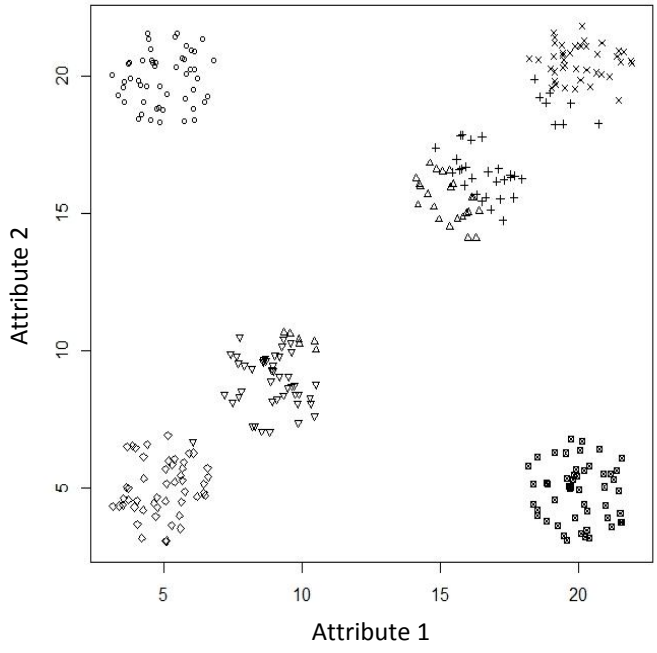
Figure 5.5: (a) Original/True solution and (b) Clustering using FRC-NSGA-II (MS = 0.65) for the data set Sph_5_2 (K = 6).

- **Non-gene Data Set: Sph_6_2**

Figure 5.6(a) shows the original/true solution of clustering the data set Sph_6_2 where number of cluster is 6 and Figure 5.6(b) shows the clustering performed by the proposed fuzzy relational clustering algorithm, FRC-NSGA-II where number of cluster is 7 and the MS is 0.47.



(a)



(b)

Figure 5.6: (a) Original/True solution and (b) Clustering using FRC-NSGA-II (MS = 0.47) for the data set Sph_6_2 (K = 7).

Figure 5.1 to Figure 5.6 shows the comparison of data partitioning by the proposed method FRC-NSGA-II with the original data partitioning for six data sets: AD_5_2, AD_10_2, Square-1, Square-4, Sph_5_2 and Sph_6_2. The comparison shows that FRC-NSGA-II provides better performance for compactness, separation and overlapping of clusters of the data sets. From the figures above, it is shown FRC-NSGA-II works effectively for the well-separated data points of AD_10_2 (Figure 5.2) and Square-1 (Figure 5.3). Data points may be misclassified by overlapping where borderline between clusters contains the data points.

Analyzing the results of different multi-objective techniques it is found that FRC-NSGA-II provides better accuracy than other techniques which is obvious by the Minkowski Score (MS) value and the optimal cluster number as shown in Table 5.2 and Figure 5.1 to Figure 5.6. Here, Minkowski Score is used to obtain the best solution from the non-dominated solution set. Experimental results on benchmark data sets demonstrate that the FRC-NSGA-II is capable of determining well-separated, hyperspherical and overlapping clusters.

5.2.2 Gene Expression Data Sets

- **Gene Expression Data Set: Leukemia**

For Leukemia data set, the proposed method was compared with SVM[42], Naive Bayes(NB)(Lytvynenko, 2014)[42], KNN[43], GA+KNN[44], Decision Tree[45] and others existing methods like Li et *al*. 2000, Furey et *al*.(Cho & Won, 2003)[46], Dudoit et *al*.[47]. Figure 5.7 shows that FRC-NSGA-II provides clustering solution with higher accuracy than other methods.

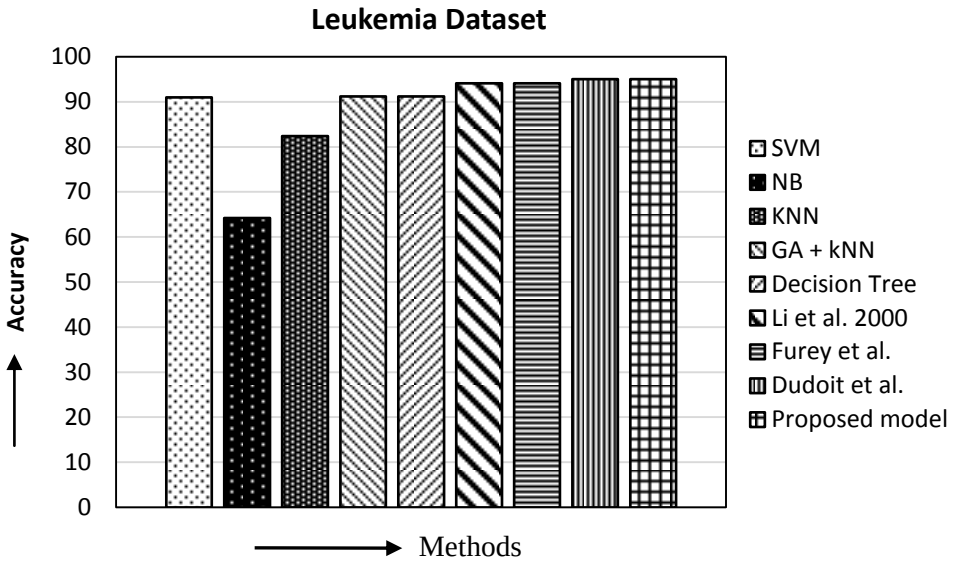


Figure 5.7: Performance evaluation of the proposed method with respect to existing methods on Leukemia data set.

- **Gene Expression Data Set: Lymphoma**

For Lymphoma data set, the proposed system was compared with Decision Tree(Hijazi & Chan, 2013)[48] and others existing methods like Elena[49], Li et *al*. (KNN)[50], Dudoit et *al*., Li et *al*. (Cho & Won, 2003)[46], ELM(R. Zhang, Huang, Sundararajan, & Saratchandran)[51]. Figure 5.8 shows that FRC-NSGA-II provides more accurate clustering solution than other methods except ELM method.

