A project report on

ENTROPY BASED AUTOMATIC CLUSTERING

Submitted in partial fulfilment of the Requirement for the award of degree of

Master of Technology

In

Information Systems

Submitted By

KAVITA

(2K15/ISY/11)

Under the Guidance of

Dr. Anil Singh Parihar

(Assistant Professor, Department of Information Technology, DTU)

2015-2017

Department of Information Technology

Delhi Technological University

Bawana Road, Delhi – 110042

CERTIFICATE

This is to certify that **Ms. Kavita (2K15/ISY/11)** has carried out the major project titled “**Entropy Based Automatic Clustering**” as a partial requirement for the award of Master of Technology degree in Information Systems by Delhi Technological University. The major project is a bonafide piece of work carried out and completed under my supervision and guidance during the academic session 2015-2017. The matter contained in this report has not been submitted elsewhere for the award of any other degree.

(Project Guide)

Dr. Anil Singh Parihar

Assistant Professor

Department of Information Technology

Delhi Technological University

Bawana Road, 110042

ACKNOWLEDGEMENT

I express my gratitude to my major project guide Dr. Anil Singh Parihar, Assistant Professor, Department of Information Technology, Delhi Technological University, for the valuable support and guidance he provided in making this major project. It is my pleasure to record my sincere thanks to my respected guide for his constructive criticism and insight without which the project would not have shaped as it has. I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.

I would also like to appreciate the support provided by our lab assistants, seniors and our peer group who aided us with all the knowledge they had regarding various topics.

Kavita

2K15/ISY/11

M.Tech (Information Systems)

ABSTRACT

Cluster analysis has been a fundamental research area in data analysis and pattern recognition. In this project, an Entropy based Automatic Clustering method is purposed. It automatically determines the number and initial position of cluster centers. In this, concept of Black hole entropy is used. It calculates the entropy for each data points in a dataset and then select minimum entropy as cluster center. Minimum entropy is chosen because it is the point which is more connected by other data points. Next, it eliminates all those data points which having a criteria greater than threshold value. Again, choose next cluster center in remaining dataset having minimum entropy. This process is repeated until no data points remain in the dataset. Now we will get the appropriate number of centers with their initial location. In order to avoid drawbacks which were occur in EFC, again we check the similarity measure with these obtained cluster centers and put the data point to that cluster for which its similarity value is higher. In this way, we will get better clusters and for making center at mean point, we will calculate the mean of all data points within a cluster and represent it by center of that cluster. There is one parameter which requires to handle i.e. threshold value which is easy to specify. In this method, there is no need to revised entropy value for each data point after cluster center is determined. So, it is simple in nature and also no need of user constraints. Experimental results shows how this method is good for predicting cluster centers. Results are also compared with other clustering algorithms like K-mean and FCM method. Complexity of this method is lesser than standard FCM method. It also handles large dataset very well.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER – 1

INTRODUCTION

1.1 Data Mining

We are living in a world full of data. Each day we come across with large amount of information which are supplementary processed and store in form of data [1]. Data mining is used in almost all area to analysis the data like in government and military systems, market research, medical area, and geographical information systems, etc. One of the major means of dealing with these data is to analyse the data objects which are in same classes or simply to group the objects. In general, Classification is a process related to labelling, in which facts or objects are determined, segregated and understood. In terms of Machine Learning, Classification is process of recognizing that an idea or object goes to which set of groupings, on the basis of already known observations. Classification is measured as Managed (Supervised) Learning where a known observation is existing properly. On the other side, unsupervised learning is known as clustering which clusters the objects based on some similarity measures. Data mining can be categories in following types:

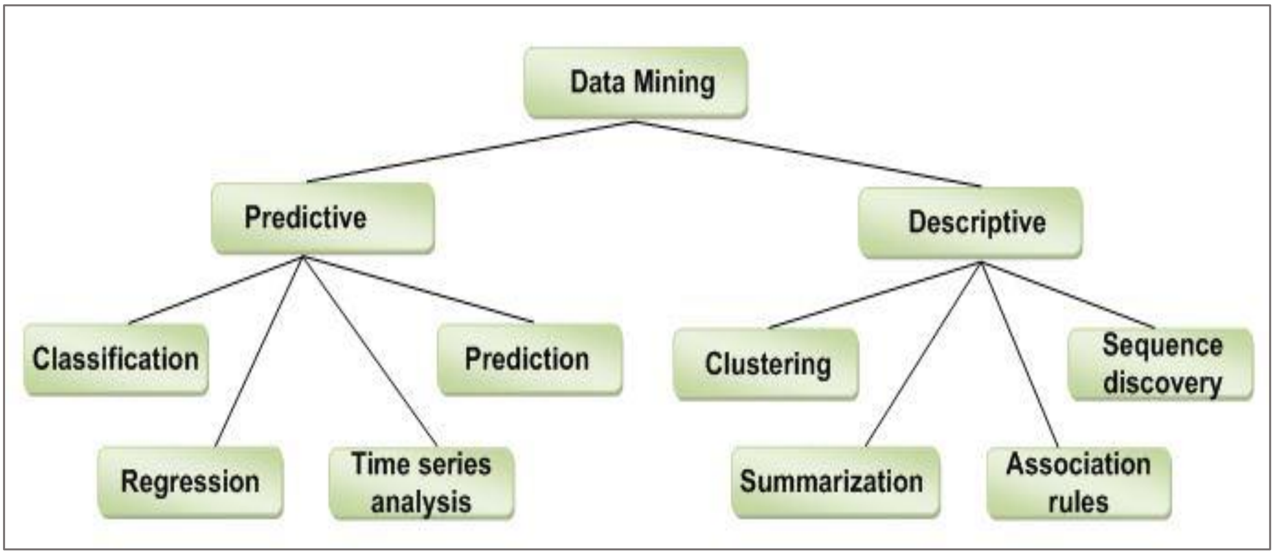


Figure 1.1 Categories of Data Mining

1.1.1 Classification

Classification is one of the measures of data mining to analyse the data. It is predictive type of data mining. In classification, we classify the patterns into their known labelled classes. Here we know the type of a class for a pattern and have to make such algorithms which can correctly classify the objects. It is also known as supervised learning. In this we have training set which will train the data based on specifies rules and other is testing set which determines the quality of a classification method. Training set is applied on the dataset until the error is reduced up to some threshold. Testing set is not familiar to algorithm but it having same characteristics as of trained set. In this way classification methods work for categories the patterns.

1.1.2 Clustering

Clustering is the process of assigning the same type of objects in same group and other types of objects in another group. Main objective of clustering is to increase intra cluster likeness and to decrease inter cluster likeness. It is also known as un-supervised learning as pre-observation regarding clusters are not given. Main difference between clustering and classification is that, in classification we classify the objects or pattern to already known classes whereas in clustering natural groups are formed for unlabelled data objects. Clustering is a descriptive method of data mining. The grouping issue can be expressed as takes after: given an arrangement of n information focuses (x_1 ; :::; x_k) in d dimensional space Rd and a whole number k, segment the arrangement of information into k disjoint bunches to limit some misfortune work. Clustering process contains following general steps:

- Extract features
- Cluster into categories
- Consolidation

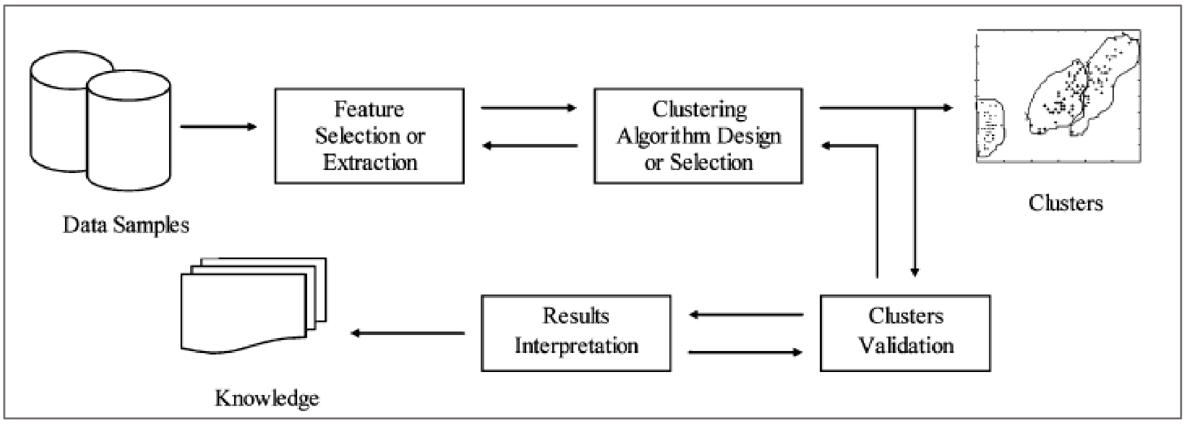

Data Samples
Feature Selection or Extraction
Clustering Algorithm Design or Selection
Clusters
Clusters Validation
Results Interpretation
Knowledge

Figure 1.2 Process of Data Clustering

1.1.3 Similarity of data

Similarity is the measurement for clustering. It is a quantity that replicates the connection or bond between two data items, it embodies the level of likeness of two patterns. It signifies the degree of belonginess among objects across all the appearances used in analysis. This similarity measure can be used in both view of separation and gathering capacities, for example, Euclidean separation, Manhattan remove, Murkowski separate, Cosine similitude, and so on to gathering objects in clusters [2] . The clusters are shaped such that any two information questions inside a group have a base separation esteem and any two information protests across over various groups have a greatest separation esteem. Clustering utilizing separation capacities, called remove based grouping, is an extremely well-known procedure to bunch the items and has given great outcomes.

