Design a new Multiobjective Algorithm for Data Clustering

Project Report submitted in partial fulfillment of the requirement for the degree of

Master of Technology

in

Computer Science & Engineering

Under the supervision of

Dr. Yugal Kumar and Dr. S.P.Ghrera

By

Kshamta Chauhan

172205

Jaypee University of Information Technology

Waknaghat, Solan – 173234, Himachal Pradesh

CERTIFICATE

This is to certify that project report entitled **“Design a new Multiobjective Algorithm for Data Clustering”,** submitted by **Kshamta Chauhan** in partial fulfillment for the award of degree of Master of Technology in Computer Science & Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been made under my supervision.

This report has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date:	**Supervisor’s Signature** **Dr. Yugal Kumar** **Assistant Professor (Senior Grade)**
Date:	**Co-Supervisor’s Signature** **Dr. Satya Prakash Ghrera** **Professor, Brig (Retd.)** **Head, Dept. of CSE and IT**

ACKNOWLEDGMENT

I take this opportunity to acknowledge all who have been a great sense of support and inspiration throughout the project work. There are lots of people who inspired me and helped, worked for me in every possible way to provide detail about various related topics, thus, making of report work success. My first gratitude goes to our head of department **Dr. S.P. Ghrera** for his guidance, encouragement and support.

I am very grateful to **Dr. Yugal Kumar** (Asst. Professor), for all his diligence, guidance, and encouragement and helped throughout the project work. I also thank him for the time that he spared for me, from his extremely busy schedule. His insight and creative ideas are always the inspiration for me during the dissertation work.

Date: Signature:

Kshamta Chauhan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBRIVATIONS

PSO	Particle Swarm Optimization
MOVPS	Multiobjective Vibrating Particle System
GA	Genetic Algorithm
DE	Differential Evolution
MOPSO	Multiobjective Particle Swarm Optimization
MOO	Multiobjective Optimization
XB	Xie and Beni
AMOSA	Archived Multiobjective Simulated Annealing
MOITLBO	Multiobjective Improved Teaching-Learning-Based Optimization Algorithm
ITLBO	Improved Teaching-Learning-Based Optimization Algorithm
DB	Davies Bouldin
GD	Generational Distance
C	Convergence
SD	Space metric
MS	Maximum spread
HV	Hyper Volume
VPS	Vibrating Particle System
GP	Good particle
BP	Bad particle
HB	Historical Best Position
MABC	Multiobjective Artificial Bee Colony

ILMOFCCD	Incremental Learning Based Multiobjective Fuzzy Clustering for Categorical Data
SSE	Sum of Square Error
WBC	Breast Cancer Wisconsin
NSGA II	Elitist Non-Dominated Sorting Genetic Algorithm

ABSTRACT

Clustering is an unsupervised learning technique. It is a collection of objects that are grouped together on the basis of distance measure. As the number of population increases the data is also increasing, so we need to organize this data based on their similarities. The problem in clustering is single-objective because due to vast data, results are not accurate and performance are not that much good. In this project, clustering is seen as multi-objective rather than single-objective. In multi-objective clustering, more than one objective is optimized simultaneously and aim of multi-objective is to improve the performance of data clustering. Vibrating Particle System (VPS) algorithm is used for optimization in multiobjective clustering. The results of the multiobjective clustering algorithm are more accurate than that of the single-objective algorithm. Two objectives that are optimized is compactness and connectedness. The first objective is intra-cluster variance, we have to compute the distance of the object to the nearest cluster center and we can also call that overall deviation of a partitioning. The second objective is connectedness of the cluster, neighboring data objects have to identify whether they belong to the same cluster or not. Using these two objectives we will try to achieve a more accurate result, better performance, and efficiency.

CHAPTER 1

INTRODUCTION

Clustering is an unsupervised learning technique. It is a collection of objects that are grouped together on the basis of distance measure [1, 18]. As the number of population increases, the data is also increasing, so we need to organize this data based on their similarities. A good clustering algorithm identifies the arbitrary shape of clusters. Various clustering methods have developed various clusters on the same dataset.

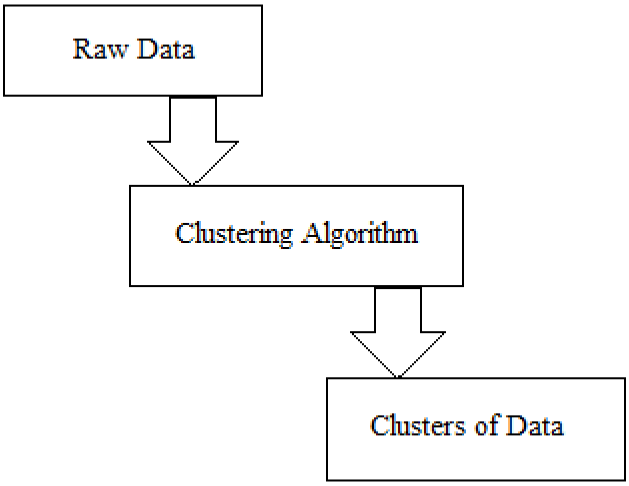


Figure 1.1: Stages of Clustering

Mainly, there are two types of attributes related to the input data in clustering algorithm i.e. numerical and categorical attributes. The numerical attribute has a finite or infinite number of structured values such as the age of a person and categorical attribute have finite unstructured values such as the occupation of a person.

A good cluster should be [1]:

1. **Intra-cluster distance:** Sum of the distance between the object in the same cluster should be minimum.
2. **Inter-cluster distance:** Distance between two clusters should be maximum.

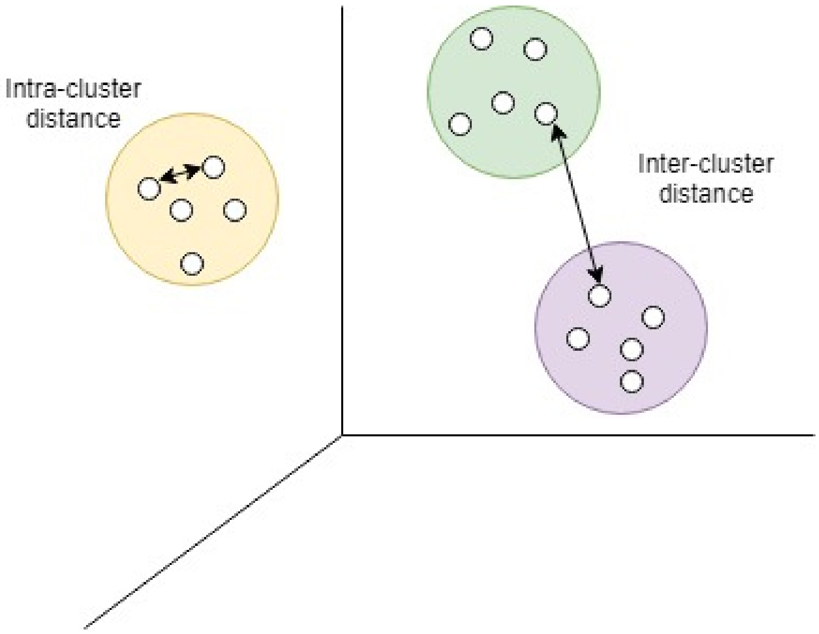


Figure 1.2. shows intra and inter-cluster distance

1.1 Characteristics of Clustering [28]:

1. **Well separated clusters:** Any point in a cluster should be close to any other point in the same cluster as relate to any other point i.e. not present in the cluster.
2. **Contiguous clusters:** A point in a cluster should be closest to all the other point in a cluster rather than all point that are not present in the cluster.
3. **Centred-based clusters:** An object in a cluster is closest to its centroid of the cluster as relate to another cluster centroid.
4. **Shared property or Conceptual clusters:** Make a group of clusters that share the same functionality.
5. **Density based clusters:** In a density based cluster, clusters are formed according to the low-density region and high-density region.

1.2 Requirements for clustering algorithm [18]:

1. **Scalability:** The clustering algorithm must be compatible to work on a large amount of data.

2. **Ability to deal with noisy data:** Database has missing, noisy and error data. Some clustering algorithm is delicate about this data and generates poor quality of clusters.

3. **Detection of the arbitrary shape of clusters:** Clustering algorithm should have to find clusters of arbitrary shape. Distance measure should not only be constrained to make a cluster. The algorithm should find rounded clusters of small amount.

4. **High dimensionality:** Clustering algorithms must be capable to handle both high and low dimensionality of data.

5. **Interpretability:** Cluster result should be interpretable, usable and comprehensible.

1.3 Types of Clustering [16, 17, 18]:

1. **Partitioned base clustering:** In this clustering type, it divides the data into many subsets. Each subset or partition will show a cluster and k≤n. The data are classified into k groups and every group consists of at least one object.