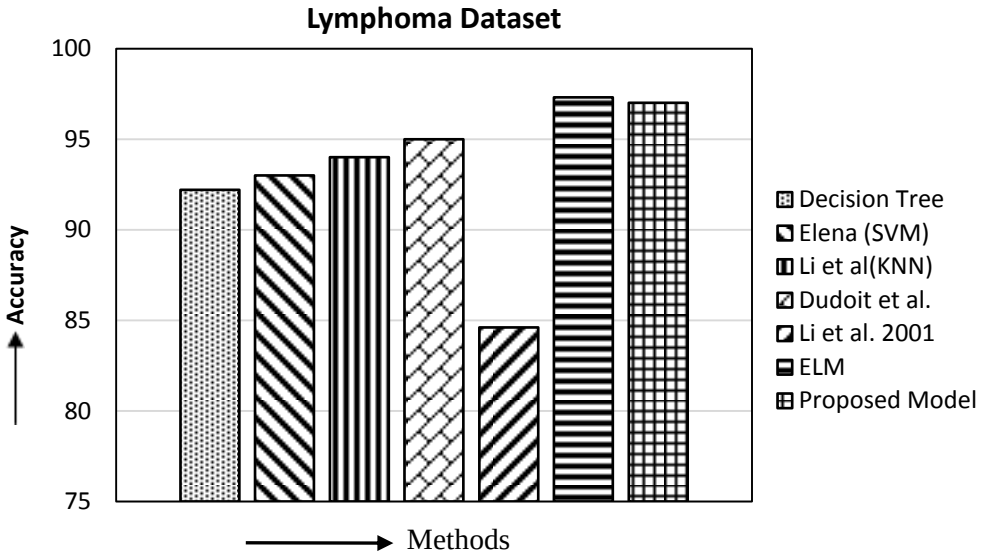


Figure 5.8: Performance evaluation of the proposed method with respect to existing methods on Lymphoma data set.

- **Gene Expression Data Set: Prostate Tumor**

For Prostate Tumor data set, the proposed system was compared with Decision Tree(A. C. Tan & Gilbert, 2003)[45], SVM(G. Cong, K. Tan, A. K.H. Tung, & Xin Xu, 2005)[52], KNN(Singh et *al*., 2002)[41], Bayes Network, Naive Bayes classifier(NB)(Lytvynenko, 2014)[42]. Figure 5.9 shows that FRC-NSGA-II performs clustering with higher accuracy than other methods.

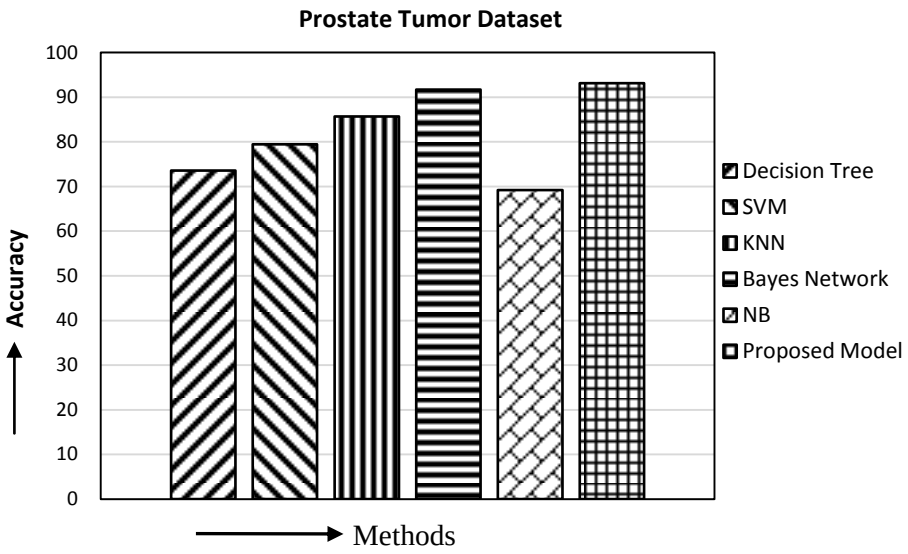


Figure 5.9: Performance evaluation of the proposed method with respect to existing methods on Prostate tumor data set.

- **Gene Expression Data Set: Colon cancer**

For colon data set, the proposed method was compared with SVM, Naive Bayes classifier(NB), Bayes Network(Lytvynenko, 2014)[42], KNN[43] and other existing methods like Ben-Dor et *al*.[53], Dettling et *al*.[54]. Figure 5.10 shows that FRC-NSGA-II performs clustering more accurately than other methods except SVM.

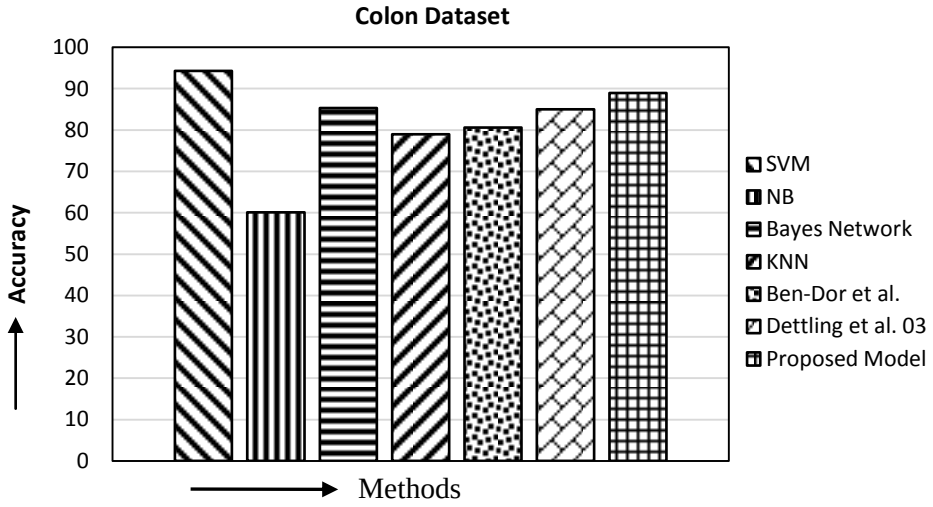


Figure 5.10: Performance evaluation of the proposed method with respect to existing methods on Colon data set.

Figure 5.1 to Figure 5.6 above shows the comparison of classification accuracy by FRC-NSGA-II with other methods for four gene expression data sets. The comparison shows that the proposed model performs more accurate clustering than most of the other methods.