1.1 Motivation

Cluster analysis is one of the most widespread techniques used in data mining. The idea behind cluster analysis is to find expected groups with data in such a way that each element in the group is homogenous to each other as possible. At the same time, the groups are heterogenous to other groups as likely.

Main motivation behind this report is to efficiently calculate the required number of clusters in a given dataset using some entropy. As in Fuzzy C Mean Clustering and K-Mean Clustering we explicitly have to give number of cluster to continue the process. Our assumption of giving clusters may be wrong or right up to some perspective. The whole clustering procedure is based on the number of centres we want to create. So, this process should be carry out by focusing on dataset and how many clusters should be sufficient in a given dataset.

1.2 Objective and Scope

Two utmost significant benefits of clustering are as follows:

1. can detect and remove noisy or outliers' data objects
2. ability to deal with large datasets and can detect initial location of cluster centre automatically

The objectives of the thesis is

- To realize an approach to cluster the data efficiently and to speed up the entire process
- To improve the strategy of initial selection of centroids to automatic centres
- To completely eradicate the problem of local-minima

- To determine which data belongs to which cluster by the use of similarity matrix
- To check the performance of the proposed algorithm on different input datasets and comparing their results.
- To develop a comparative study of the proposed algorithm and the existing algorithms.

1.3 Organization of Thesis

The thesis in all consists of seven chapters and references

Chapter 1 is the introduction. It evokes the motivation of the work, aim of thesis and structure of thesis

Chapter 2 description of work done by various people and their contribution. It also explains what clustering is and different techniques of data clustering.

Chapter 3 Introduction and explanation of related algorithms. It also reports the benefits and respective limitations of the existing data clustering algorithms

Chapter 4 This chapter discusses the purposed Algorithm which was taken as an inspiration for the research design and architecture of the proposed black hole clustering algorithm

Chapter 5 Experimental results

Chapter 6 Conclusion and Future work

In the end, bibliography is being included.

CHAPTER – 2

LITERATURE REVIEW

In this chapter, a review on most used clustering methods will be purposed. Here we describe the different categories of clustering with their respective algorithms and comparison.

Clustering methods are un-supervised because we don't know the nature of clusters and the chosen dataset. Training and testing part is not provided here, like classification. In this case clustering parameters are computed from learning data. Unsupervised learning is a sort of machine learning procedures used to draw derivations from datasets comprising of information without marked reactions. There are two sorts of learning: Off-line or Batch learning for information accessible in pieces and On-line learning for information that arrive consecutively.

Off-line learning gives the best results but not applicable for real time application and very large data like social data or internet-data. Clustering methods are more appropriate for real, heterogeneous and large data sets with many attributes [3].

Cluster examination becomes a complex problem because of following issues:

1. Active comparability or similarity measures
2. Criteria function
3. Algorithms

These are become an integral factor in formulating a very much tuned clustering strategy for a given grouping issues. In addition, it is well known that no clustering technique can satisfactorily deal with a wide range of group structures i.e. shape, size and density. There is no universal clustering algorithm which can apply and gives satisfactory outcomes to any dataset. All techniques are application based and having their own pros and cons.

2.1 Categories of Clustering:

Followings are the main categories of clustering:

1. Partitional Clustering
2. Density based Clustering
3. Hierarchical clustering
4. Grid Based Clustering
5. Graph Oriented Approach
6. Soft Computing Based Approaches

2.1.1 Partition Clustering

The mostly used clustering method is Partition clustering for its simplicity and robustness nature. It separates the dataset in different cluster which have much dissimilarity among each other and high similarity within it. For n number of data points, it clusters dataset in K number of clusters where (k<=n). This separation is done on the bases of some objective function which set the criteria of clustering and it is different for each partitioning technique.

This method should fulfil the following necessities:

- At least one element should belong to a cluster
- A element can only refers to one cluster at a time.

Most applications accept widespread exploratory methods like in the k-means algorithm, in which mean value of the objects represents the cluster.

Typical Methods: K-means, K-medoids, PAM (Partition Around Medoids) and CLARA are some of the partition clustering algorithms.

a) K-mean Algorithm: k-means algorithm is in use since 1965 [4] and is the most used clustering algorithm because its easiness of implementation and its efficiency. Euclidean Distance is used to clusters the dataset and the objective function J for k-mean algorithm is as follows:

$$J(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (|| x_i - k_j ||)^2 \qquad \text{..................(2.1.1)}$$

Where k: cluster centre

c: number of clusters

$n_{i:}$ number of elements in cluster i.

Partitioned clustering approach in K-mean:

a) A centre is associated for every cluster

b) Each object is allotted to its nearest cluster centre

c) Number of clusters, K, must be specified [5].

b) Variants of K-mean Algorithm [6]: Several improvements have been purposed in basic K-mean algorithm due to some limitation in original algorithms. These upgrading are done on the basis of using other distance as compared to Euclidean distance as it only identifies spherical clusters or some variance in picking centre point. Some of advancement in initial algorithm are as given below:

➢ K-Medoids or general K-Mean algorithm which choose medoid as data centre and uses Manhattan Norm to define distance between data points. A medoid can be explain as the object of cluster whose mediocre distinction to all the objects in the cluster is minimal. It is the most central located point or object in the cluster.

- K-mode and K-median are another popular variant of K-mean where median is used in place of mean of calculating centre of clusters in K-median algorithm. And in case of K-modes, simple matching similarity measure matrix is in place to Euclidean distance.
- Kernel K-mean Algorithm: This is another variant of k-mean which uses kernel method as distance measure as compared to Euclidean method. It gives better results as compared to original one but having large time complexity and also complex in nature. It appropriate for real data.
- Hierarchical K-mean Algorithm: In this method, k-mean algorithm is applied on dataset for some limited iteration after that the resulted centred points are executing using hierarchical technique. Then resulted centroids are used as initial centres in k-mean method.

2.1.2 Density Based Clustering:

The conventional partitioning methods can form only spherical-shaped cluster as they are distance based, these techniques cannot group random shaped clusters in efficient way. Density based clustering methods are established based on the belief of density. Their basic concept of density based clustering is to form a cluster until the density of neighbourhood crosses the threshold point. For understanding, each data point within a given cluster must have minimum number of points in their neighbourhood. Doing this, we can exempt from the noisy and outliers data point in a dataset and can also formed uninformed shaped clusters. Classically, these methods consider restricted clusters only, and do not deliberate fuzzy clusters.

Typical Methods: DBSCAN and OPTICS are density based methods that produce clusters based on the density.

a) DBSCAN Method (Density Based Spatial Clustering of Application with Noise): The core concept of this method is, there must be exist a minimum number of objects or points for every point with a giver radius of neighbourhood in a given cluster. R*-Tree is implement for this process [7]. Basically, it uses the idea of Density reachability and Density connectivity.

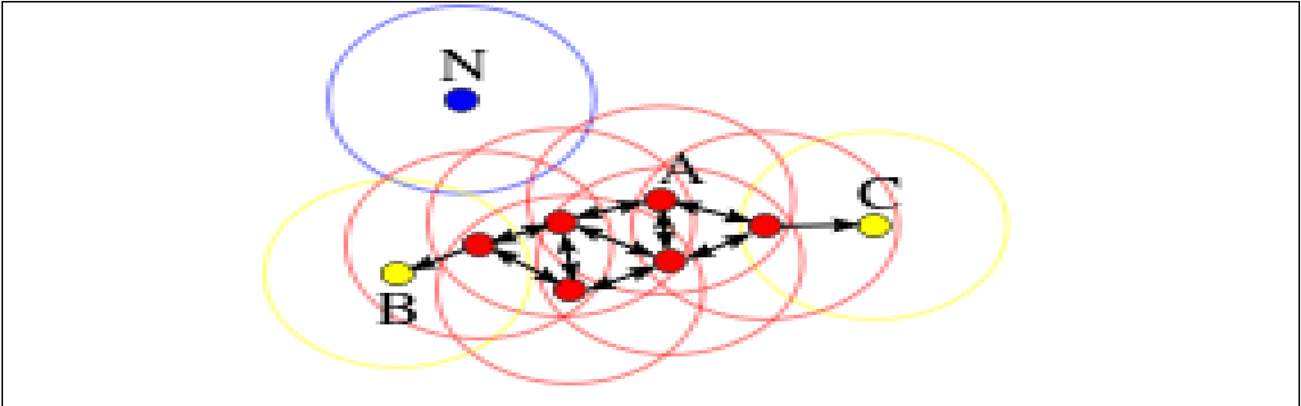


Figure 2.1 DBSCAN Method, B and C are connected by several points with A as core point and N is a Noise point

DBSCAN divides data points into three groups which are

1. Core points
2. Border points and
3. Noise points

Core points are the interior objects and Border point are exterior point where as Noise points are outliers i.e. neither core points nor border points. DBSCAN method is well suited for identify noise and outliers and also able to find arbitrary clusters. But it is not perfect for high dimensional datasets.

b) DENCLUE Method (DENsity based CLUstEing): This method uses solid mathematical foundation or density distribution function. It is faster than DBSCAN and also good for large noisy datasets. This method is based on influence of impact factor for each object which can be modified using mathematical functions also called as influence function which determine the impact factor with respect to other neighbour data points. The overall density factor can also be modelled analytically. Clusters are identified using these density measures. These all needs a large number of mathematical parameters which cause complexion.

c) OPTICS Method (Ordering Points To Identify the Clustering Structure) [8]: This method is similar to DBSCAN with improvement in calculating criteria for centres. It provides an efficient cluster with ordering the data objects with reachability value and connectivity value. It can be represented graphically or using visualizing techniques.

d) DBCLASD (A Distribution Based Clustering Algorithms for Mining for Large Spatial Databases): This is an incremental method in which a point is allocated to the cluster that managed inclemently without acknowledged the cluster [9]. It also discovers the clusters of uninformed shape. The efficiency of this algorithm is very high for large scale spatial databases.