2. **Hierarchical clustering:** This types of clustering build a hierarchy of clusters. The hierarchical method is classified on the base of how hierarchical decomposition is formed. Hierarchical clustering is classified into two types:

   a) **Agglomerative approach:** This approach follows a bottom-up approach. In this, a separate group is formed by each object. It merges the groups that are nearest to each other and carry on doing until all the groups are combined into one.

   b) **Divisive approach:** This approach follows the top-down approach. All the objects in this approach are under the same cluster. And in every loop, cluster is divided into sub clusters until each and every object in a single cluster.

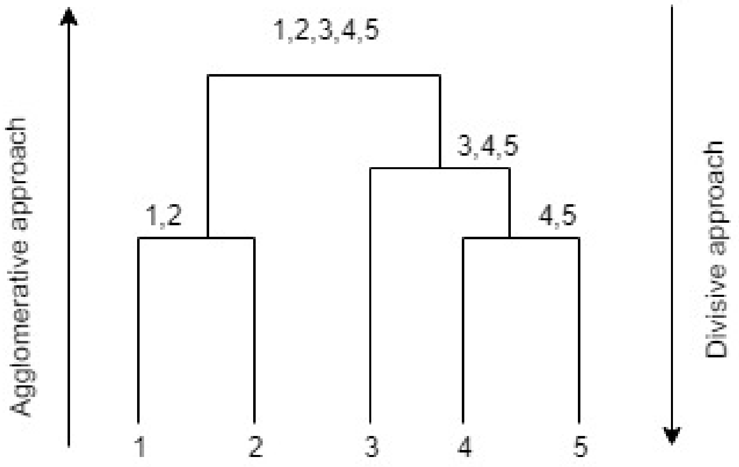


Figure1.3 shows the agglomerative and division approach

3. **Density-based clustering:** In this type, a higher density area formed a cluster than remaining of the dataset. Objects that are placed in those areas are recognized as noise. Using local distribution of nearest neighbours both connectivity and density is measured.

   - **Core points:** Core point are those points that are present around the center of the cluster and it has more than the stated number of points within the defined radius.

   - **Border points:** It is in the neighbourhood of core points and has less than a specified number of points within the radius.

   - **Noise point:** Those point that are neither come under border point nor in a core points.

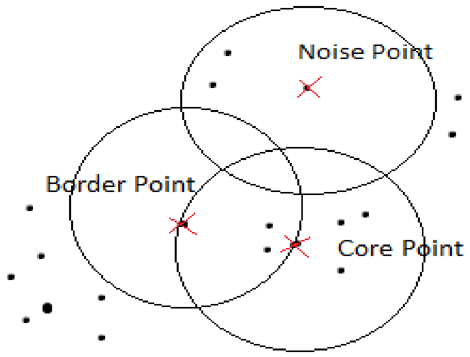


Figure 1.4. Shows density based clustering

4. **Grid based clustering:** In this clustering type, data area is quantized towards a fixed number of cells and those cells formed a grid structure and later clustering is performed on that grids. Grid based clustering processing is very fast and it depends on the capacity of the grid instead of the capacity of data.

5. **Model-based clustering:** In model-based clustering, it finds the best fit of the data and hypothesized it for each cluster for a given model. By using density function it locates the clusters and clustered them. It is a very robust clustering method.

1.4 Application of clustering [29]:

1. **Marketing:** Clustering is used by market researchers to partition the general population of consumers into market section so that they can understand the relationship between distinct groups of the customer and use it in new product development, market segmentation and for selecting test markets.

2. **Social science:** Clustering can be used in the criminal analysis for identifying the same type of crime happened in over a period of time. It is used to analyze areas where the crime happened on a greater scale.

3. **Wireless sensor network:** Clustering plays a very big role in the wireless sensor network to solve the energy efficiency problem and to enhance network lifetime.

4. **Medicine:** In medicine, clustering can be used to IMRT segmentation, medical imaging and for analysis of antimicrobial activity.

5. **Bioinformatics:** Clustering is used in biology and bioinformatics to find similarity of human genetic data and also used to group homologous sequence into some gene families.

1.5 Single objective clustering:

In single objective clustering, only one validity index is optimized as an objective function. But for a large amount of data, single validity indices does not work properly for all dataset because of the different data property. Thus, it is mandatory to optimize two or more objectives together to capture the distinct properties of the data.

1.6 Multiobjective clustering:

For solving many problems in the real world there is a need for multiobjective optimization in which we can optimize them simultaneously [1]. In this, there exist more than one objective values so it is normally difficult to compare one solution with another. This is very difficult to find a single solution for all the objective functions and to define optimality for these problems. Generally, these problems have multiple solutions when the importance of the objective is anonymous the problem is recognized as acceptable and equivalent. Aim of multi-objective is to enhance the performance of data clustering. The outcomes of the multiobjective clustering algorithm are more accurate than that of single objective algorithm.

1.6.1 Particle swarm optimization [1, 13]:

PSO is a random generation optimization method dependent on population like attitude of a fish in search of food in delusional adaption in schools or a bird in a flock. To find the global optimal solution, particle swarm optimization has a population of particle swarm close to best field of search space. A population is initialized with the random solution and renew generations by inspecting for optima. In PSO there is not any evolution operator like crossover or mutation. Particles, also called as potential solutions, travel via problem space and follows the particles who are currently optimum. In problem space, every particle has a record of its location inside the problem space which dependent on the optimum solution attained by it. It is known as pbest value. Lbest is the best value attained by any of its neighbouring particle. Gbest is a best value which is attained when whole population in particle's topology is of its neighbours. Velocity towards lbest and pbest locations changes every time. By a random term, acceleration is weighted, and independent numbers are generated towards pbest and lbest locations. There are some steps in figure 1.5.

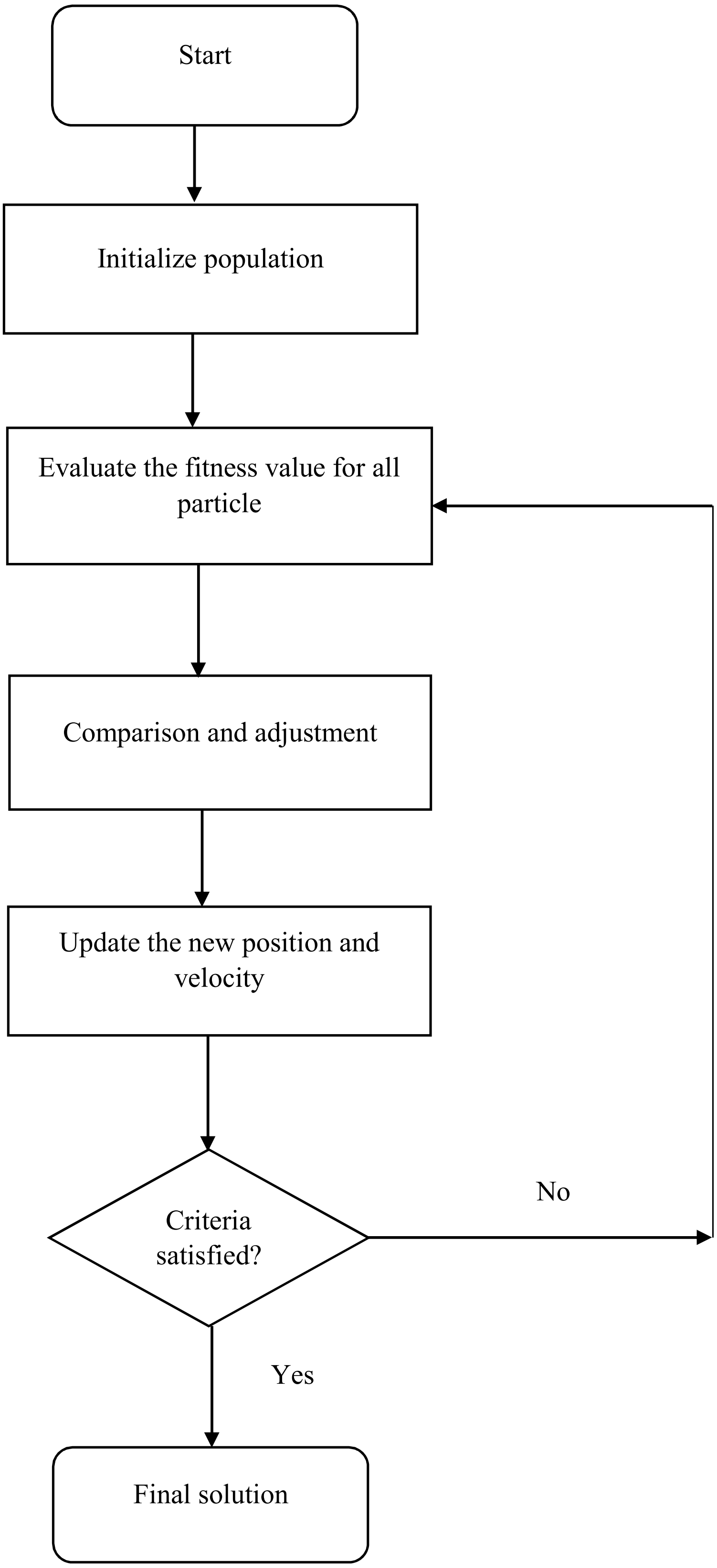

Start
Initialize population
Evaluate the fitness value for all particle
Comparison and adjustment
Update the new position and velocity
Criteria satisfied?
No
Yes
Final solution

Figure 1.5. Flow chart of PSO

1.6.2 Genetic Algorithm [15]:

GA is called a modifying heuristic search algorithm. It is based on the idea of natural selection and genetic. This algorithm is used to achieve a great trait of solution for optimization and search space. It is a bio activated operator like mutation, selection and crossover. In figure 1.6 there are some steps of the genetic algorithm.