5.3 Results of Non-dominated Solutions

5.3.1 Non-gene Expression Data Sets

This section shows the non-dominated solutions acquired by FRC-NSGA-II using non-gene artificial and real life data sets. Figure 5.11 to Figure 5.16 shows non-dominated solutions for Iris, Wine, AD_5_2, AD_10_2, Sph_5_2 and Sph_6_2 data sets.

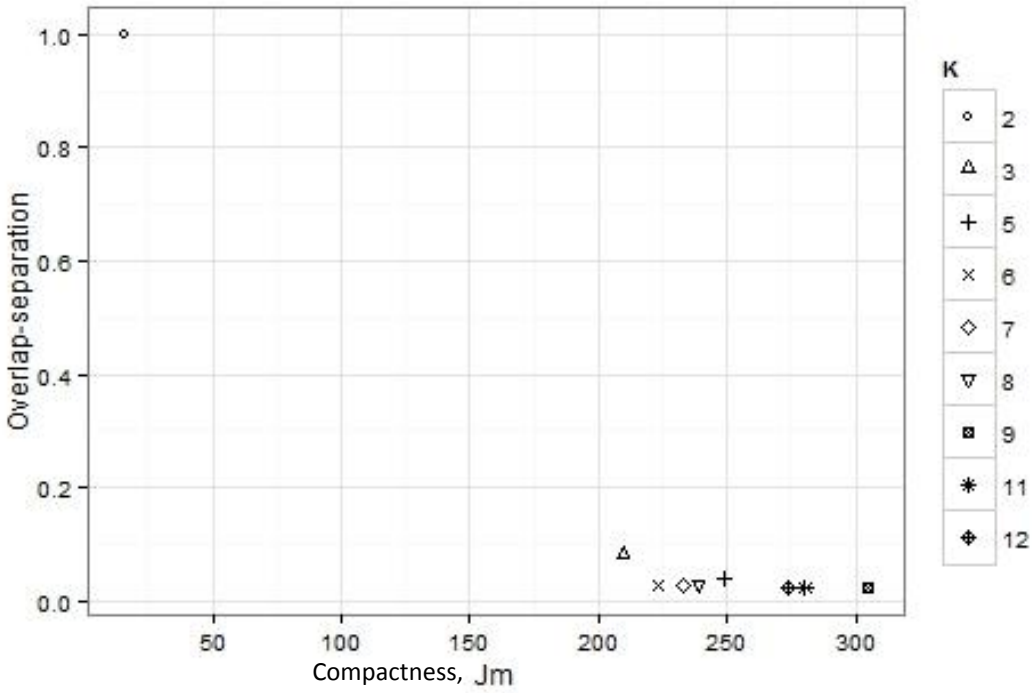


Figure 5.11: Non-dominated solutions found while applying FRC-NSGA-II on Iris data set.

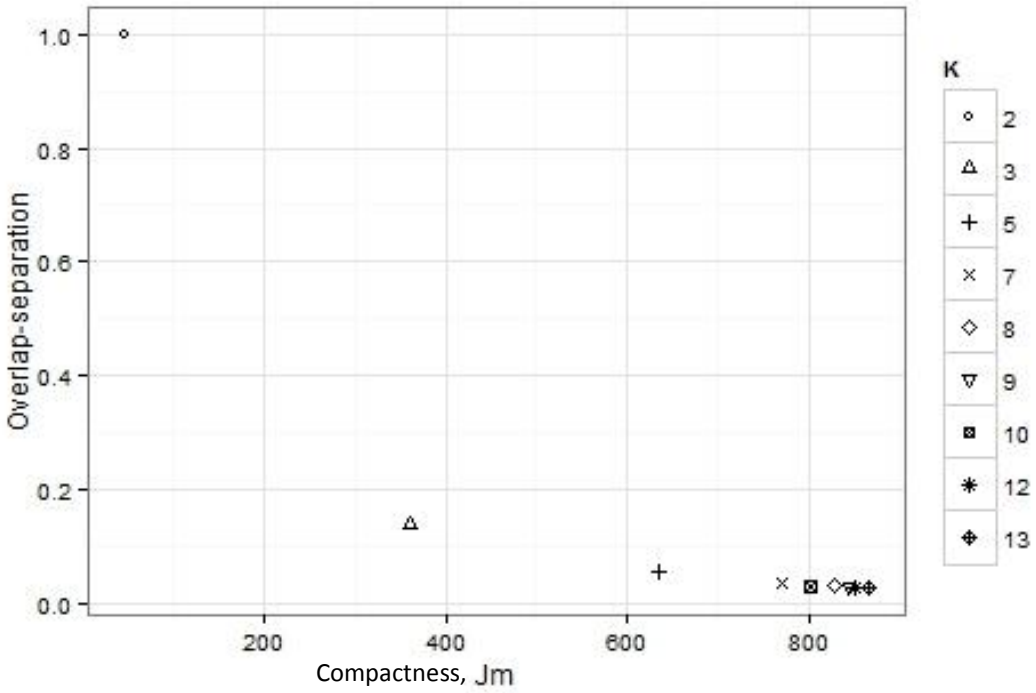


Figure 5.12: Non-dominated solutions found while applying FRC-NSGA-II on Wine data set.