2.1.3 Hierarchical Clustering

These kind of technique makes a various levelled disintegration of the given dataset objects. Based on the decomposition it can be classified into agglomerative or division approach. In the agglomerative approach, bottom up strategy is followed i.e. starting from a data object and forming a separate group. It then untermittedly merges the objects of the group close to each other until all of the groups merges into one, till the termination condition is met or the one in the topmost hierarchy is reached. On the other hand, divisive approach is top-down approach, the iteration begins with all the data points in one cluster and in consecutive iteration a cluster is fragmented into smaller clusters. This process goes on until each element is a separate cluster or the termination condition is met. These levelled techniques encounter an issue from the way that once a consolidation or a split is done it can never be fixed i.e. they can't correct wrong choices. Typical Methods: Diana, Agnes and BIRCH are some Hierarchical Clustering techniques.

a) BIRCH Method (Balanced Iterative Reducing and Clustering using Hierarchies): It produces best outcomes for given resources like memory and time constraints. Maximum of the cases, BIRCH only needs single examination of database. It is a main memory based algorithm. Main concept of this clustering is clustering features and CF tree (cluster feature tree). A CF tree is a height-balanced tree which contain the information about clusters. Limitation of previously clustering methods was that they can’t handle large dataset i.e. which are not fitted in main memory at a time. While BIRCH makes the full efforts to use given resources to produce best outcomes of clusters in a given dataset. It is also an incremental process which does not needs whole dataset for processing. And it is very suitable for discrete and continuous attributed data clustering problems. Generally, BRICH having following two phases [10]:

   1) Scan the database in initial memory to create a CF tree.
   2) Use a random clustering method to cluster the leaf or last nodes of CF tree.

   Main negative point about this algorithm is that it is more suitable for only numeric database and also depends on the structure of the database.

b) CURE (Clustering Using REpresentatives): It use sample point variant as the cluster representative rather than every point in the cluster. Traditional clustering techniques are better for spherical shaped cluster whereas this method is more robust to outliers and non-spherical clusters and other variants [11]. In this method, firstly we have to set a target sample number ***c***. Than try to select **c** well scattered sample points from the cluster. The chosen scattered points are contracted toward the centroid in a portion of α where $0 \leq \alpha \leq 1$. It follows the property of both centroid based and all data point algorithm. This method cannot be directly applied to large database as its time complexity is high i.e. $O(n^2 \text{ logn})$. Architecture of CURE can be shown as:

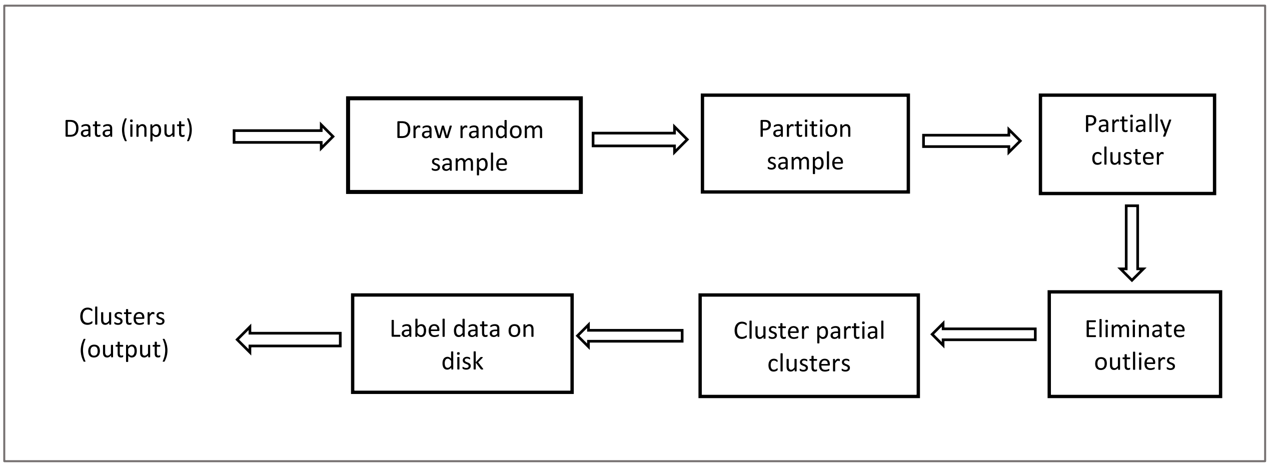


Figure 2.2 Architecture of CURE Technique

Different enhancement is done on this method. Some are, CURE clustering with map reduce technique, Random sampling and partition method to handle large database, CORE method using Hadoop and etc.

c) ROCK (Robust Clustering algorithm for Categorical attributes): It group dataset with quality attributes. It uses “link” as measure to describe relation between data point and its neighbour points. Link(p,q) means number of common neighrbour data points. If this counting is large that means both point belongs to same cluster. In this way measurement is calculated. Here also, a random sampling is used to handle large database.

2.1.4 Grid Based Clustering:

Main concerned of this method is with value space of data points and not with the data point itself as we read in previous all clustering. A typical grid based clustering consists following steps:

1) Dividing the data space into fixed number of cells or grid
2) Calculating density for each cell i.e. cell density
3) Sort all cell densities
4) Identify the centre of cluster
5) Reallocation of neighbour cell.

This type of clustering is depending on the grid cell not on the data points as earlier said. Due to this reason, these types of clustering having fast processing speed.

Typical Methods: STING, WaveCluster, OptiGrid and CLIQUE.

a) STING Method (a STatistical INformation Grid approach): The method is used for answering the diferent kinds of spatial queries. A grid with several levels depends on the different levels of resolution is being formed on the spatial area. Statistical information of each cell is pre-computed and stored in hierarchical structure. Statistical parameters may be like mean, variance, maximum or minimum and type of distribution i.e. numerical feature of a cell [12]. It uses top-down method to answer the spatial queries. Time complexity of STING method is O(n) where n is total number of grid cells. Main negative point of this method is that it can only identify horizontal and vertical boundaries can't detect diagonal boundaries.

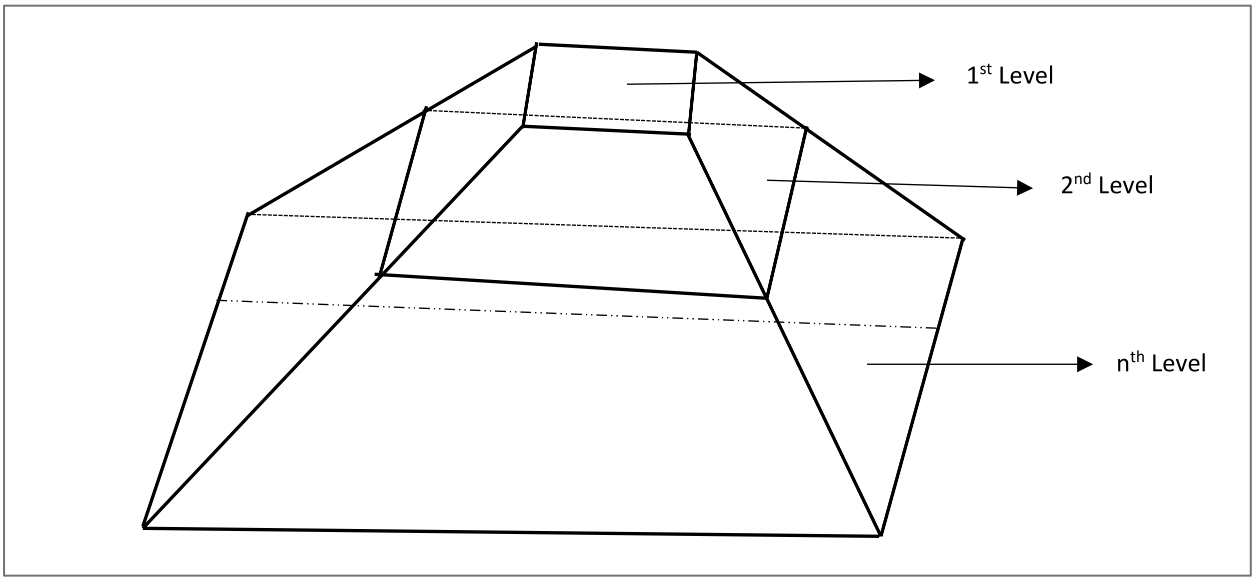


Figure 2.3 Grid structure of STING Method

b) CLIQUE (CLustering in QUEst): This method can be measured both as Grid based as well as Density based. It partitions the data space into grids as property of grid based method and a cluster is chosen as maximum set of connected dense cell in the subspace. It uses multi-resolution grid data structure. This method is suitable for high dimensional database. Major steps are:

1) Identify subspaces that contains cluster using apriori function
2) Identify clusters using dense units
3) Determine minimal cover of each cluster

Performance of this method is directly depends on the size of input and also maintain the correctness after increment the dimensions. Strength of the method is that it robotically finds classes of the highest dimensionality with high density clusters exist in those subclasses.

c) WaveCluster Method: It is used to find cluster in very large scale dataset. Firstly, it divides the dataspace into multidimensional space grid and then in order to find dense space it transform the origin feature by using wavelet transformation. Apply wavelet transformation multiple times to get fine cluster to coarse cluster. A wavelet transformation is a signal processing technique that convert a signal into different frequency sub-band. This method has time complexity of O(n) and also sensible to outliers. But it is not appropriate for high dimensional database. WaveCluster is well competent of discovering discretionary silhouette groups with composite structures. In this we have to give expected number of clusters and wavelet transformation as well number of application of wavelet transform as input parameters. Knowing exact number of clusters is not necessarily [13].