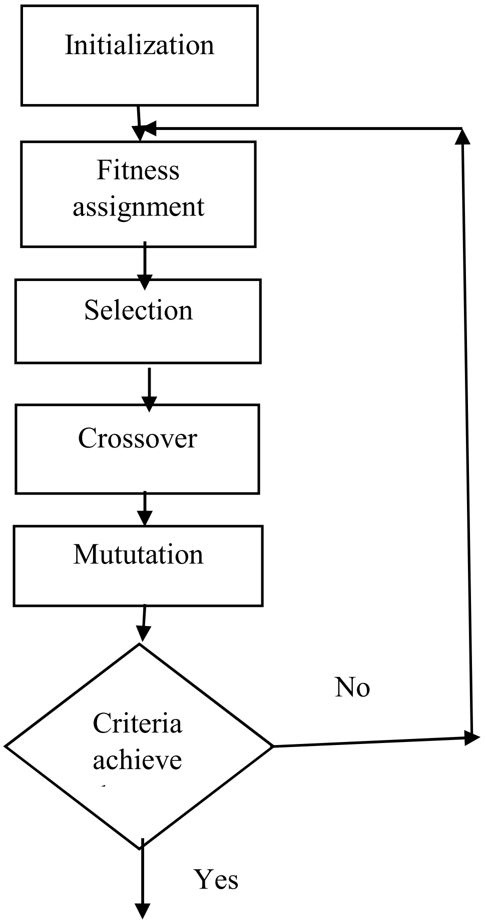


Figure 1.6: Flow chart of Genetic algorithm

1.6.3 Differential evolution [12]:

Differential evolution is a part of an evolutionary algorithm. It is used for an optimization problem over the repeated domain. Each variable in DE is defined in the chromosome by the real number. Differential evolution is part of subclasses of the genetic algorithm which uses the exact operation of mutation, crossover and selection. These operations are applied to a population to reduce the objective function over the course of succeeding generation.

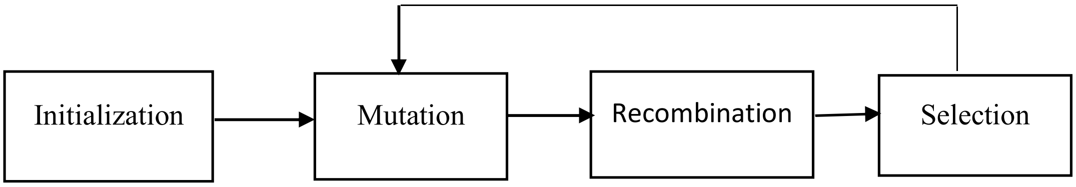


Figure 1.7: Evolutionary Algorithm Procedure

CHAPTER 2
LITERATURE REVIEW

2.1 Literature review of multiobjective clustering

Prakash et al. [1] proposed an approach for hard partitional clustering using multiobjective. TSMPSO was introduced to improve exploratory capability of the MOPSO by introducing crossover operator in Genetic algorithm. In this work, compactness and connectedness are considered as objective functions. The simulation results are taken on seven real data sets. It is observed that proposed TSMPSO gives better accuracy.

Bandyopadhyay et al. [2] proposed an algorithm using fuzzy clustering with validity and stability based on multiobjective. It has optimized each and every appliance separately by using MOO because combining different appliances into one is difficult and sometimes it is unsuited. The results shows that it performed better than other in terms of capability to find appropriate clusters and match with true cluster.

Fabre et al. [3] proposed an enhanced and more expandable evolutionary algorithm for multiobjective clustering. In this paper, k-determination approach with automatic and enhancement of multiobjective clustering algorithm was proposed. This approach enhance the scalability of predecessor in several ways, main adjustment are linked to the use of specialized initialization routine and two alternative reduced-length representation. In this study, author analysed and found that new solution reduced the overhead, improve the capability and overall performance of process.

Saha et al., [4] designed a clustering approach for automatic development of clusters based on symmetry multiobjective. Authors introduce a balance based multiobjective automatic clustering. In this paper, two cluster validity indices, XB index, Euclidean distance and newly grew point symmetry distance establish cluster validity index and Sym-index is optimized together. This algorithm detect the cluster having point symmetry property. The designed algorithm was applied on six real-life dataset and seven artificial dataset and identify the good number of cluster and partition the dataset.

In this paper, author also proposed new semi-supervised approach to choose only solution from concluding Pareto optimal front of the aimed algorithm.

Bara'a Ali Attea [5] designed a fuzzy multiobjective algorithm for efficient clustering. In this paper, the author proposed fuzzy multi objective PSO in an inventive way for clustering. In this study, author resolve the confusion in the dataset that how object that share same functionality to more than one cluster and then for this, various test are performed in numerical and categorical real life data sets. The simulation results present that designed algorithm perform better than other in terms of both effectiveness and efficiency.

Saha et al. [6] proposed a multiobjective approach for automatic clustering algorithm. An advanced multiobjective clustering technique named generalized automatic clustering algorithm was designed. In this work, author take three objective functions such as total compactness established on Euclidian distance, total symmetry of the cluster and connectedness are optimized simultaneously using AMOSA. This proposed algorithm identify the correct number of cluster and correct partitioning from dataset having either well-classified clusters or having point symmetric clusters.

Dutta et al. [7] designed the two version of new absolute coded combination elitist MOGA for MOGA_Feature_Selection (Homogeneity, Separation) and K-clustering (MOGA (Homogeneity, Separation)). Three benchmark dataset is selected by author from any dataset repository that contain only categorical features. In this work, author achieve data mining function of clustering by identifying modes of clusters and consider categorical features only.

Saha et al. [8] give multiobjective fuzzy clustering for categorical data based on incremental learning. In this paper, author designed a fuzzy algorithm for clustering that are used on categorical data based on incremental learning using multiobjective. This algorithm simultaneously optimize two contradictory objectives like J_m and XB index. The Pareto-optimal solutions are applied in this learning to obtain the group training set to divide the remaining points by applying stochastic forest classifier. Statistical test are performed to make effectiveness on proposed ILMOFCCD technique and this algorithm does better job than other existing algorithm. This new designed technique is used in many areas of prediction, web mining, biometry and forecasting etc.

Saha et al. [9] proposed a multiobjective optimization technique using feature selection and semi-supervised clustering. In this article, semi-supervised clustering and automatic feature selection using multiobjective optimization problem is resolved. In this work, four objective function are used among which one measures the compactness based on Euclidean distance, second entire symmetry of the clusters then evaluating the resemblance among the accessible labelled information and achieve partitioning and the final one calculating the present features. This proposed algorithm performance is equated with Euclidean distance and obtain that this algorithm is adequate to identify the convenient feature solution and convenient partitioning from datasets.

Kotinis et al. [10] proposed two clustering approach K-medoids and fuzzy adaptation to enhance differential evolution optimizer in multiobjective. Multi-objective optimization algorithm established on differential evolution conception mixed with *K*-medoids clustering and Mamdani-type (FLCs) was proposed. Combining the fuzzy adaptation clustering and K-medoids decrease the consuming time of setting algorithmic parameters and improves the algorithmic performance. This proposed algorithm is tested against many problems to examine the cluster and parameter adaption to check under several condition such as issue of high spatiality, issues of discontinuous Pareto fronts and issue of non-convex Pareto fronts.

Alok et al. [11] designed a novel semi supervised clustering approach based on multi-objective optimization. In this work, multiobjective optimization uses semi-supervised clustering, objective mentioned in this paper have more or less supervised and unsupervised information. In the beginning of three objectives, mean the goodwill of partitioning in conditions of Euclidean distance, entire symmetry in the cluster based on Sym-index and connectedness based on Con-index. The last objective represent distinct external cluster validity indices, adjust rand index, a lately developed minimum-maximum distance based MMI index, Minkowski Score and NMMI index. This algorithm was used on five real-life data and some artificial data sets and give better performance on both unsupervised and supervised information.

Omar S. et al. [12] proposed a self- adaptive differential evolution algorithm for data clustering using c-means multiobjective .The three data clustering validity indices are introduced in multi-objective C-means data clustering algorithm applying Self-adaptive Differential evolution. In this study, three objective are proposed symmetry-index to increase the resemblance among clusters, the compactness index to increase difference within clusters, and the validity Silhouette index to enhance the validity of data clustering. This algorithm was implemented on twenty dataset and perform better other five data clustering algorithm MOSAC-Means, GenClustMOO, MOCK, VGAPS, and GenClustPESA2.