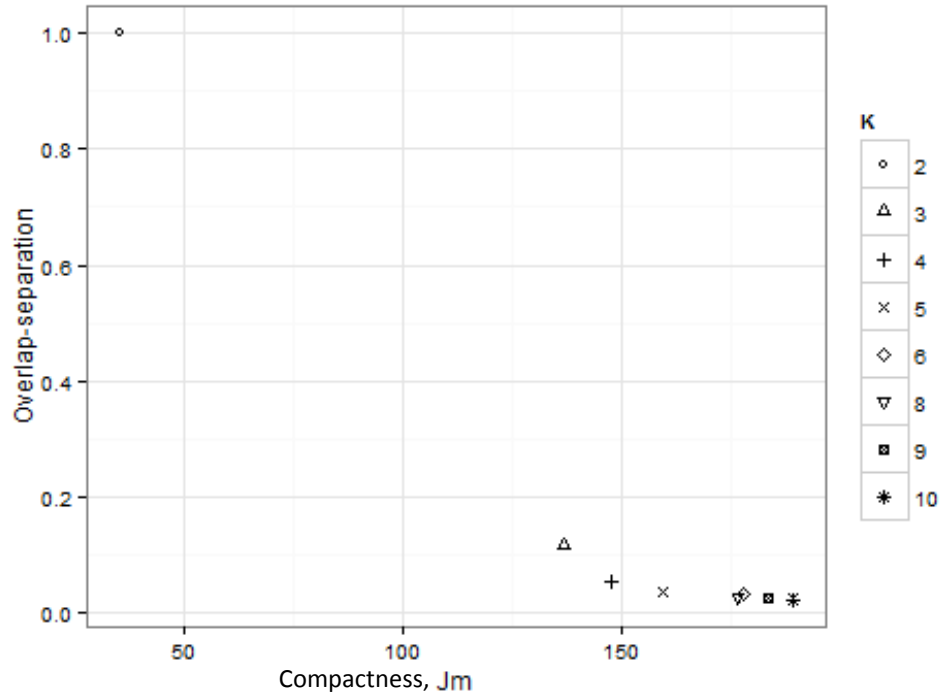


Figure 5.13: Non-dominated solutions found while applying FRC-NSGA-II on AD_5_2 data set.

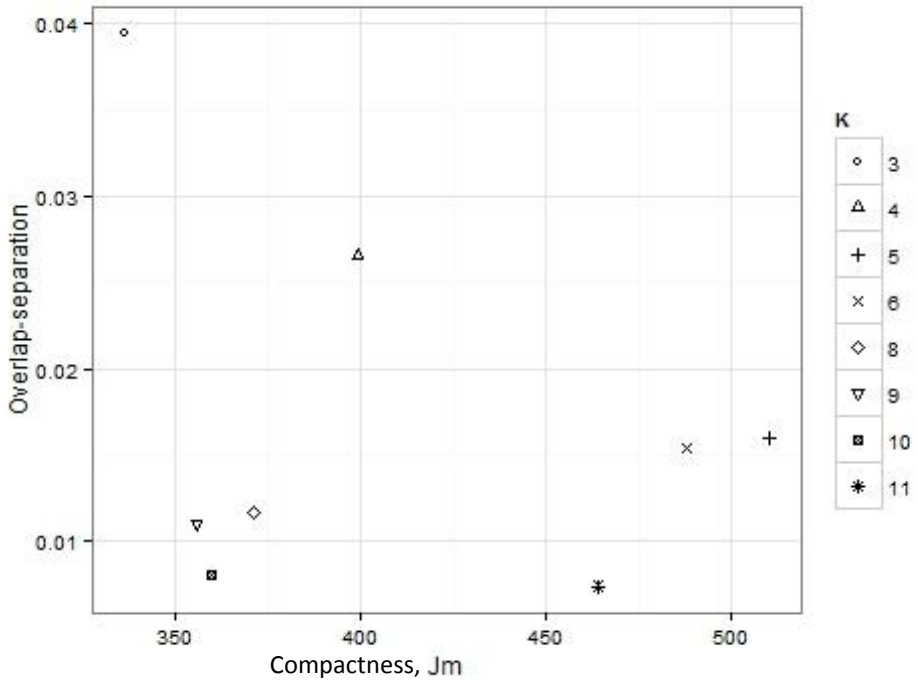


Figure 5.14: Non-dominated solutions found while applying FRC-NSGA-II on AD_10_2 data set.

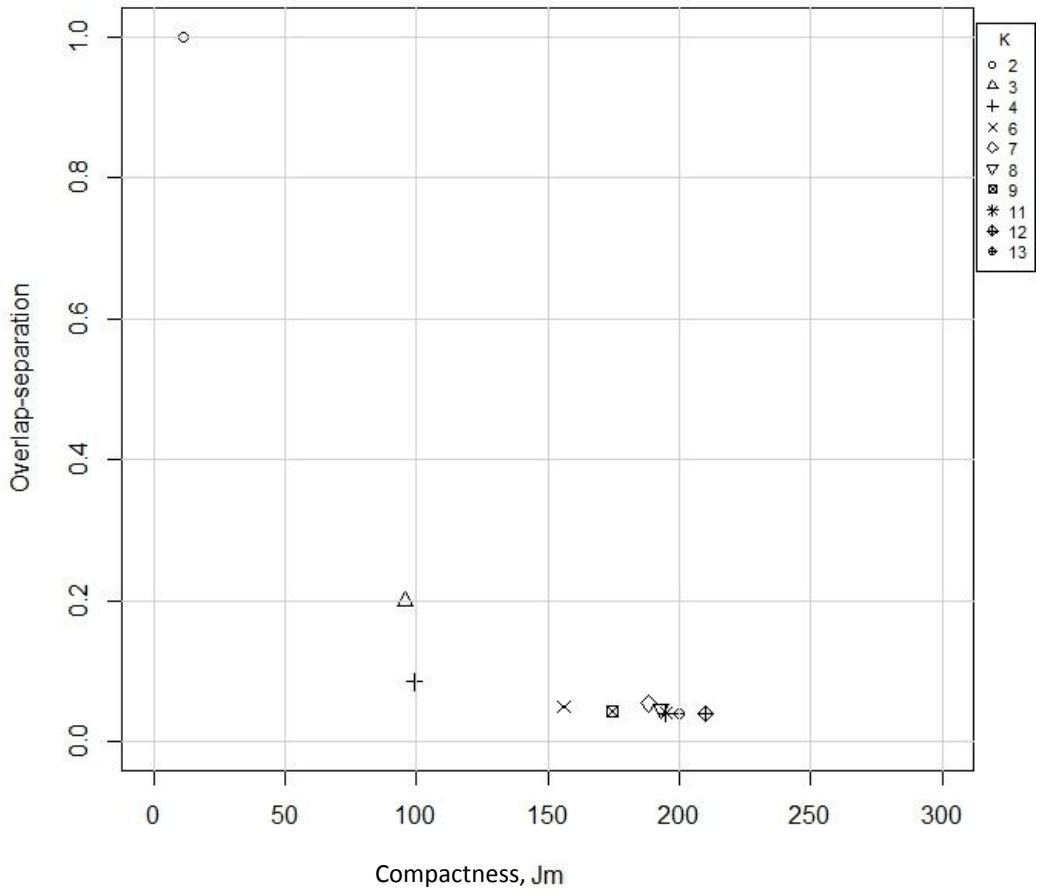


Figure 5.15: Non-dominated solutions found while applying FRC-NSGA-II on Sph_5_2 data set.