2.1.5 Graph Oriented Approach:

As graph theory provides more detailed information about the structure of database in terms of cliques, cluster, centre and outliers. Mostly popular clustering techniques are based on graph-theory. These algorithms are based on the MST i.e. minimum spanning tree of a dataset. MST can be defined as the subgraph which contains all the nodes with minimum weight connected to each other without cycle. The hierarchical approach is also somewhat based on graph theory with hierarchical structure. Single-link clusters are the sub-graph of MST of the data. Minimum spanning tree can be constructed either using Kruskal's or Prim's algorithm. Graphs are faster to access the information regarding the data as compared to other methods. Graph is a mathematical tool to represent the network and relationship between data. Here data or objects represents the node or vertex and distance represents the edge weight in the graph. Basic steps in graph-theoretic approach is

1) Construct the MST of the data
2) Delete the maximum length edges to generate clusters

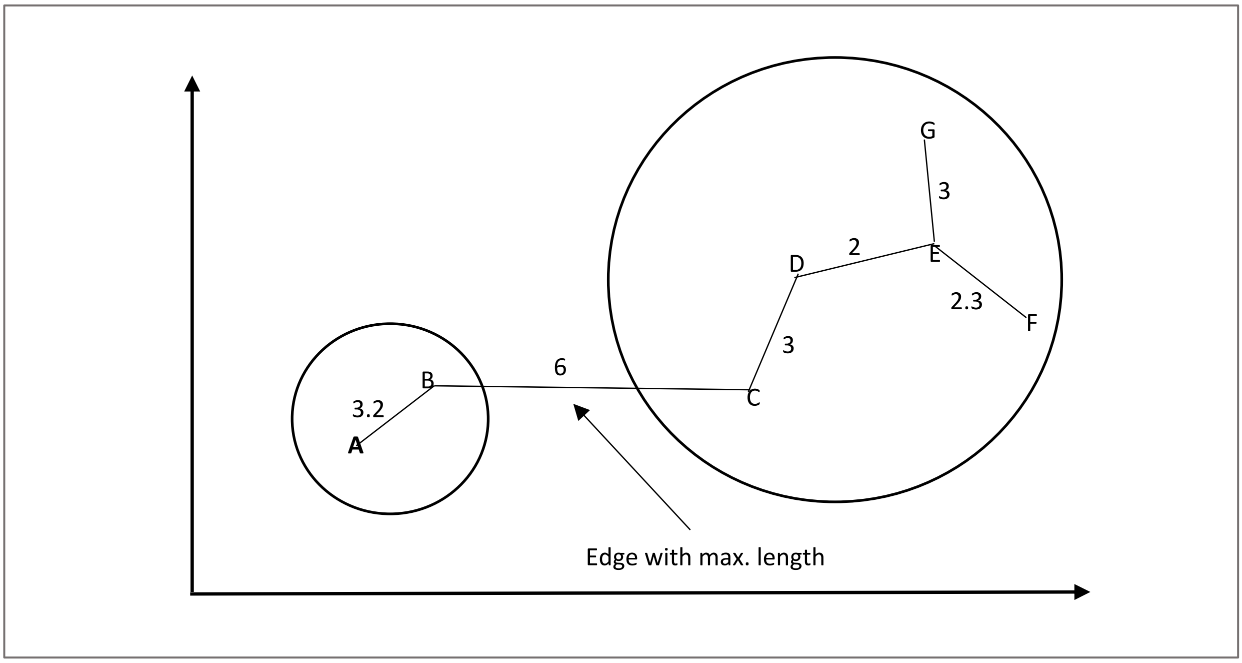


Figure 2.4 Example of Graph oriented method

This technique is highly efficient and scalable. Pre-knowledge of total number of clusters is not required. It is well applicable for social media database which can be explicit or implicit groups. In the simplest case, clusters are the connected components of the graph. Different graph clustering algorithms are: K-spanning tree, k-Spanning Tree, Shared Nearest Neighbor, Betweenness Centrality Based, Highly Connected Components, Maximal Clique Enumeration, Kernel k-means etc.

2.1.6 Soft Computing Based Clustering Approaches:

All of the above discussed methods are hard clustering techniques. As the name implies soft computing is different from hard computing. Soft computing paradigm includes computing using fuzzy logics, neural networks, probabilistic reasoning and genetic algorithms. Soft computing solutions lie between truth and false prediction. It can be partial truth where as hard computing is always truth or false. There is a difference between possibility and soft computing. Possibility is used where we don't know much about solution of a problem but in case of soft computing, problem itself has not so clearly defined. These types of problem are generated from human mind, which are with doubts and emotions. These techniques are based on logical systems i.e. sentential logic and predicate logic, on the other hand hard computing is based on binary logics, crisp system. Although soft computing gives approximate results whereas hard computing gives precise answer. Soft computing algorithms are efficient in speed and cost. They produce faster results and also require less resources.

a) Fuzzy Clustering Method [14]: As it is a soft computing approach so it follows the idea of partial set membership. In all hard clustering method, a data point only belongs to one cluster at a time but in fuzzy system a data point may belongs to several clusters at the same point with some degree of membership as shown in figure 2.5.

Membership function in fuzzy system can ne define as the degree of belonginess of a data point to other respective sets. It lies between 0-1 value. The fuzzy itself state the fact that anything cannot always be in true or false state, it may be partial truth. One of the best fuzzy clustering method is FCM i.e. fuzzy c-mean where c are the number of cluster centres.

General steps are:

- Choose the number of clusters
- Initialize membership matrix
- Compute centres as per giver criteria
- Update the membership matrix also
- Repeat the steps 3 & 4 until criteria matches

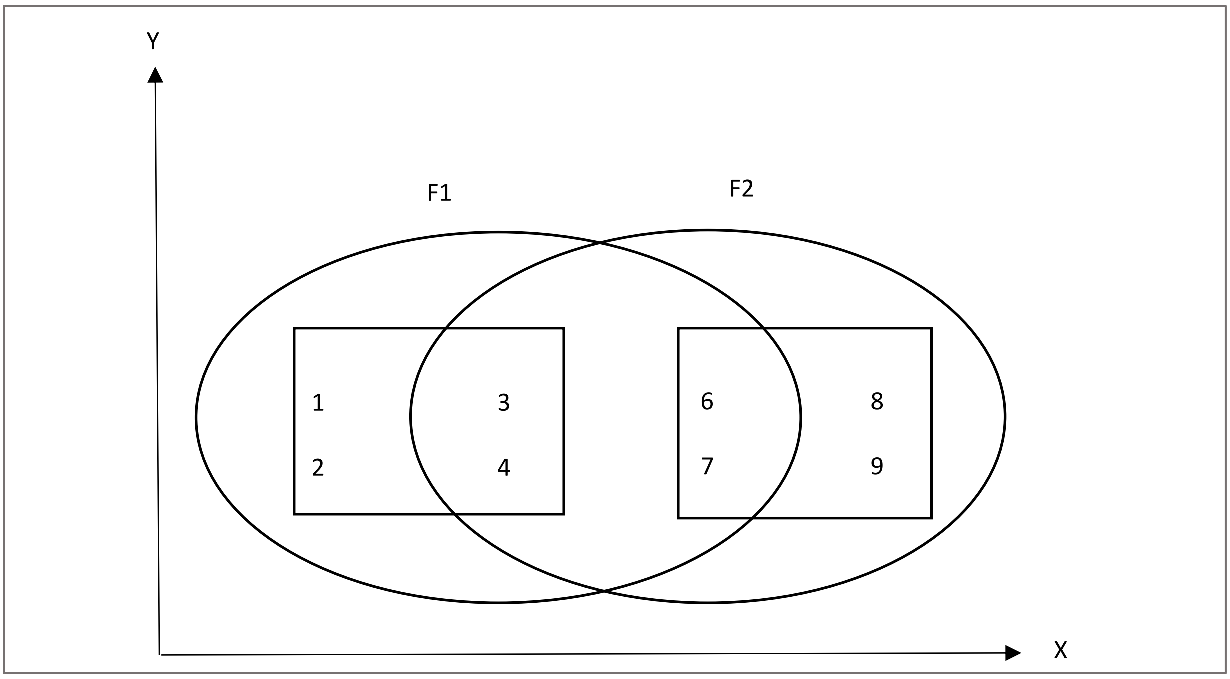


Figure 2.5 Partially distributed classes in Fuzzy Clustering

Above are the basic steps of fuzzy based algorithm. Fuzzy system algorithms are widely used in image processing for segmentation, as pattern recognition technique in bioinformatics, in marketing etc.

b) Neural Network based Clustering [15]: Neural network or artificial neural network (ANN) is a computational model which follows the structure of biological neural connection and widely used in machine learning paradigm. As soft computing methods are used for solving human problems, in the same way neural network are used to solve the problem as a human would. Neural networks are basically used in supervised learning technique where neurons are given under training and testing phase. In unsupervised learning, ANN is used in respect to get better representation of input data rather than giving clusters. These are used to find the pattern between input and output data. SOM (self-organized map), Adaptive resonance theory(ART), neural gas are some neural network clustering techniques.

c) Evolutionary approach for clustering: In order to select appropriate cluster centres in data space we use genetic algorithms for clustering. Traditional calculus based methods are not well suitable for clustering discontinuous and multi-model datasets. GA (genetic algorithms) are used to provide robust search in such spaces [16]. These algorithms are based on natural genetics. These algorithms have basic three operations which are: 1) Selection 2) Cross-over and 3) Mutation. There is also a fitness function according to which these operations are performed. It takes the solution as input and generates the suitability of the solution as output. Sometimes it is also called as object function.