Armano et al., [13] designed multiobjective clustering analysis using particle swarm optimization. In this paper, authors take two objective function and these objective function are applied on automatic grouping of large unlabelled data set. Using connectivity and cohesion we judge the trait of collected cluster and result are well separated, connected and robust as compared to the result of early clustering techniques.

Esfahani et al. [14] proposed a multi-objective approach to fuzzy clustering using ITLBO algorithm. Multi-objective enhances the optimization algorithm based on teaching-learning [14] (MOITLBO) and used to perform compactness and connectedness. These two objectives namely J_m and XB indices using MOITLBO algorithm give better performance as compared to single-objective algorithms. When this algorithm is compared with another multiobjective algorithm it will give the better result in noisy dataset.

Arkeman et al. [15] proposed a multi-objective genetic algorithm based on optimized K-means clustering. In this paper, writer present the improvement of K-means with Pareto front approaches by using a multiobjective genetic algorithm. In this study, author compare the K-means and K-means genetic algorithm that was performed on Iris and Wine data. To find the best solution minimum Davies Bouldin index and the desired number of clusters was used. The simulation results shows that K-Means GA perform better than K-Means and find the minimum index value.

Liu et al. [23] proposed an algorithm that are based on two mutation of synergy for multiobjective fuzzy clustering. In this work, two novel technologies has been proposed compactness validity index that are based on exponential function and synergy of two mutations. The simulation outcomes shows that the new approach outperforms better as measure to other technologies using distinct datasets.

Abbasian et al. [24] presented a gravitational based search algorithm for multiobjective clustering. In this work, the Pareto principle and non-dominated solution front are used to select the archive for elitism. In order to have a suitable trade-off between exploitation and exploration, the archive is grouped into some cluster and that cluster is randomly chosen for each population member as kbest set. The simulation outcome display that CA-MOGSA is a good-trained multiobjective interpretation of GSA when applied on eight basic benchmark.

George et al. [25] proposed a multiobjective approach named cuckoo search for data clustering that are used in medical field. In this paper, objective functions (Sym-Index, Fuzzy DB-Index and Xie and Beni-Index) to measure the fitness for the burrow achieve in fractional cuckoo search, in order to fix the centroids to cluster and to choose the dimensions to cluster the data in the multidimensional dataset. The results of proposed work is that it reduced the dimensionality of the high dimensional data without affecting the clustering accuracy.

Garcia-Piquer et al. [26] proposed a multiobjective technique to retrieve high achievement solution in clustering. This technique designed an automatic retrieval system that handle multiobjective clustering problem on pareto based and also improve pareto quality and shape. This proposed technique minimize the computational time by the individual strategies and improve the accuracy on every solution.

Morik et al. [27] proposed Multi-objective frequent term-set clustering. In this paper, author gives genetic algorithm to solve issue of identifying alternative high-trait architecture for navigation in a big number of high-spatial data. The experimental effect the pareto- optimal solution in which users may select their favourite form of a design for navigation by assembling or search the distinct views given by the distinct optimal solutions.

2.2 Review table for above-mentioned literature work

Table 1.1 shows the literature review

References	Source	Year	Objectives	Methodology	Performance evaluation
Ref [1]	ME (Springer)	2015	In this two objective function is used that is Compactness and connectedness	TSMPSO is used to improve a variety of mechanism in MOPSO	Precision, Recall and F-measure
Ref [2]	IEEE Transactions	2011	Cluster validity and cluster stability are integrated and optimize them simultaneously	Proposed algorithm AMOSA is used for the optimization	Accuracy And XB indices
Ref [3]	IEEE Transactions	2016	It betters its predecessor in various ways, but main changes are related to efficient, specialized initialization routine.	Proposed the bettered version of the multiobjective clustering using automatic k-determination algorithm	Overall performance and convergence behaviour
Ref [4]	ASC	2012	Proper partitioning and good clusters.	A point symmetry based multi-objective	F-measure

				clustering technique	
Ref [5]	ME	2010	Proposed algorithm gives better performance than others and provide better efficiency and effective data cluster.	Proposed multiobjective PSO using fuzzy	MS evaluation and CP score
Ref [6]	ASE	2013	Total symmetry of the cluster, total compactness and connectedness	Give the multiobjective model for automatic clustering algorithm.	Precision, Recall and F-measure
Ref [7]	Research Gate	2013	Homogeneity and Separation	The multiobjective genetic clustering algorithm is used for categorical feature reduction	DB index, Dunn index, and C-index to measure the goodness of cluster
Ref [8]	IS	2014	Jm(Global cluster division) and XB(Minimum separation)	For categorical data incremental learning established multiobjective	MS measure, DB index and CP

				fuzzy clustering.	
Ref [9]	Springer	2014	Compactness, total symmetry of the cluster and counting the number of features.	Semi-supervised clustering in feature selection using multiobjective optimization	Minkowski score for seven datasets
Ref [10]	SC	2013	Improve problem of high dimensionality, problem of discontinuous pareto front and problem of non-convex pareto front	Using fuzzy version and K--medoids clustering a multiobjective differenitial evolution optimizer is improved	HV metric
Ref [11]	AI	2015	The goodness of partitioning, total symmetry and Sym-index	Proposed semi-genetic clustering multiobjective optimization	Sym-index, con index and I index
Ref [12]	IJSER	2015	Maximum similarity, compactness and improve the validity of cluster	Design self-adaptive differential evolution using multiobjective C-means data clustering algorithm	Silhouette index and symmetry-index and

Ref [13]	ESA	2016	Cohesion and connectivity these two objectives are optimized	Particle swarm optimization in multiobjective clustering	Accuracy and average for 40 independent run
Ref [14]	AIDM	2017	Compactness and connectivity	Using ITLBO algorithm to fuzzy clustering a multiobjective approach is used	Jm and XB index
Ref [15]	IJECS	2012	The minimum variation within clusters and maximum variation between cluster	Multiobjective genetic algorithm is used with K-means clustering optimization	Davies Bouldin validity index
Ref [23]	KLIS	2015	The optimal number of cluster and compactness	Automatic fuzzy clustering algorithm based on the synergy of two mutation using multiobjective	PBM index and Compactness validity index
Ref [24]	Research Gate	2015	The solution of Pareto optimal diversity and pareto optimal convergence	A gravitational searches algorithm based on archive multiobjective	Convergence metric, SP, GD and MS

				clustering algorithm	
Ref [25]	JMIHI	2015	Compactness, proper number of clusters and distance between data object and centroid	Fractional cuckoo search for multiobjective data clustering	Sym index, XB index and DB index.
Ref [26]	IS	2015	Deviation and connectivity	Multiobjective clustering used automatic retral system to supervise pareto front	Accuracy, DB index
Ref [27]	KIS	2012	Find different high-trait structures in a big collection of high-dimensional data for the purpose of navigation.	Frequent termsets multiobjective clustering	Completeness, cluster set depth and child count.

2.1.1 Proposed work

During the literature survey, it is observed that there are many problems in the clustering filed such as feature section, optimized algorithm for single objective as well multi objective clustering, etc. In this work, a VPS algorithm based multi objective clustering algorithm is proposed. This algorithm is based on the concept of free vibration and forced vibration. Free vibration of the freedom systems is single degree with viscous damping which activate this algorithm. VPS has recently newly developed by Kaveh and Ilchi Ghazaan. Two objectives that are optimized in this paper is compactness and connectedness. First objective is intra cluster variance, we have to compute the distance of object to the nearest cluster centre and we can also called that overall deviation of a partitioning. The second objective is connectedness of the cluster, neighbouring data objects have to identify whether they belongs to same cluster or not. Using these two objective we are trying to achieve more accurate result, better performance and efficiency. To achieve these performance parameters using two objective functions we have done literature survey on multiobjective clustering algorithm in which we have seen many objectives.

CHAPTER 3
PROBLEM IDENTIFICATION

This chapter explains the problems that are observed during a literature survey. Most of the problem that is appeared is on single objective clustering. Some single objective problems are:

- The single objective function is partisan for a special criterion.
- Single objective function is not efficient for better performance of data clustering.
- Single objective clustering does not form a proper number of clusters and not give better accuracy.

CHAPTER 4
PROPOSED SOLUTION

In our work, we will be proposing a multiobjective vibrating particle system clustering algorithm for data clustering problems. This approach is based on the conception of vibrations. VPS algorithm has two types forced vibration and free vibration [20,21]. Free vibration of freedom systems is single degree with viscous damping which activates this algorithm. Initially, the location of particles is randomly chosen in a k-dimensional search space.

$$X_j^i = x_{min} + \mathrm{rand}(\,) \times \left(\mathrm{x_{max}} - \mathrm{x_{min}}\right), \qquad (1)$$

In this, X_{ij} that is present in equation (1) is the ith variable of the particle j. X_{max} and X_{min} are the maximal and minimal vectors, [0,1] is the range of random number.