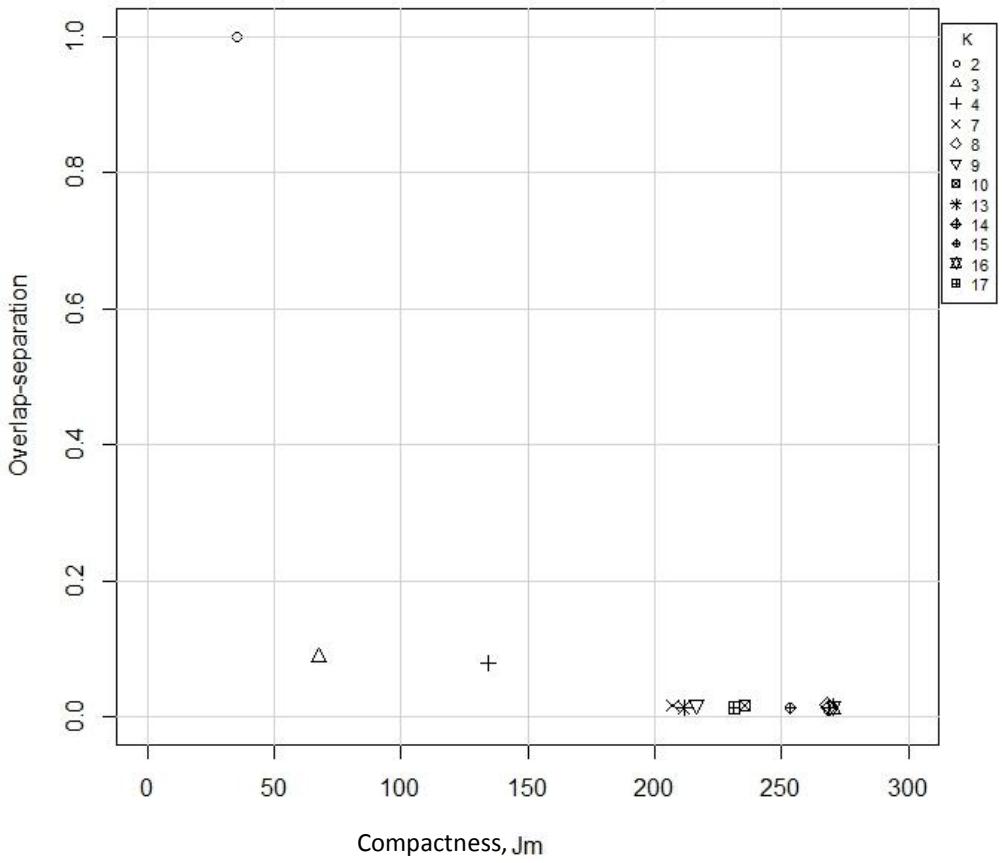


Figure 5.16: Non-dominated solutions found while applying FRC-NSGA-II on Sph_6_2 data set.

5.3.2 Gene Expression Data Sets

In this study, the proposed model was analyzed on four different microarray gene expression data sets. With the different cluster numbers, multiple solutions are generated in multi-objective evolutionary algorithm. In FRC-NSGA-II, it generates a set of non-dominated solutions where it simultaneously optimizes two objective functions: compactness and overlap-separation. A best solution among all the non-dominated solutions is selected by the minimum value of Minkowski Score (MS) [56].

For Leukemia data set, FRC-NSGA-II partitions the data samples into two clusters automatically. Figure 5.17 and Figure 5.18 shows non-dominated solutions for Leukemia and Lymphoma data sets respectively. It is also found that some cluster number is missing for some range of cluster number K. Final solution is selected from the non-dominated solution set by the minimum value of Minkowski score (MS). The MS values obtained for different gene expression data sets are given in Table 5.2.

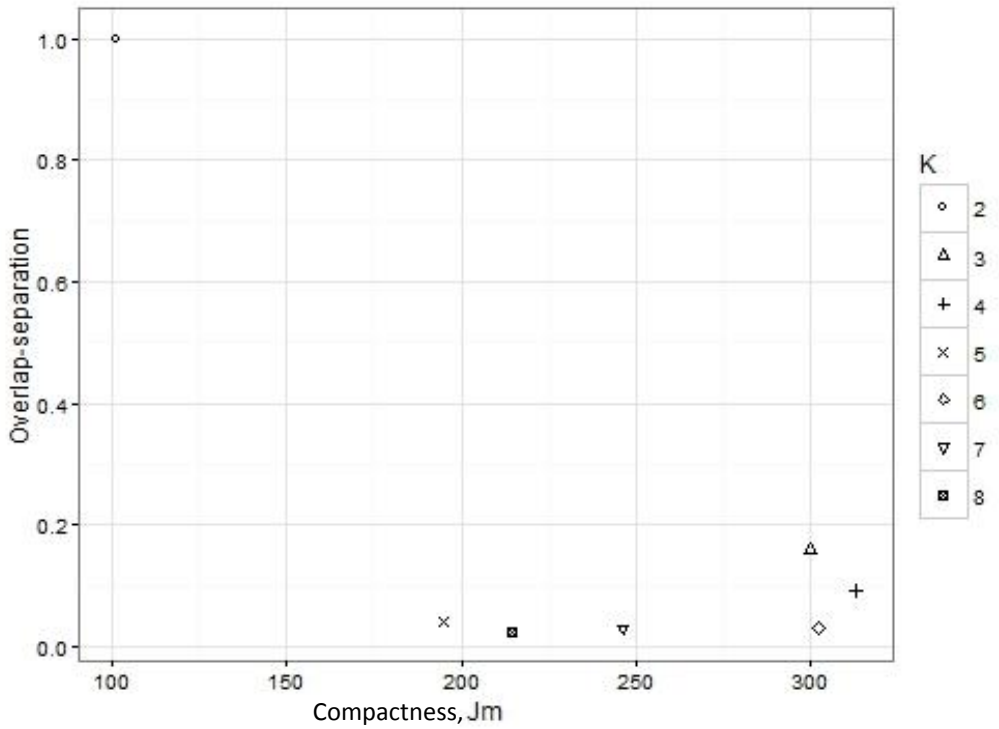


Figure 5.17: Non-dominated solutions found for Leukemia data set while applying FRC-NSGA-II.