It calculates approximate k value and resolves the best assemblage of the documents into k clusters.

Flow chart for general steps in Evolutionary approach:

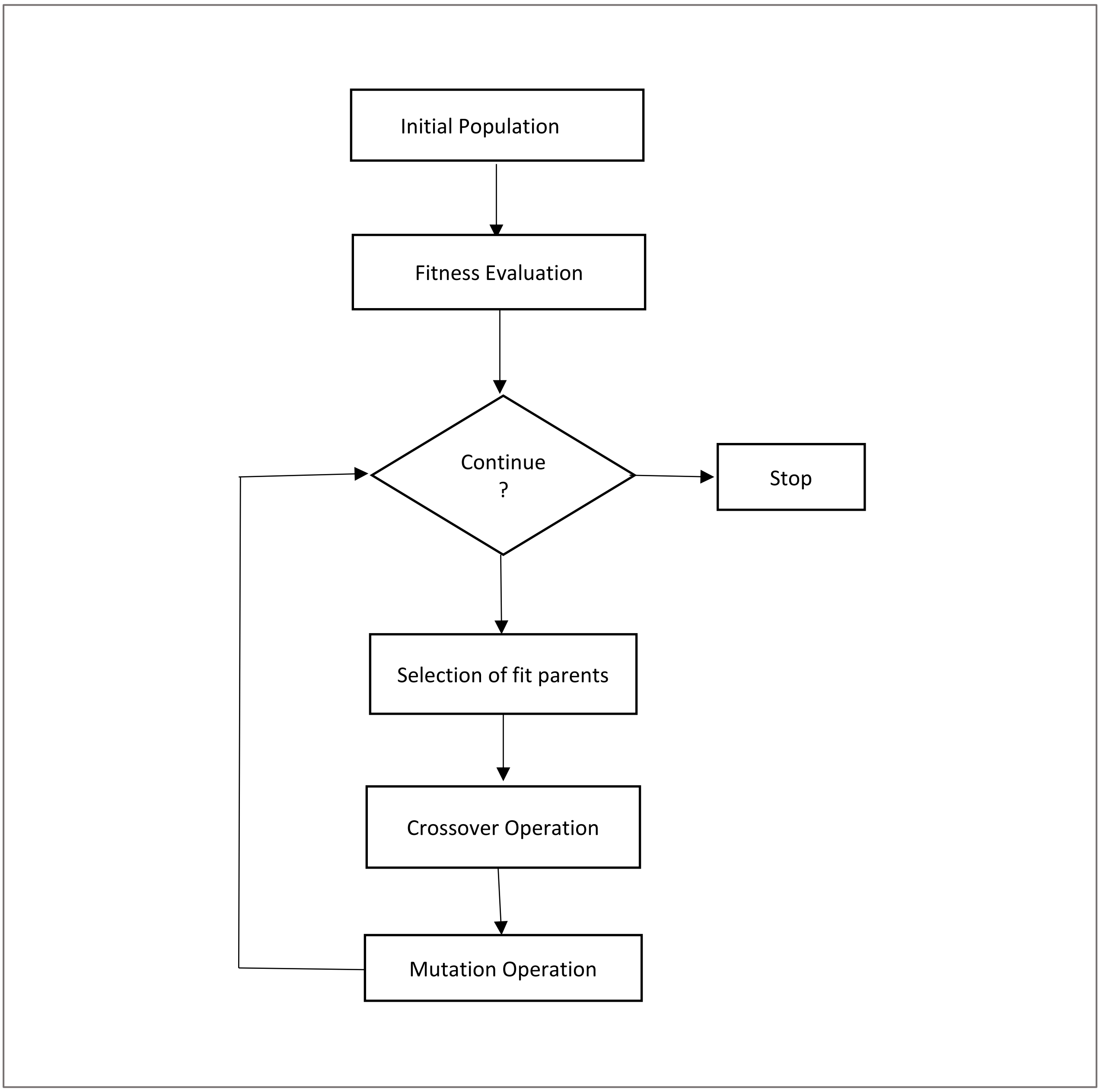


Figure 2.6 Steps in Evolutionary Approach

CHAPTER – 3

RELATED ALGORITHMS

3.1 K-Mean Clustering Technique

Clustering is the process of dividing a finite set of data points into a finite number of clusters with finite objects. K-mean clustering method is type of partition method. It is very easy to implement and the complexity of this algorithm is O(mn) where m are the total number of iteration and n are data points in a dataset. Objective function J(k) with k numbers of clusters for K-means algorithm is to minimize the following equation:

$$J(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\| x_i - k_j \|)^2 \qquad \text{..................(3.1.1)}$$

Where k: cluster centre

c: number of clusters

$n_{i:}$ number of elements in cluster i.

Algorithmic steps:

Let X be the set of n data points and K be the set of c cluster centres

1) Initially choose c cluster centres randomly
2) Determine the distance between all data points to these cluster centres
3) Now, allot the data points to cluster with minimum distance to centre
4) Update the centres with following way

$$k_i = (1/c_i) \sum_{j=1}^{c_i} x_j \qquad \text{..................(3.1.2)}$$

5) Again, calculate the distance of each data points to new centers and assign to minimum separated center cluster
6) Repeat the whole process until no changes are discovered.

Advantages of K-Mean Method:

1) Easy to understand and fast algorithm

2) Having time complexity O(nmc) where n are total data points, m is dimensional and c are number of clusters

3) Gives best results for well separated data clusters or spherical shaped clusters.

Disadvantage of K-Mean Method:

1) Doesn’t identify random shaped clusters efficiently

2) Requires pre-specification of number of centers

3) Randomly selection of center points can’t produce better results

4) Noise and outliers are not handle properly

5) Euclidean distance measure is not a best similarity measure

3.2 FCM Clustering Technique

Fuzzy C-mean clustering is a soft computing technique in which a data point can belongs to all clusters with specific membership value. It is based on the minimization of objective function given as [17]:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \, || \, x_i - c_j \, ||^2 \qquad for \; 1 \le m < \infty \quad ....(3.2.1)$$

Where μ_{ij} : membership matrix of data point x_i in jth cluster

N: total number of data elements in dataset

C: number of clusters

m: real number with value >1

Algorithmic Steps:

1) Initializing phase: initialize the membership matrix with random values between 0 – 1, select c number of clusters and also take the value of m (usually taken as 2)

2) Updating phase:

a) Calculate the centers of clusters as :

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m . x_i}{\sum_{i=1}^{N} u_{ij}^m} \qquad ....(3.2.2)$$

b) Update membership matrix as:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{|| \, x_i - c_j \, ||}{|| \, x_j - c_j \, ||} \right)^{\frac{2}{m-1}}} \qquad ....(3.2.3)$$

3) Stopping Criteria: Perform the updating part till

$$\max_{ij}\left\{|\, u_{ij}^{(k+1)} - u_{ij}^{(k)}\right\} < \varepsilon \qquad ....(3.2.4)$$

4) Now this algorithm gives the efficient centers in the dataset, after we have to allot data point to that particular center cluster whose membership degree is highest.

Advantages of FCM:

1) Gives better outcomes as compare to K-mean clustering technique
2) Unlike all other hard clustering methods, in this a data element can belongs to more than one cluster with assigned membership values
3) It also diminishes the intra-cluster differences.

Disadvantages of FCM:

1) Like k-mean clustering, it also requires pre-knowledge of initial cluster centers
2) Due to above reason, it also having local minima problem
3) Gives higher membership values to noise and outliers which leads to inefficient clustering process
4) Not suitable for high dimensional datasets.

3.3 PFCM Clustering Technique

Possibilistic Fuzzy C-Mean clustering is the advancement of FCM which is mixture of FCM and PCM (Possibilistic C-Mean) clustering. In PCM algorithm, typicality matrix is there which will efficiently deals with outliers. Objective function for PCM is as follows:

$$p_m(T,V;X,\gamma) = \sum_{i=1}^{n}\sum_{k=1}^{c} t_{ik}^{m} d_{ki}^{2} + \sum_{i=1}^{c}\lambda_i \sum_{k=1}^{n}(1-t_{ik})^{m} \qquad ....(3.3.1)$$

Where T is typicality matrix, V is centre matrix, X is dataset and ϒ is user defined constant.
PFCM includes all characteristics of both methods which will help to overcome the limitations of FCM technique.
The objective function for PFCM is given by the following equation

$$J_{m,n}(U,T,V:Z) = \sum_{i=1}^{c}\sum_{k=1}^{n} a\mu_{ik}^{m} + bt_{ik}^{m} * \|\, z_k - v_i \,\|^2 + \sum_{i=1}^{c}\delta_i \sum_{k=1}^{n}(1-t_{ik})^{\eta} \; ......(3.3.2)$$

where $\sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ik} = 1 \qquad \forall k, 0 \le \mu_{ki}, t_{ki} \le 1, a > 0, b > 0, m > 1$

and a, b are user-defined parameters which specify the relation between membership matrix and typicality matrix. Algorithmic steps for PFCM are same as FCM with following change in updating phase:

$$\mu_{ik} = \left(\sum_{j=1}^{c} \left(\frac{D_{ikA}}{D_{jkA}} \right)^{\frac{2}{(m-1)}} \right)^{-1} \qquad \text{for} \quad 1 \le i \le c \, ; 1 \le k \le \eta \qquad \text{..............(3.3.3)}$$

$$t_{ik} = \frac{1}{1 + (b(D_{ik}^{2}) / \delta_{i})^{\frac{1}{\eta-1}}} \qquad \text{for} \quad 1 \le i \le c \, ; 1 \le k \le \eta \qquad \text{............(3.3.4)}$$

$$v_{i} = \frac{\sum_{k=1}^{n} (a\mu_{ik}^{m} + bt_{ki}^{\eta}) x_{k}}{\sum_{k=1}^{n} (a\mu_{ik}^{m} + bt_{ki}^{\eta})}, k > 0 \qquad \text{for} \quad 1 \le i \le c \qquad \text{............(3.3.5)}$$

Advantages of PFCM:

1) Overcomes the noise and outlier problem of FCM
2) Also solve the coincident cluster problem in PCM.