Now, with three positions three distinct weights are set and the location of every particle is modified by studying them during each generation. (H.B) is the historical best location of the long-term population, (G.P) is the good particle and (B.P) is the bad particle. Sort the latest population according to the objective values in an ascending order to choose the BP and GP for each candidate and then commonly choose the GP and BP from the first and second half. Modify rules in the VPS algorithm are given by

$$x_i^j = w_1.\left[D.A.rand1 + HB^j\right] + w_2.\left[D.A.rand2 + GP^j\right] + w_3.\left[D.A.rand3 + BP^j\right] \qquad (2)$$

$$K = \left[w_1.\left(HB^j - x_i^j\right)\right] + \left[w_2.\left(GP^j - x_i^j\right)\right] + \left[w_3.\left(BP^j - x_i^j\right)\right] \qquad (3)$$

$$w_1 + w_2 + w_3 = 1 \qquad (4)$$

Where, w1, w2, and w3 are the relative importance of GP, BP and HB. Evenly distributed random numbers within the range of [0,1] are rand1, rand2, and rand3. Bad particle is used for modifying the position formula to accomplish the quicker and efficient merging in the Vibrating particle system. Further, in this algorithm, our main aim will be to optimize two objective functions simultaneously which are compactness and connectedness.

Fitness function

The first objective is compactness (intra cluster variance) based on the Euclidean distance, the object must belong to the nearest cluster center.

$$\text{SSE (M)} = \sum_{i=1}^{p} \sum_{x_{k \in d_i}} \| x_k - s_i \|^2 \tag{5}$$

Here M is the solution, x_k indicates the kth object of the dataset and s_i is the centroid of ith object.

Second objective is connectedness, object in the neighbourhood must belong to same cluster. Using this, arbitrary shape of cluster is easy to identify.

$$\text{Connectedness (M)} = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} x_{i,ss_i(j)} \right) \tag{6}$$

Where M is the solution, m represents amount of elements resides inside of dataset, I indicates ith object in given dataset. ssi(j) indicates the jth nearest object of i xi,ssi(j) add the penalty with (1/j) if jth nearest neighbour is not present in same cluster. Otherwise, there will be no penalty if they both are in the same cluster. In more ways, for every data point i amount of penalties are identical to amount of penalties obtained in correspondence to value (1/j) because of every closest neighbour j when j and i do not reside in similar cluster; m shows the amount of closest neighbours to apply; it can be observed that penalty value found step by step diminishes for far neighbour. With the help of given both objective functions, similarity among clusters is evaluated using sum of square error (SSE) and dissimilarity within the clusters is evaluated completely by connectivity.

4.1 Proposed Methodology

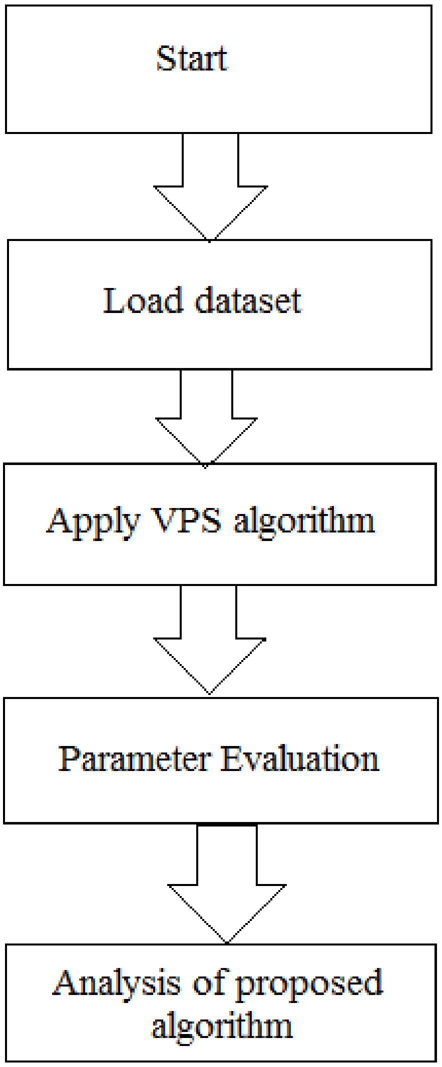


Figure 4.1 Proposed methodology

Figure 4.1. shows the methodology that is used in this work. Initially, it starts with loading the dataset such as iris, wine, glass, zoo and cancer. Then after loading the dataset applies VPS clustering algorithm on these datasets. The objective function is computed such as connectedness and compactness and check whether the proposed algorithm gives better clusters or not.

4.2 Flow chart of VPS algorithm

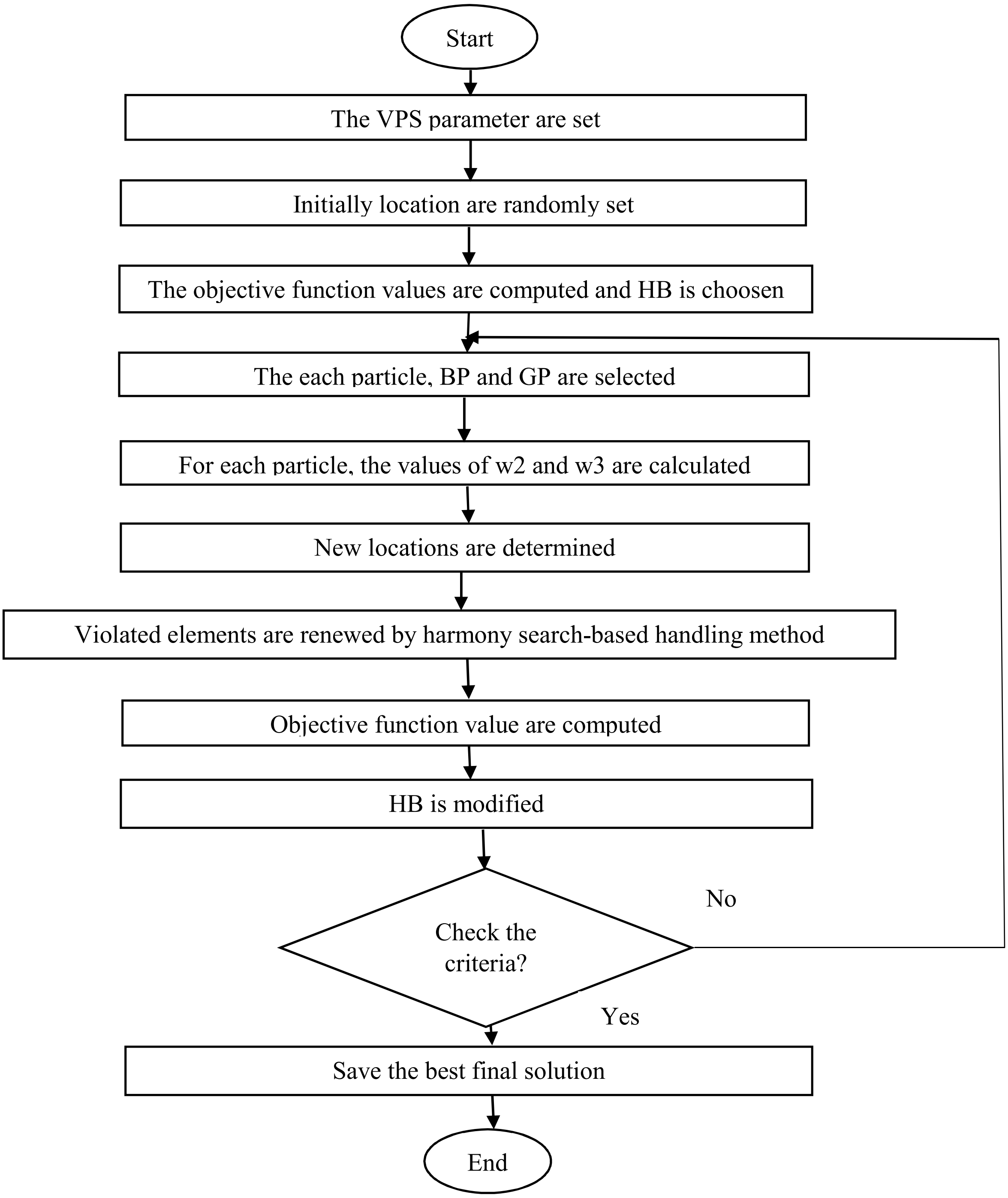

Start
The VPS parameter are set
Initially location are randomly set
The objective function values are computed and HB is choosen
The each particle, BP and GP are selected
For each particle, the values of w2 and w3 are calculated
New locations are determined
Violated elements are renewed by harmony search-based handling method
Objective function value are computed
HB is modified
Check the criteria?
No
Yes
Save the best final solution
End

Figure 4.2 Flow chart of proposed algorithm

CHAPTER 5
IMPLEMENTATION AND EXPERIMENT

5.1 Software requirement

Here experiments are performed in Matlab R2016a using 8GB RAM, Window OS and corei5 processor. Proposed algorithm optimize the two conflicting criteria connectedness and compactness (intra cluster distance) simultaneously. Using these two objectives efficiency and effectiveness of proposed algorithm is measured. To test the performance of this algorithm we use six real datasets from UCI machine learning repository.