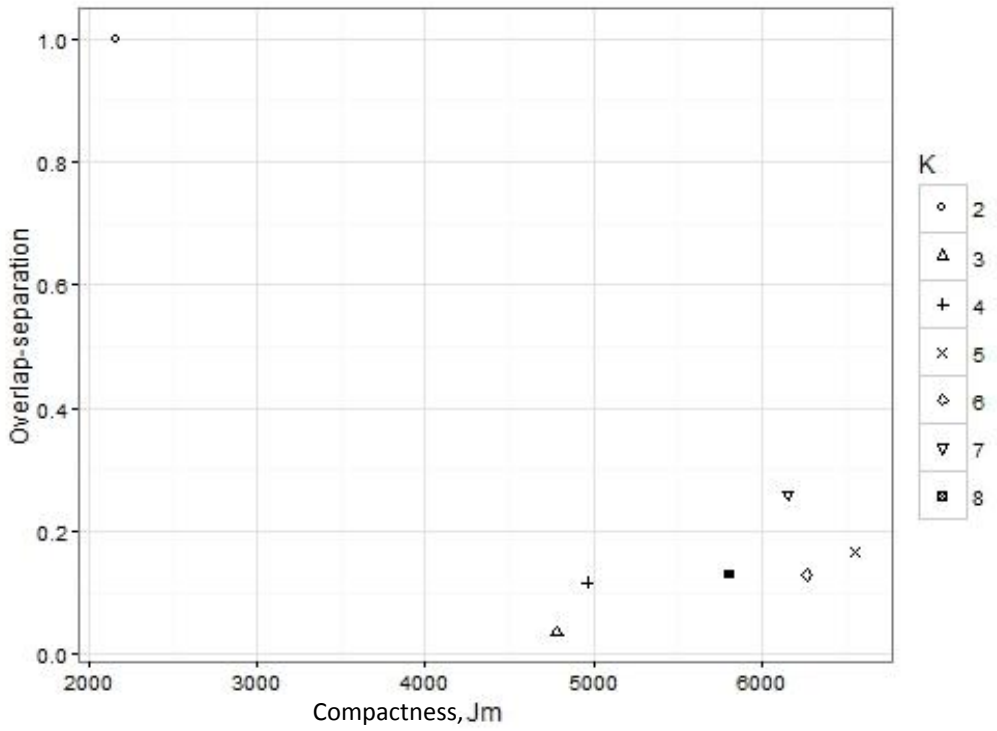


Figure 5.18: Non-dominated solutions found for Lymphoma data set while applying FRC-NSGA-II.

Table 5.2: Minkowski score(MS) values for different gene expression data sets

Data Set	Minkowski score (MS)
Leukemia	0.43
Lymphoma	0.28
Prostate tumor	0.50
Colon	0.60

5.4 Comparative Analysis on Non-gene Expression Data Sets

In order to validate the proposed fuzzy relational clustering technique, FRC-NSGA-II, it is compared with other existing multi-objective clustering approaches VAMOSA [8], MOCK [7], FCM-NSGA [56] and with a single objective clustering approach VGAPS [57]. The comparison is performed in terms of the number of clusters and the values of MS as shown in Table 5.3, where the results of VGAPS, VAMOSA and MOCK are obtained from [8] and the result of FCM-NSGA is obtained from [56]. Table 5.3 shows that the proposed model determines appropriate number of clusters for all the data sets and it provides solution with

minimum MS value for most of the data sets with comparison to other approaches and it performs superior to other approaches for Wine and Iris data sets.

Table 5.3: Comparative analysis of FRC-NSGA-II with other existing methods

DataSet	**Actual K**	**Frc-NSGA-II**		**FCM-NSGA**		**VAMOSA**		**MOCK**		**VGAPS**	
		K	**MS**	**K**	**MS**	**K**	**MS**	**K**	**MS**	**K**	**MS**
AD_5_2	5	5	0.28	5	0.29	5	0.25	6	0.39	5	0.25
AD_10_2	10	10	**0.12**	10	0.13	10	0.43	6	1.01	7	0.84
Square-1	4	4	**0.18**	4	0.18	4	0.19	4	0.19	4	0.20
Square-4	4	4	0.50	4	0.49	4	0.51	4	0.60	5	0.52
Long-1	2	2	**0.0**	2	0.0	2	0.68	2	0.0	3	1.00
Glass	6	6	**0.97**	-	-	6	1.33	6	1.34	6	1.11
Wine	3	3	**0.52**	3	0.93	3	0.97	3	0.90	6	0.97
Iris	3	3	**0.38**	3	0.57	2	0.80	2	0.82	3	0.62
LiverDisorders	2	2	0.98	2	0.98	2	0.98	3	0.98	2	0.98

In Table 5.3, the bold marked values indicate that FRC-NSGA-II generates better results than other methods. Table 5.3 shows that the proposed model generates solutions with lower MS values for the well separated clusters from AD_10_2 (Figure 5.2) and Square-1 (Figure 5.3). The proposed approach generates solutions with higher MS values for highly overlapped clusters of AD_5_2 (Figure 5.1) and Square-4 (Figure 5.4). Single-objective clustering method VGPAS cannot detect overlapping clusters. VGAPS gives solutions with higher MS values for almost all data sets. It gives solution with lower MS value for well separated clusters from Square-1 data set.

MOCK's performance is not well for overlapping clusters. It cannot estimate the number of cluster properly for overlapping clusters (AD_5_2, AD_10_2, iris data sets). Here, the objective functions fail to detect the overlapping structures. VAMOSA shows good performance for different types of data sets except Glass and iris data set. For iris data set, it cannot identify the number of cluster correctly. The performance of VAMOSA is not satisfactory for highly overlapped clusters. Table 5.3 shows that the proposed model provides solutions with minimum MS values for most of the data sets (AD_10_2, Square-1, Long-1, Glass, Wine, Iris, Liver Disorders) with comparison to other approaches and it performs superior to others approaches for AD_10_2, Glass, Wine and Iris data sets.

5.5 Conclusions

This chapter focuses on the performance analysis of the proposed method. The implementation results show that the proposed approach performs better in comparison with other methods for both non-gene and gene expression data sets. It is also shown that the proposed approach performs well in detecting highly overlapped clusters.