3.4 Entropy Based Fuzzy Clustering Technique

In context to the information theory, entropy is the mean or average of the information contain in the message. In this technique entropy value for a data point is calculated and according to its value all calculations are performed. All above discussed algorithms having a week point of selection number of clusters manually. We have to give initial number of clusters to function then it continues all its calculation. But what if we initially give wrong prediction of clusters. This will tend to wrong implementation of the process. In EFC method entropy of a data with respect to other data is define as [18]:

$$E = -S \log_{2} S - (1-S) \log_{2}(1-S) \qquad \text{..................(3.4.1)}$$

Where S is the similarity measure and E is entropy of a data object.

The total entropy E_i for a data point i to all respective data points is define as:

$$E = -\sum_{j \in x}^{j \ne i} (S_{ij} \log_{2} S_{ij} + (1-S_{ij}) \log_{2}(1-S_{ij})) \qquad \text{.............(3.4.2)}$$

In this method, distance based equation is used as similarity measure whose value lies under the interval of 0 - 1. Values near to zero means data point are more closer and one where data points having distance near to mean distance of all the pair of data points. The similarity measure S_{ij} between points i and j is given as:

$$S_{ij} = e^{-\alpha D_{ij}} \qquad \text{................(3.4.3)}$$

Where D_{ij} is the distance between points i and j and α is a constant which can be calculated by assigning the S_{ij} value as similarity of 0.5 (mean value) when two data points have separation of mean distance of all data points from each other i.e. D. So, in this way α can be considered as:

$$\alpha = -\ln(0.5/D) \qquad \text{...............(3.4.4)}$$

Algorithmic steps:

In this algorithm, there are n data points with m dimensions

1) Calculate the distance metric with size n*n from one data point to another with Euclidean distance formula.
2) Determine the similarity measure S_{ij} according to equation no. 3.4.4
3) Determine total entropy for each data point and select minimum entropy i.e. E_{min} as the first cluster centre.
4) Select all the points with having similarity measure as greater than threshold value T (in this we take T as 0.7) and assign all these points to respective cluster.
5) Repeat steps 3 and 4 until all data points are covered.

- Drawback of EFC method:

  The main limitation of this algorithm is that as it selects all points within threshold value to that particular cluster centre, but it may possible that selected data points may have larger threshold value for another cluster centre. This makes this algorithm to wrongly classify the objects in a cluster.

CHAPTER – 4

PURPOSED ALGORITHMS

4.1 Black Hole Theory

Black hole is actually such a heavenly body in which its gravitational force is so strong that even light cannot pass through it. Anybody that crosses the boundary of black hole is gulped by the black hole and the body vanishes with the speed of light. Black hole is having massive gravitational pull but cover small area. When the fuel of a star comes up short then it is no possible to holds the weight. The weight from the stars layers of hydrogen push down driving the star to get shrink. Eventually star will become smaller than an atom [19]. As we can't see a black hole, their presence can be feel by strong gravitational force.

Basic structure of Holes has three layers:

1) Event horizon
2) Singularity
3) Ergosphere

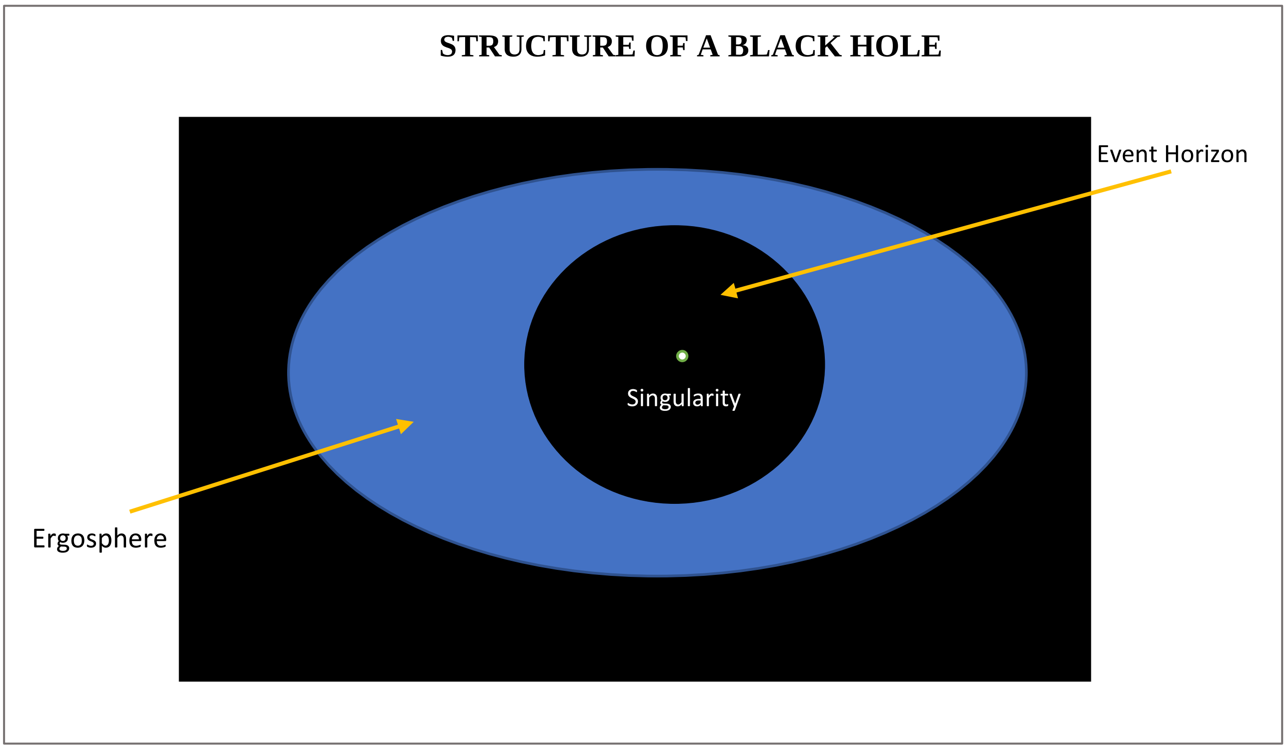


Figure 4.1 Structure of Black Hole

- The inner boundary of black hole region as shown in fig. 4.1 from where nothing can pass is Event horizon. Once an element goes inside the event horizon, it cannot leave. Gravity is constant all over the event horizon.
- The white dot in inner region of a black hole as shown in above fig, is known as its singularity. It is the point where mass of black hole lies.
- Ergosphere can also be called as outer event horizon, it is the area from one can extract energy and mass from this region

4.2 Black Hole Entropy for Clustering

BHE plays an important role in astrophysics, due to its statistical significance. One can also use this nature of entropy for clustering purpose. The basic equation for BHE Entropy is as follows [20]:

$$H = A/4 + \gamma \ln(A/4) + \beta \qquad \text{..........................(4.2.1)}$$

Where H is the entropy, A defines the area of horizon, γ and β are the constants through which we can mould the results of entropy based on our convenience.

4.3 Purposed Method

In the previous chapter, description of many methods is given in which most of are not automatic in cluster centre calculation. They need pre-knowledge of number of clusters formed in a dataset which is not an appropriate method of clustering. As wrong initialization may result inefficient clustering. Although Entropy based Clustering is an automatic clustering but it also has its limitation which we will discuss further.

So, main motive of this method is to calculate efficient number of cluster without pre-requirement of cluster centres. In this method, we use Black Hole Entropy (BHE) to calculate the entropy of each element in the dataset with respect to the all other elements. The equation for total entropy for element X_i with respect to all other n number of data-points is as:

$$H(X_i) = \sum_{j=1}^{n} Siml_i(j) + \gamma \ln(Siml_i(j)) + \beta \qquad \text{..........................(4.3.1)}$$

Here $Siml_i(j)$ is the similarity measure of ith with respect to jth element. Similarity measure can be of any criteria of measuring likeness between two data elements. Its value lies between 0 - 1 and it can be of different type based on the measuring attributes as discussed in chapter 3. γ and β are the user defined constants.

In our method, Euclidean distance is taken basic of similarity measure. Firstly we calculate the distance between each point to every other points in the dataset. Then we normalize the whole distance matrix as described in pseudo code. Then we fuzzifier that normalize matrix by taking its negative exponential. Which will call as similarity matrix having values between 0-1, where 1 represents much closer points and 0 defines fartest points.

4.3.1 Algorithmic Steps for Entropy based Automatic Clustering Method

Let's a dataset X of n number of data elements

1. Initialize all user define constants
2. Assign the threshold value ε between 0 - 1 (mostly taken as 0.7)
3. Define a distance_matrix of n x n in which distance of each element with respect to all other element is calculated using Euclidean distance formula
4. Normalize the distance_matrix using

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \qquad \text{........................(4.3.2)}$$

5. Calculate Siml_matrix of n x n using eq.
6. Calculate entropy of each element using eq. no. 5.1.1
7. Select min entropy as the cluster centre and assign all dataset which having similarity value greater than ε
8. Repeat step 7 until no data element left

   Phase 2:
9. Using above approach, we will get the efficient number of cluster centres. Now again calculate distance of each element with respect to all obtained cluster centres.
10. Assign the data-elements to that cluster which having min distance value.
11. Make mean of data points in a cluster as centre.