5.2 Experimental results

5.2.1 Dataset used

Seven real datasets are used in this work such as Glass, Iris, Vowels, Wine, Zoo, and Wisconsin Breast Cancer from UCI machine learning repository. In Wisconsin Breast Cancer we use only 683 samples out of 699. Table 5.1 illustrates the dataset descriptions.

Table 5.1. Datasets descriptions:

Dataset	Number of Clusters	Number of rows	Number of columns
Iris	3	150	4
Wine	3	178	13
Zoo	7	101	16
Glass	6	214	9
Vowels	6	871	3
WBC	2	683	9

5.2.2 F-measure

F-measure[9] is used to check the accuracy of the achieved clusters. It considers both recall and precision as the harmonic mean. Fraction of recovered objects that are compatible is called precision and fraction of compatible objects that are recovered is called recall. F-measure is computed as

$$\mathrm{F-measure} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \qquad (1)$$

In this, precision=n_{ij}/n_i, n_{ij} in precision is the number of data objects that belongs to the pair of cluster p and q, n_i is the overall number of data objects present in cluster I, recall= n_{ij}/n_j, n_j in recall is the number of data objects present in class n_j. 1 is the optimum score of F-measure.

5.2.3 Implementation results and comparison

In this section, we compare the proposed algorithm with other four different algorithms MOPSO [26], MABC [27] and NSGA II [28] using six real datasets based on performance measure (F-measure). MOVPS algorithm perform well in all datasets compared to other algorithms. In Table 5.2 F-measure is calculated, algorithm with highest F-measure have better accuracy. In this table, standard deviation and mean of the best F-measure is evaluated. MOVPS has lowest standard deviation in Iris, WBC and Wine datasets which shows the robustness of the algorithm. In Wine dataset, accuracy is almost same for all algorithms and in Zoo and Vowel datasets standard deviation is better but in Glass standard deviation is high.

Table 5.2. Performance of proposed algorithm using distinct datasets based on F-measure.

Data Sets	MOVPS	TSMPSO	MOPSO	NSGA II	MABC
Iris	0.9467 (0.0132)	0.9265 (0.0182)	0.9089 (0.0142)	0.8902 (0.0353)	0.8602 (0.0556)
Glass	0.5591 (0.0169)	0.5584 (0.0144)	0.5443 (0.0255)	0.5546 (0.0162)	0.5016 (0.0495)
Vowel	0.6419 (0.0236)	0.6317 (0.0255)	0.6108 (0.0222)	0.6085 (0.0233)	0.6024 (0.0265)
WBC	0.9835 (0.0021)	0.9741 (0.0010)	0.9725 (0.0028)	0.9524 (0.0202)	0.9702 (0.0026)
Wine	0.7294 (0.0019)	0.7294 (0.0037)	0.7265 (0.0001)	0.7241 (0.0034)	0.7271 (0.0095)
Zoo	0.8317 (0.0156)	0.8131 (0.0162)	0.8017 (0.0183)	0.7852 (0.0220)	0.7905 (0.0245)

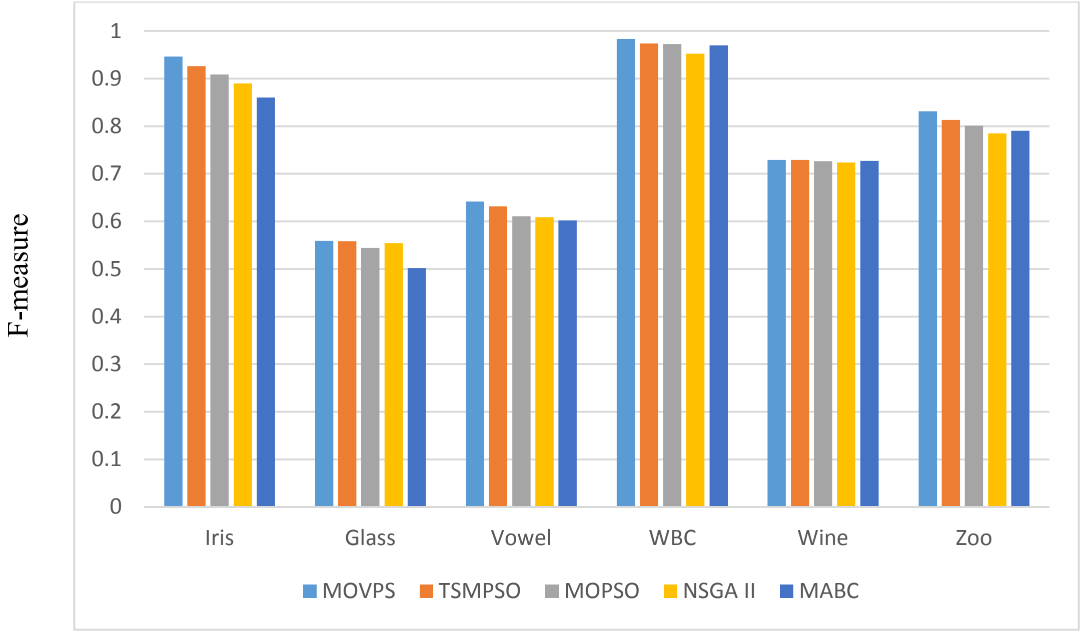
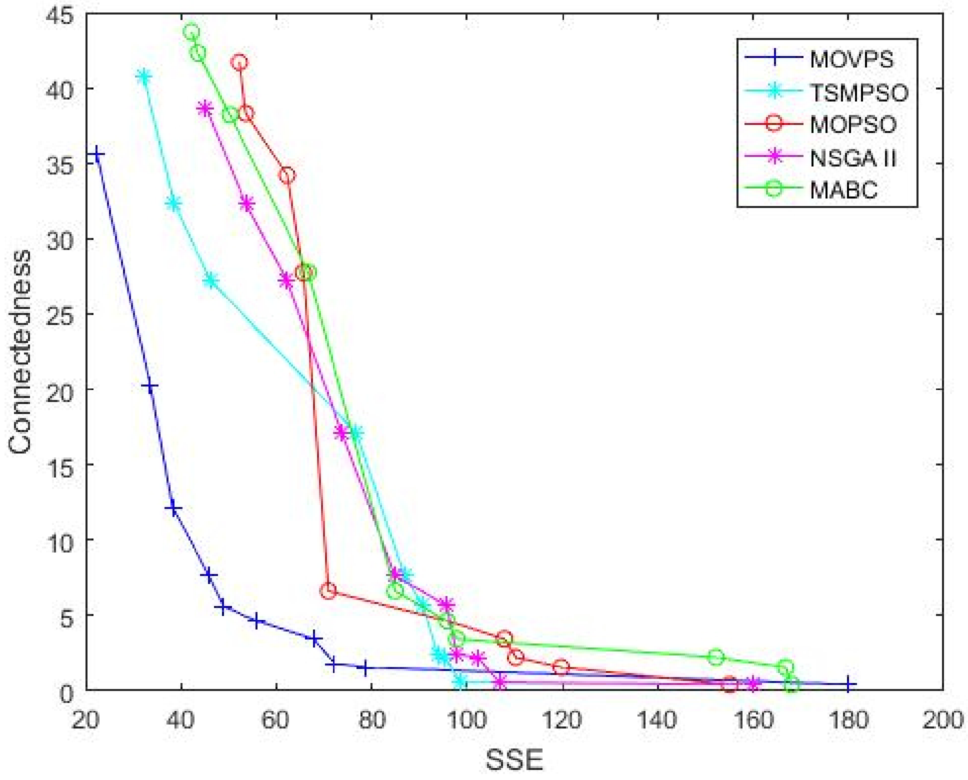