CHAPTER VI

Conclusion

6.1 Conclusions

In this work, we propose a new fuzzy relational clustering approach called FRC-NSGA-II based on multi-objective non-dominated sorting genetic algorithm (NSGA-II) when the number of cluster is unknown for (dis)similarity-based data or relational data analysis and this was successfully achieved. This hybrid clustering technique is capable to partition the given data set into groups satisfying two objective functions, compactness and overlap-separation. The proposed multi-objective non-dominated sorting genetic algorithm for relational fuzzy clustering has been statistically assessed over 15 data sets both in gene and non-gene that can also be used to automatically estimate the number of clusters in relational data. The simulation results illustrate the superior capability of this proposed hybrid technique in handling overlapped clusters than other single and multi-objective approaches. On the other-hand, compared with existing clustering approaches for relational data, FRC-NSGA-II is able to capture the underlying structures of the data more accurately and provide richer information for the description of the resulting clusters.

6.2 Future Works

Suggestions for follow-up work that may come after this thesis are:

- As a perspective for future work, the method FRC-NSGA-II can be refined to enable it to explore more possible solutions in shorter execution time by employing parallel computing technology with shared memory to improve the scalability of the algorithm. The local search, fitness evaluation and

mutation procedures are independent among the individuals of the population, so they can be distributed across multiple processors. By doing this, most of the computational burden of the algorithm could be easily made in parallel if the relational data matrix could be shared across the multiple processors.

- This research work also can be extended to design a GAs based method to construct an appropriate fuzzy classification system to maximize the number of correctly classified patterns.

- This work can be extended to implement hierarchical fuzzy relational clustering.

- It can be used to design genetic learning system in which the learning mechanism itself finds an appropriate balance between interpretability and accuracy.

- The work can be extended to improve computational scalability and robustness to deal with noisy data.

REFERENCES

[1] P. Corsini, B. Lazzerini, and F. Marcelloni, "A new fuzzy relational clustering algorithm based on the fuzzy C-means algorithm," *Soft Computing*, vol. 9, no. 6, pp. 439-447, Jun. 2005.

[2] R.N. Dave and S. Sen "Robust fuzzy clustering of relational data," *IEEE transactions on Fuzzy Systems*, vol. 10, no. 6, pp. 713-727, Dec. 2002.

[3] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Algorithms", Plenum, New York, 1981.

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182-197, 2002.

[5] E.E. Korkmaz et *al*., "Combining advantages of new chromosome representation scheme and multi-objective genetic algorithms for better clustering," *Intelligent Data Analysis*, vol. 10, no. 2, pp. 163-182, 2006.

[6] D. Dutta, P. Dutta, and J. Sil, "Clustering by multi objective genetic algorithm," *2012 1st Intl. Conf. Recent Advances in Information Technology (RAIT)*, IEEE, 2012, pp. 548-553.

[7] J. Handl and J. Knowles, "Evolutionary multiobjective clustering," in *International Conference on Parallel Problem Solving from Nature*, Springer 2004, pp. 1081-1091.

[8] S. Saha and S. Bandyopadhyay, "A symmetry based multiobjective clustering technique for automatic evolution of clusters," *Pattern Recognit*., vol. 43, no. 3, pp. 738–751, 2010.

[9] S. Bandyopadhyay and S. Saha, "GAPS: A clustering method using a new point symmetry-based distance measure," *Pattern Recognit.*, vol. 40, no. 12, pp. 3430-3451, 2007.

[10] J. J. T. Valenzuela, "A clustering genetic algorithm for inferring protein-protein functional interaction sites," INSTITUTO TECNOLO´ GICO Y DE ESTUDIOS SUPERIORES DE MONTERREY, Jul. 2009.

[11] E. Falkenauer, Genetic algorithms and grouping problems, John Wiley & Sons, Inc., 1998.

[12] D.G. Bethelmy, "Aspect Mining using Multiobjective Genetic Clustering Algorithms," Doctoral dissertation, Nova Southeastern University, 2016.

[13] R. Caballero, M. Laguna, R. Marti, and J. Molina, "Multiobjective Clustering with Metaheuristic Optimization Technology," Reporte técnico, Departamento de Estadıstica e Investigación Operativa, Universidad de Valencia, Valencia, Espana (2006).

[14] J. Handl and J. Knowles, "Multiobjective clustering around medoids," in *2005 IEEE Congress on Evolutionary Computation*, vol. 1, 2005, pp. 632-639.

[15] M.H.C. Law, A.P. Topchy, and A.K. Jain, "Multiobjective data clustering," *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognit.*, 2004, pp. 424-430.

[16] Peter Peng et *al*., "Reporting and analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data," *Knowledge-Based Systems*, vol. 56, pp. 108-122, Jan 2014.

[17] D. Dutta, P. Dutta, and J. Sil, "Clustering data set with categorical feature using multi objective genetic algorithm," *2012 Intl. Conf. Data Science & Engineering (ICDSE)*, IEEE, 2012, PP. 103-108.

[18] K.P.Malarkodi and S.Punithavathy, "A Fuzzy Based Evolutionary Multi-objective Clustering For Overlapping Clusters Detection," International Journal of Scientific & Engineering Research vol. 2, no. 9, Sep. 2011.

[19] K. Suresh et *al*., "Automatic Clustering with Multi-objective Differential Evolution Algorithms," *2009 IEEE Congress on Evolutionary Computation*, 2009, pp. 2590-2597.

[20] M. Anusha and J.G.R. Sathiaseelan, "Multi-Objective Optimization Algorithm to the Analyses of Diabetes Disease Diagnosis," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 7, no. 1, pp. 485-488, 2016.

[21] M. Anusha and J.G.R. Sathiaseelan, "Evolutionary Clustering Algorithm Using Criterion-Knowledge-Ranking for Multi-objective Optimization," *Wireless Personal Communications*, pp. 1-22.

[22] K.C. Mondal, A. Mukhopadhyay, U. Maulik, S. Bandhyapadhyay, and N. Pasquier, "Simultaneous Clustering and Gene Ranking: A Multiobjective Genetic Approach," *International Conference on Computational Intelligence for Bioinformatics and Biostatistics (CIBB'2010)*, 2010, pp. 104-114.

[23] Mukhopadhyay. U Maulik, and S. Bandyopadhyay, "Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 991-1005, 2009.