4.3.2 Flow Chart for Entropy based Automatic Clustering

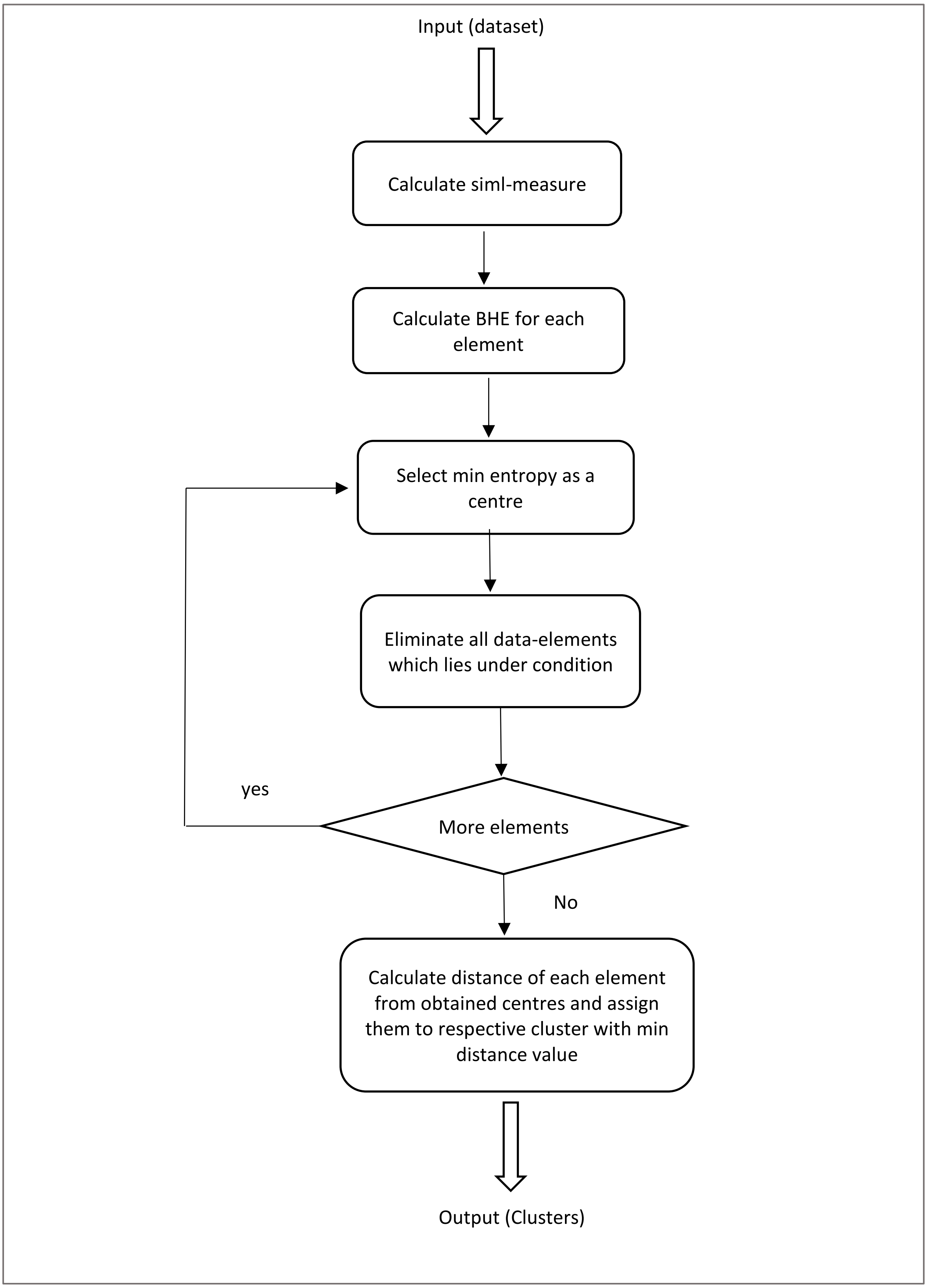

Input (dataset)
Calculate siml-measure
Calculate BHE for each element
Select min entropy as a centre
Eliminate all data-elements which lies under condition
yes
More elements
No
Calculate distance of each element from obtained centres and assign them to respective cluster with min distance value
Output (Clusters)

Figure 4.2 Flow Chart of purposed algorithm

CHAPTER – 5

Experimental Results

5.1 Datasets

The proposed algorithm is tested on five standard datasets which are taken from UCI machine learning repository.

Table 5.1 Characteristics of test datasets

S.No.	*Datasets*	*No. of clusters*	*No. of features*	*No. of data objects*
1	Iris	3	4	150 (50,50,50)
2.	Wine	3	13	178 (50,71,48)
3.	Glass	6	9	214 (70,76,17,13,9,29)
4.	CCPP	2	5	9568*
5.	Magic	2	10	19020 (12332,6688)

1.Iris Dataset: It is flower dataset with four attributes as petal length and width and sepal length and width. The dataset covers three clusters with 50 instances each.

2. Glass Dataset: The glass dataset by USA Forensic Science Service contains 6 forms of glass in relations of oxide content with 178 number of instances. There are in all ten attributes but we are using nine attributes for clustering as we are using numerical data only for clustering.

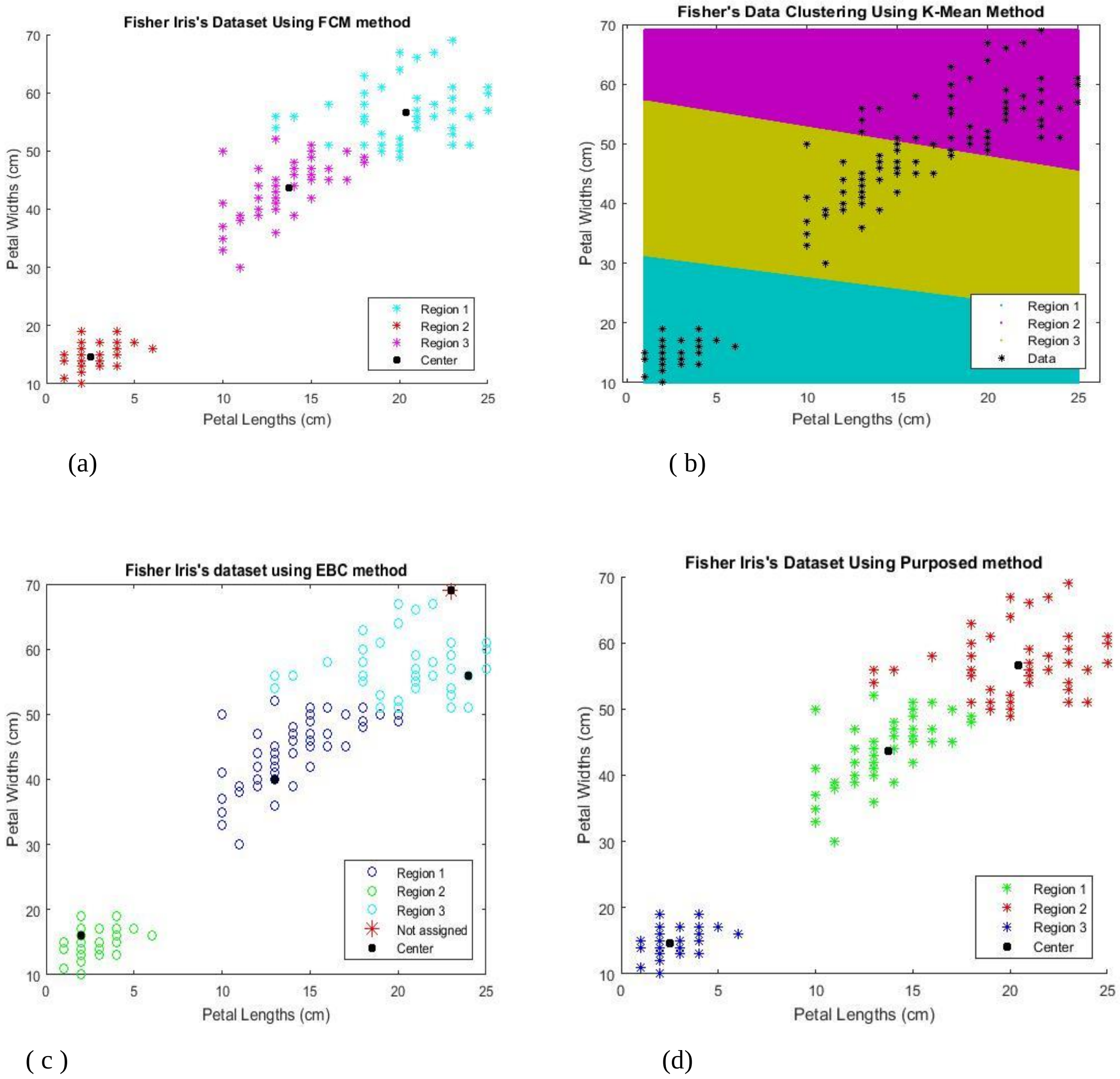
3. Wine Dataset: It is a dataset containing the data of analysis of chemical determining the origin of wines. The analysis gives the features of 13 elements found in three types of wines.