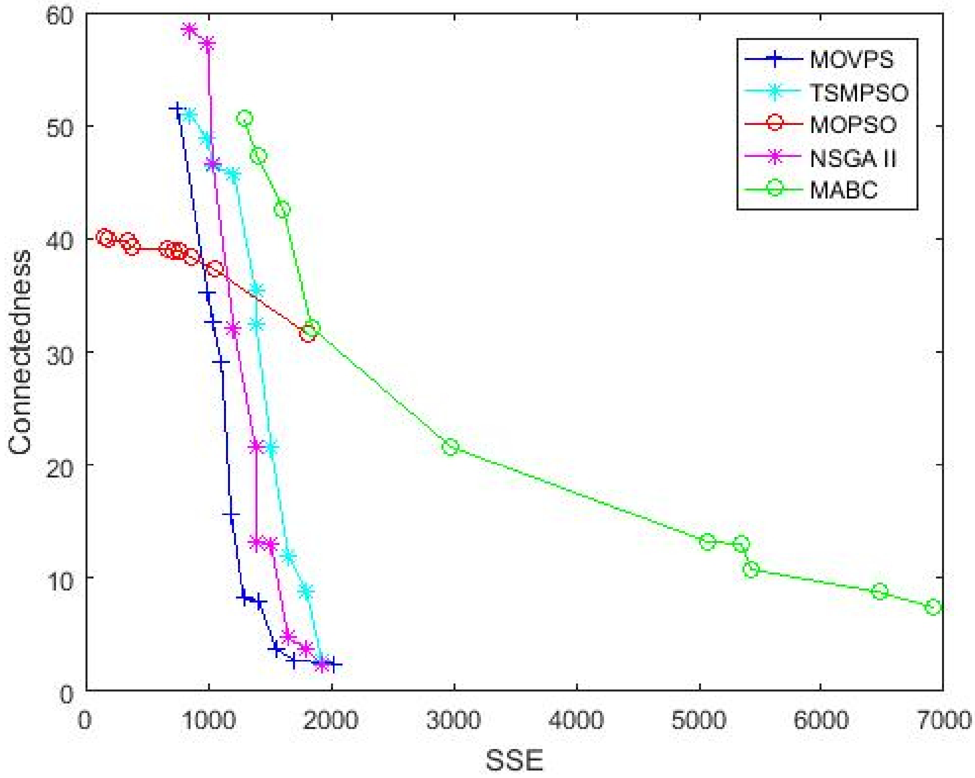
Figure 5.1. Bar Chart shows the Overall accuracy obtained in different runs.

Figure 5.1 depicts the overall accuracy of proposed algorithm obtained in different independent runs. MOVPS algorithm perform well in all the datasets as compare to other algorithms. MOVPS algorithm shows the perfection over other algorithm.

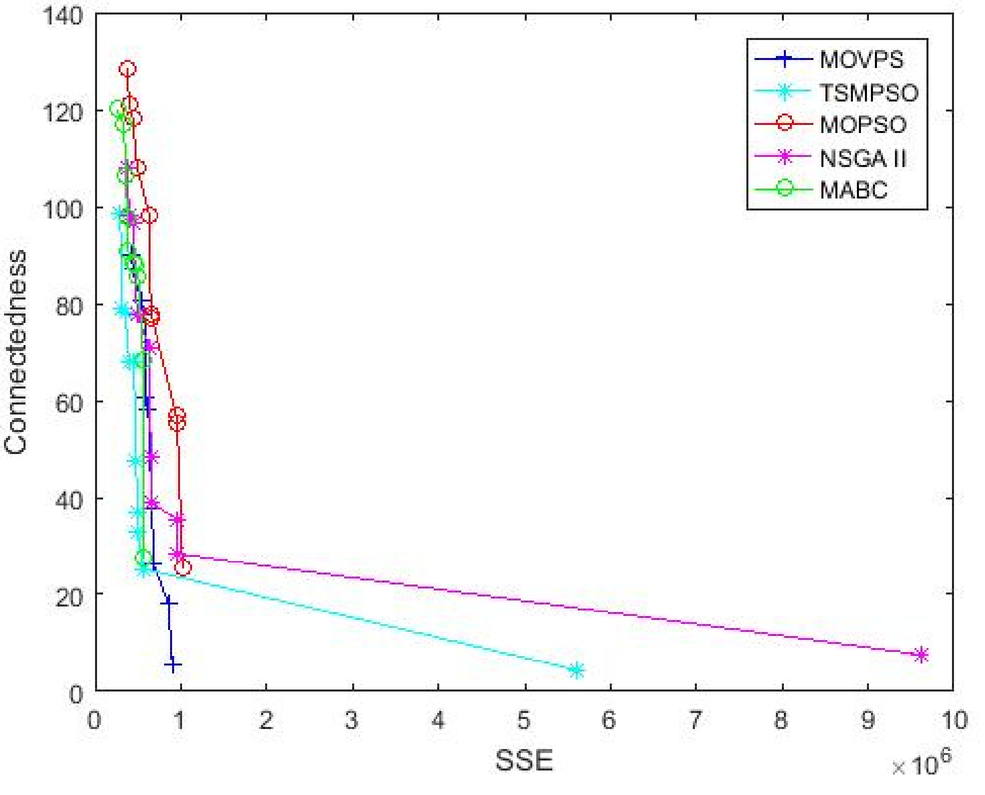
Figure 5.2 shows the Pareto fronts for all algorithm in different datasets, it certainly shows that proposed algorithm (MOVPS) performs better. Hence, we conclude that MOVPS is good to find optimal cluster centers.