[24] J. Du et *al*., "Alternative clustering by utilizing multi-objective genetic algorithm with linked-list based chromosome encoding," *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, Springer Berlin Heidelberg, 2005, pp. 346-355.

[25] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," *Bioinformatics*, vol. 23, no. 21, pp. 2859-2865, 2007.

[26] T. Özyer and R. Alhajj, "Parallel clustering of high dimensional data by integrating multi-objective genetic algorithm with divide and conquer," *Applied Intelligence*, vol. 31, no. 3, pp. 318-331, Dec 2009.

[27] Skabar and K. Abdalgader, "Clustering sentence-level text using a novel fuzzy relational clustering algorithm," *IEEE transactions on knowledge and data engineering*, Vol. 25, no. 1, pp. 62-75, Jan 2013.

[28] Dae-Won Kim, Kwang H. Lee and Doheon Lee, "On cluster validity index for estimation of the optimal number of fuzzy clusters," *Pattern Recognition*, vol. 37, no. 10, pp. 2009-2025, Oct 2004.

[29] Hoel Le Capitaine and Carl Frelicot, "A Cluster-Validity Index Combining an Overlap Measure and a Separation Measure Based on Fuzzy-Aggregation Operators," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 3, pp. 580-588, Jun 2011.

[30] James C. Bezdek, Robert Ehrlich and William Full, "FCM: The fuzzy *c*-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.

[31] H.L. Capitaine and C. Frélicot, "A Family of Cluster Validity Indexes Based on a l-Order Fuzzy OR Operator," *Joint IAPR International Workshops on Structural and Syntactic Pattern Recognition (SSPR)*, Dec 2008, pp. 612-621.

[32] L. Mascarilla, M. Berthier and C. Frélicot, "A k-order fuzzy OR operator for pattern classification with k-order ambiguity rejection," *Fuzzy Sets and Systems*, vol. 159, no. 15, pp. 2011-2029, Aug 2008.

[33] K. Deb and R.B. Agrawal, “Simulated binary crossover for continuous search space,” Complex Systems, vol. 9, no. 2, pp. 115-148, 2002.

[34] K. Deb and M. Goyal, “A Combined Genetic Adaptive Search (GeneAS) for Engineering Design,” *Computer Science and Informatics*, vol. 26, pp. 30-45, 1996.

[35] K.S.N. Ripon and N. Kwong, “A real-coding jumping gene genetic algorithm (RJGGA) for multiobjective optimization,” *Information Sciences*, vol. 177, no. 2, pp. 632-654, Jan 2007.

[36] Dong-Xia Chang, Xian-Da Zhang and Chang-Wen Zheng, “A genetic algorithm with gene rearrangement for K-means clustering,” *Pattern Recognition*, vol. 42, no. 7, pp. 1210-1222, Jul 2009.

[37] S. Bandyopadhyay and U. Maulik, “Genetic clustering for automatic evolution of clusters and application to image classification,” *Pattern Recognition*, Vol. 35, pp. 1197-1208, 2002.

[38] S. Bandyopadhyay and S. K. Pal, *Classification and Learning Using Genetic Algorithms: Applications in Bioinformatics and Web Intelligence*, Springer, 2007.

[39] TR Golub et *al*., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, pp. 531-537, Oct 1999.

[40] Michael C O'Neill and Li Song, “Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect,” BMC bioinformatics, 4, 13, 2003.

[41] D. Singh et *al*., “Gene expression correlates of clinical prostate cancer behavior,” *Cancer Cell*, vol. 1, no. 2, pp. 203-209, Mar 2002.

[42] V. Lytvynenko, “Hybrid swarm negative selection algorithm algorithm for DNA-microarray data classification,” Academic Journals & Conferences, *Computer Sciences and Information Technologies*, Vol. 800, pp. 134–148, 2014.

[43] L. Sun, D. Miao, and H. Zhang, “Gene Selection with Rough Sets for Cancer Classification,” *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, IEEE, Vol. 3, 2007, pp. 167-172.

[44] M. Banerjee, S. Mitra, and H. Banka, “Evolutionary Rough Feature Selection in Gene Expression Data,” *IEEE Transactions on Systems, Man,*

and Cybernetics, Part C (Applications and Reviews), Vol. 37, no. 4, pp. 622-632, Jul 2007.

[45] A.C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Appl. Bioinformatics;2(3 Suppl)*, 2003.

[46] Sung-Bae Cho and Hong-Hee Won, "Machine learning in DNA microarray analysis for cancer classification**,**" *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, Vol. 19, 2003, pp. 189-198.

[47] S. Dudoit, J. Fridlyand, and P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American statistical association*, vol. 97, no. 457, pp. 77-87, Mar 2002.

[48] H. Hijazi and C. Chan, "A classification framework applied to cancer gene expression profiles," *Journal of healthcare engineering*, Vol. 4, no. 2, pp. 255-283, 2013.

[49] E. Marchiori and M. Sebag, "Bayesian learning with local support vector machines for cancer classification with gene expression data," In *Workshops on Applications of Evolutionary Computation*, Springer, pp. 74-83, 2005.

[50] Li-Yeh Chuang, Chao-Hsuan Ke, Hsueh-Wei Chang, and Chang-Hsuan Ho, "A Two-Stage Feature Selection Method for Gene Expression Data," *OMICS A journal of Integrative Biology*, Vol. 13, no. 2, pp. 127-137, 2009.

[51] R. Zhang, G. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, Vol. 4, no. 3, pp. 485-495, Jul 2007.

[52] G. Cong, K. Tan, A.K.H. Tung, and X. Xu, "Mining top-k covering rule groups for gene expression data," *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, ACM, Jun 2005, pp. 670-681.

[53] Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *Journal of computational biology*, Vol. 6, no. 3-4, pp. 281-297, 1999.

[54] M. Dettling and P. Buhlmann, "Boosting for tumor classification with gene expression data," *Bioinformatics*, Vol. 19, no. 9, pp. 1061-1069, 2003.

[55] U. Alon et *al*., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*," vol. 96, no. 12, pp. 6745-6750, 1999.

[56] Siripen Wikaisuksakul, "A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering," *Applied Soft Computing*, vol. 24, pp. 679-691, Nov 2014.

[57] S. Bandyopadhyay and S. Saha, "A Point Symmetry-Based Clustering Technique for Automatic Evolution of Clusters," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, no. 11, pp. 1441-1457, 2008.