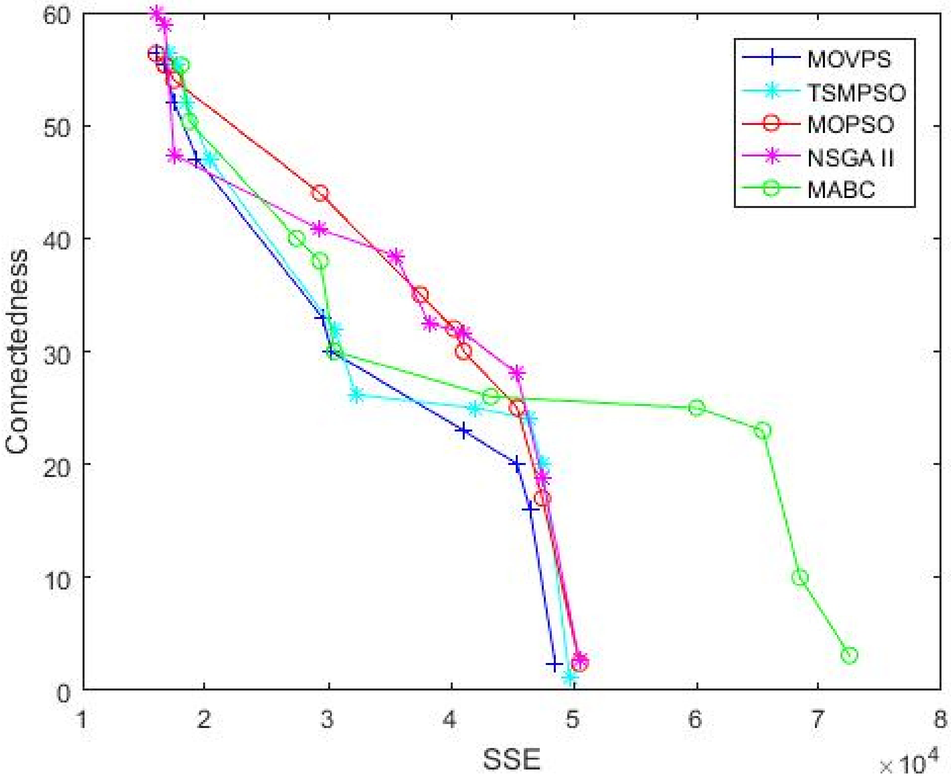
(a) Iris



(b) Glass

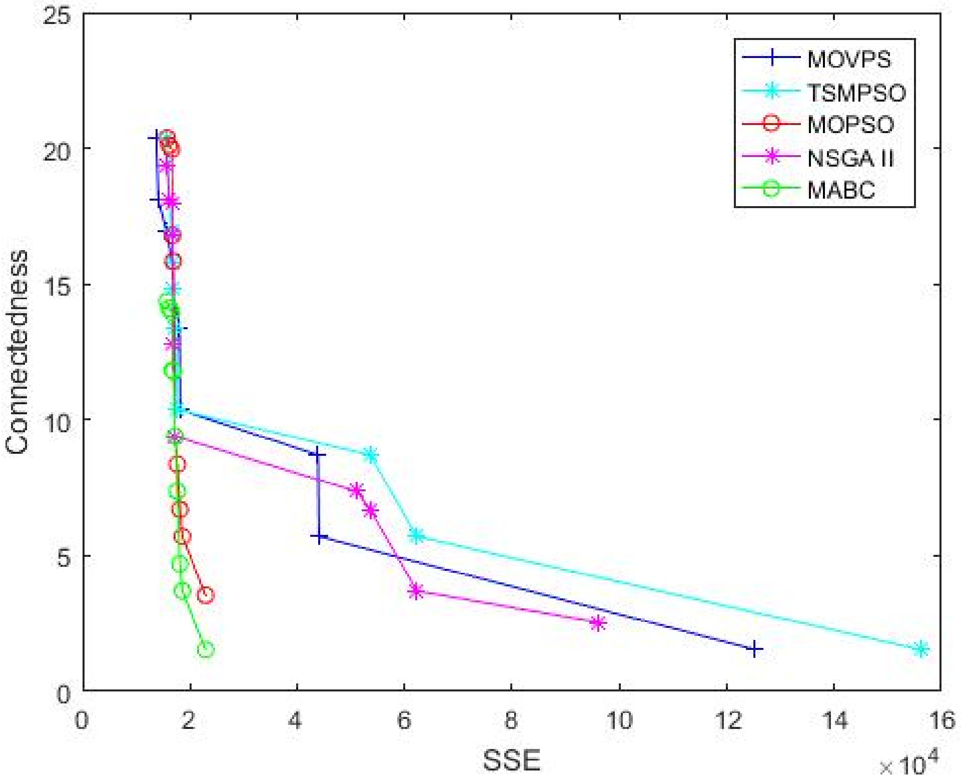


(c) Vowel



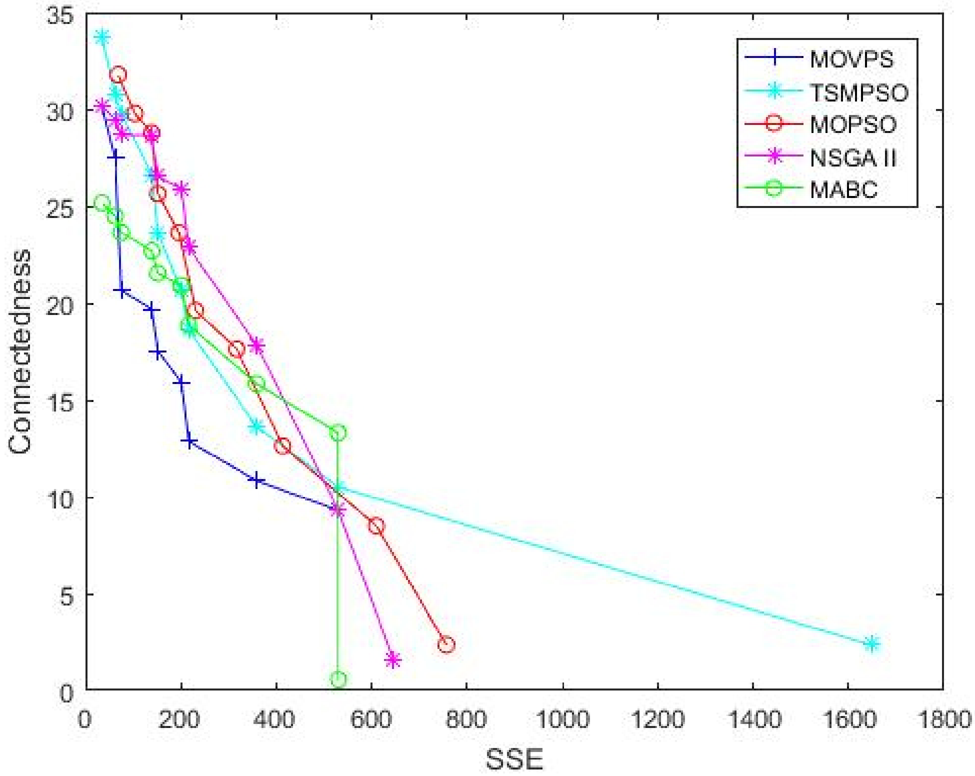
Connectedness
SSE
×10^4
0
10
20
30
40
50
60
1
2
3
4
5
6
7
8
MOVPS
TSMPSO
MOPSO
NSGA II
MABC

(d) WBC



Connectedness
SSE
×10^4
0
5
10
15
20
25
0
2
4
6
8
10
12
14
16
MOVPS
TSMPSO
MOPSO
NSGA II
MABC

(e) Wine



(f) Zoo

Figure 5.2 Results of the competing algorithms for all datasets. a) Iris; b) Glass; c) Vowel; d) WBC; e) Wine; f) Zoo.

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

The initial part of this report target on related work of various multiobjective clustering algorithm and comparison among them. Every proposed model has its advantages and disadvantages in the context of some particular concerns. Multiobjective can perform better than single objective clustering. A multiobjective optimization approach is natural way to optimize more than one objective function by cutting down the biases of a single objective. It optimize multiple objectives at the same time and helps when there is large amount of data. In this project, we are using multi objective vibrating particle system in which we will achieve two objective functions, connectedness and compactness. In future work, we can simultaneously optimize more than two or three objective functions.

References:

1. Prakash, J., & Singh, P. K. (2015). An effective multiobjective approach for hard partitional clustering. *Memetic Computing*, *7*(2), 93-104.
2. Bandyopadhyay, S. (2011). Multiobjective simulated annealing for fuzzy clustering with stability and validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *41*(5), 682-691.
3. Garza-Fabre, M., Handl, J., & Knowles, J. (2017). An improved and more scalable evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*.
4. Arkeman, Y., Wahanani, N. A., & Kustiyo, A. (2012). Saha, S., & Bandyopadhyay, S. (2010). A symmetry based multiobjective clustering technique for automatic evolution of clusters. *Pattern recognition*, *43*(3), 738-751.
5. Attea, B. A. A. (2010). A fuzzy multi-objective particle swarm optimization for effective data clustering. *Memetic Computing*, *2*(4), 305-312.
6. Saha, S., & Bandyopadhyay, S. (2013). A generalized automatic clustering algorithm in a multiobjective framework. *Applied Soft Computing*, *13*(1), 89-108.
7. Dutta, D., Dutta, P., & Sil, J. (2013). Categorical Feature Reduction Using Multi Objective Genetic Algorithm in Cluster Analysis. In *Transactions on Computational Science XXI* (pp. 164-189). Springer, Berlin, Heidelberg.
8. Saha, I., & Maulik, U. (2014). Incremental learning based multiobjective fuzzy clustering for categorical data. *Information Sciences*, *267*, 35-57.
9. Saha, S., Ekbal, A., Alok, A. K., & Spandana, R. (2014). Feature selection and semi-supervised clustering using multiobjective optimization. *SpringerPlus*, *3*(1), 465.
10. Kotinis, M. (2014). Improving a multi-objective differential evolution optimizer using fuzzy adaptation and $$ K $$ K-medoids clustering. *Soft computing*, *18*(4), 757-771.
11. Alok, A. K., Saha, S., & Ekbal, A. (2015). A new semi-supervised clustering technique using multi-objective optimization. *Applied Intelligence*, *43*(3), 633-661.
12. Soliman, O. S., & Saleh, D. A. (2015). Multi-objective C-means Data

Clustering Algorithm using Self-Adaptive Differential Evolution.

13. Armano, G., & Farmani, M. R. (2016). Multiobjective clustering analysis using particle swarm optimization. *Expert Systems with Applications*, *55*, 184-193.
14. Shahsamandi Esfahani, P., & Saghaei, A. (2017). A multi-objective approach to fuzzy clustering using ITLBO algorithm. *Journal of AI and Data Mining*, *5*(2), 307-317.
15. Arkeman, Y., Wahanani, N. A., & Kustiyo, A. (2012). Clustering K-Means Optimization with Multi-Objective Genetic Algorithm. *International Journal of Electrical & Computer Sciences IJECS-IJENS*, *12*(05), 61-66.
16. Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data mining and knowledge discovery*, *10*(2), 141-168.
17. Pedrycz, W., & Waletzky, J. (1997). Fuzzy clustering with partial supervision. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *27*(5), 787-795.
18. Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, *41*(8), 578-588.
19. Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, *2*(2), 169-194.
20. Kaveh, A., Hoseini Vaez, S. R., & Hosseini, P. (2017). Enhanced vibrating particles system algorithm for damage identification of truss structures. *Scientia Iranica*.
21. Kaveh, A., & Ghazaan, M. I. (2017). Vibrating particles system algorithm for truss optimization with multiple natural frequency constraints. *Acta Mechanica*, *228*(1), 307-322.
22. Filippone, M., Camastra, F., Masulli, F., & Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern recognition*, *41*(1), 176-190.
23. Liu, Ruochen, Lang Zhang, Bingjie Li, Yajuan Ma, and Licheng Jiao. "Synergy of two mutations based immune multi-objective automatic fuzzy clustering algorithm." *Knowledge and Information Systems* 45, no. 1 (2015): 133-157.
24. Abbasian, Mohammad Amir, Hossein Nezamabadi-pour, and Maryam Amoozegar. "A clustering based archive multi objective gravitational search algorithm." *Fundamenta Informaticae* 138, no. 4 (2015): 387-409.

25. George, Golda, and Latha Parthiban. "Multi Objective Fractional Cuckoo Search for Data Clustering and Its Application to Medical Field." *Journal of Medical Imaging and Health Informatics* 5, no. 3 (2015): 423-434.
26. Garcia-Piquer, Alvaro, Andreu Sancho-Asensio, Albert Fornells, Elisabet Golobardes, Guiomar Corral, and Francesc Teixidó-Navarro. "Toward high performance solution retrieval in multiobjective clustering." *Information Sciences* 320 (2015): 12-25.
27. Morik, Katharina, Andreas Kaspari, Michael Wurst, and Marcin Skirzynski. "Multi-objective frequent termset clustering." *Knowledge and information systems* 30, no. 3 (2012): 715-738.
28. Mann, Amandeep Kaur, and Navneet Kaur. "Survey paper on clustering techniques." *International Journal of Science, Engineering and Technology Research* 2, no. 4 (2013): pp-0803.
29. Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." *ACM computing surveys (CSUR)* 31, no. 3 (1999): 264-323.