4. CCPP (Combined Cycle Power Plant) Dataset: The dataset is gathered from a Combined Cycle Power Plant (CCPP) which is collected more than 6 years (2006-2011) to work with full load. It contains five characteristics which formed two unique groups.

5. Magic Dataset: It is a Magic gamma telescope dataset generated to simulate process of high energy gamma particles in an atmospheric telescope. Again, we are using only 10 attributes from 11, as we only need numerical data for clustering.

5.2 Pictorial View of Clustering

1. Fisher Iris DataSet:



(a)

(b)

(c)

(d)

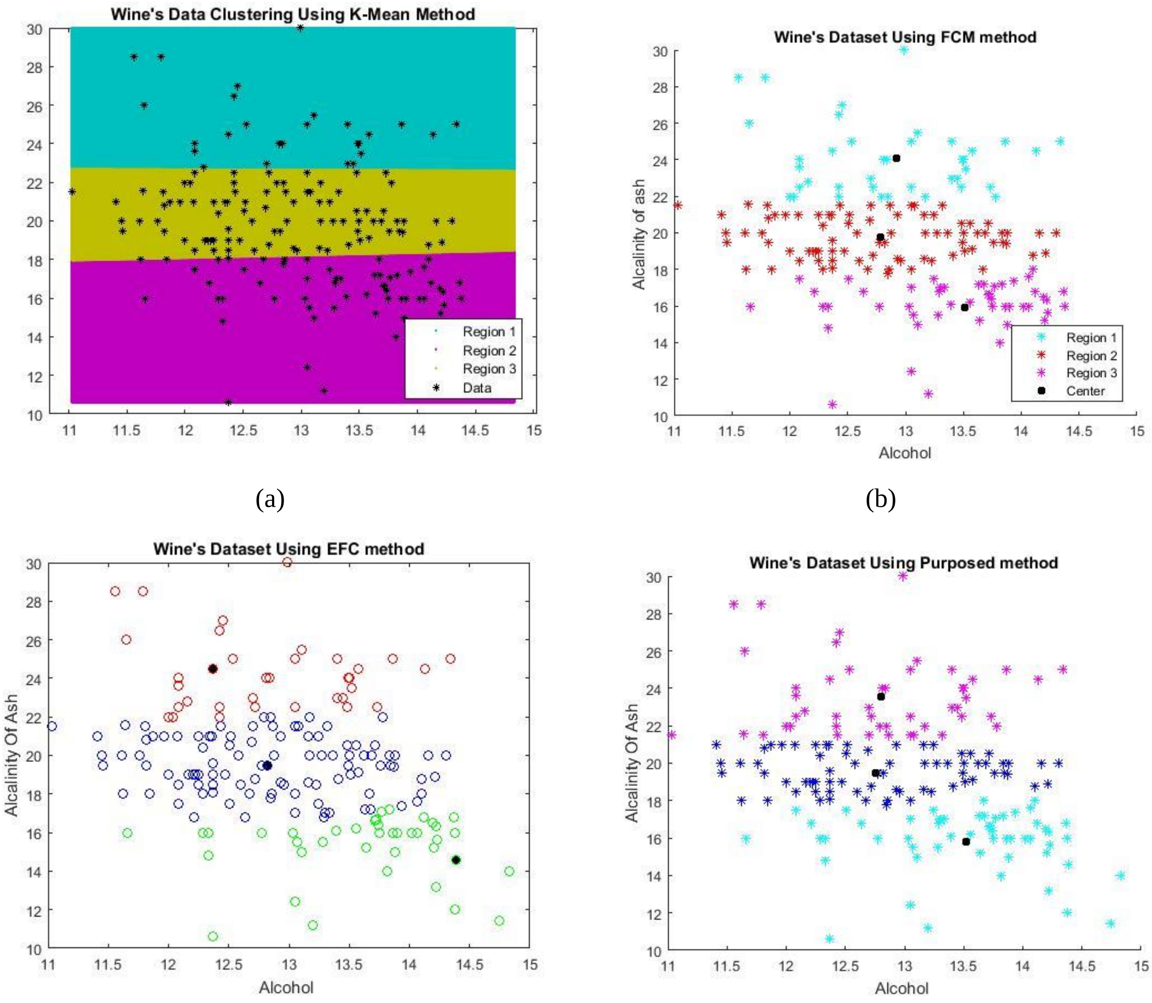
Figure 5.1 Clustering results on a Iris Dataset; (a) using k-means algorithm; (b) using FCM algorithm; (c) using EFC algorithm; (d) using Purposed method

Here, we cluster two attributes of Fisher Iris dataset i.e. Petal Length & Petal width in cm. with all four described algorithms. K-mean, FCM and purposed method classified data almost with same results, whereas EFC creates four clusters with one data point in forth cluster and having threshold value as 0.7. Purposed method takes 0.79 as threshold value. The centres and number of data points in each cluster obtained by above algorithm is described in below table:

Table 5.2.a Clustering Analysis on Iris Dataset

Algorithm	*Center1*	*Center2*	*Center3*	*DataPoints in Clusters*
K-Mean Method	13.44 42.82	2.46 14.62	19.88 55.92	[50,50,50]
FCM Method	13.76 43.58	2.50 14.72	20.36 56.69	[45,50,55]
EFC Method	13 40	2 16	24 56	[59,43,39,1]
Purposed Method	13.73 43.63	2.49 14.70	20.42 56.66	[44,50,56]

2. Wine DataSet:



(a)

(b)

(c)

(d)

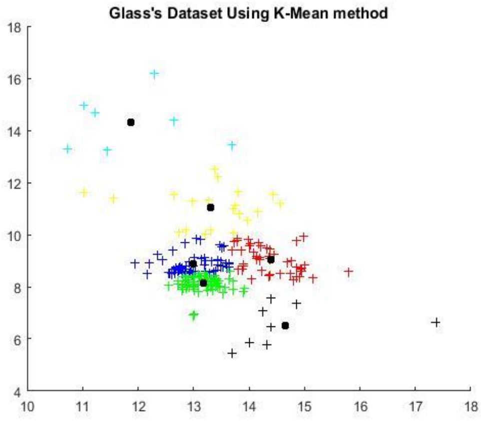
Figure 5.2 Clustering results on a Wine Dataset; (a) using k-means algorithm; (b) using FCM algorithm; (c) using EFC algorithm; (d) using Purposed method

In this, wine dataset with attributes 1. Alcohol and 2. Alkalinity Of Ash having 178 data points with three number of clusters is being classified with K-mean, FCM, EFC and purposed method. Table of obtained centres and clusters is as shown below:

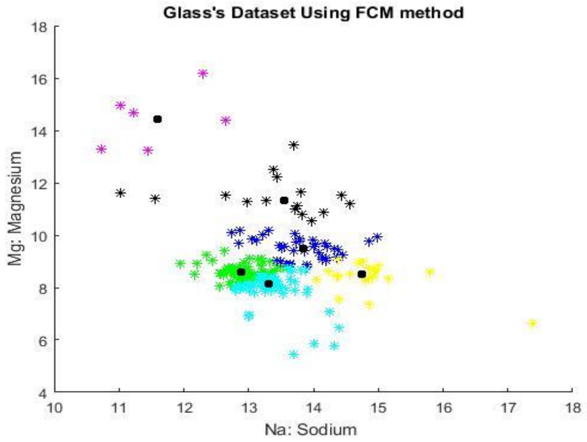
Table 5.2.b Clustering Analysis on Wine Dataset

Algorithm	*Center 1*	*Center 2*	*Center 3*	*DataPoints in Clusters*
K-Mean Method	12.73	13.53	12.88	[51,40,87]
	19.64	15.68	24.02	
FCM Method	12.78	13.51	12.92	[53,40,85]
	19.76	15.94	24.10	
EFC Method	12.81	13.64	12.84	[36,39,103]
	19.45	15.21	24.35	
Purposed Method	12.75	13.51	12.80	[52,41,85]
	19.49	15.80	23.55	

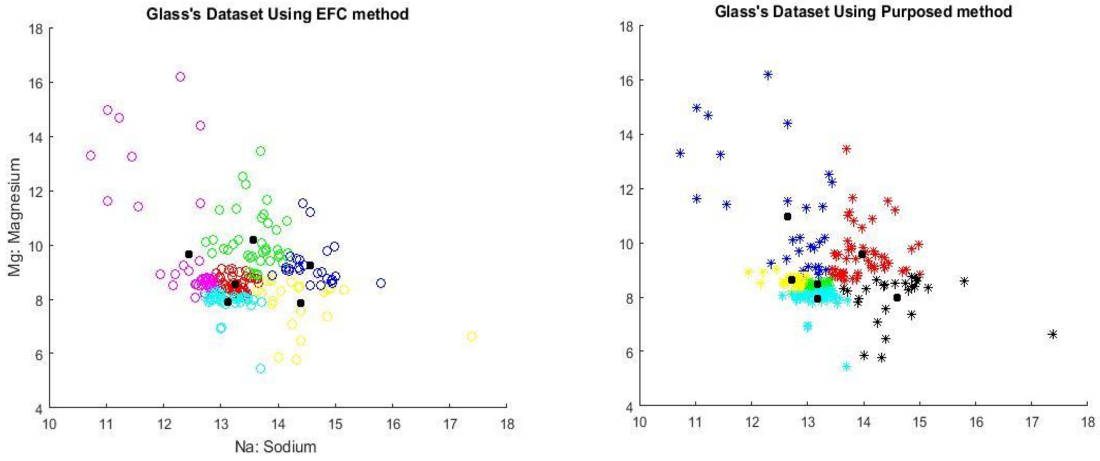
3. Glass DataSet:



(a)



(b)



(c) (d)

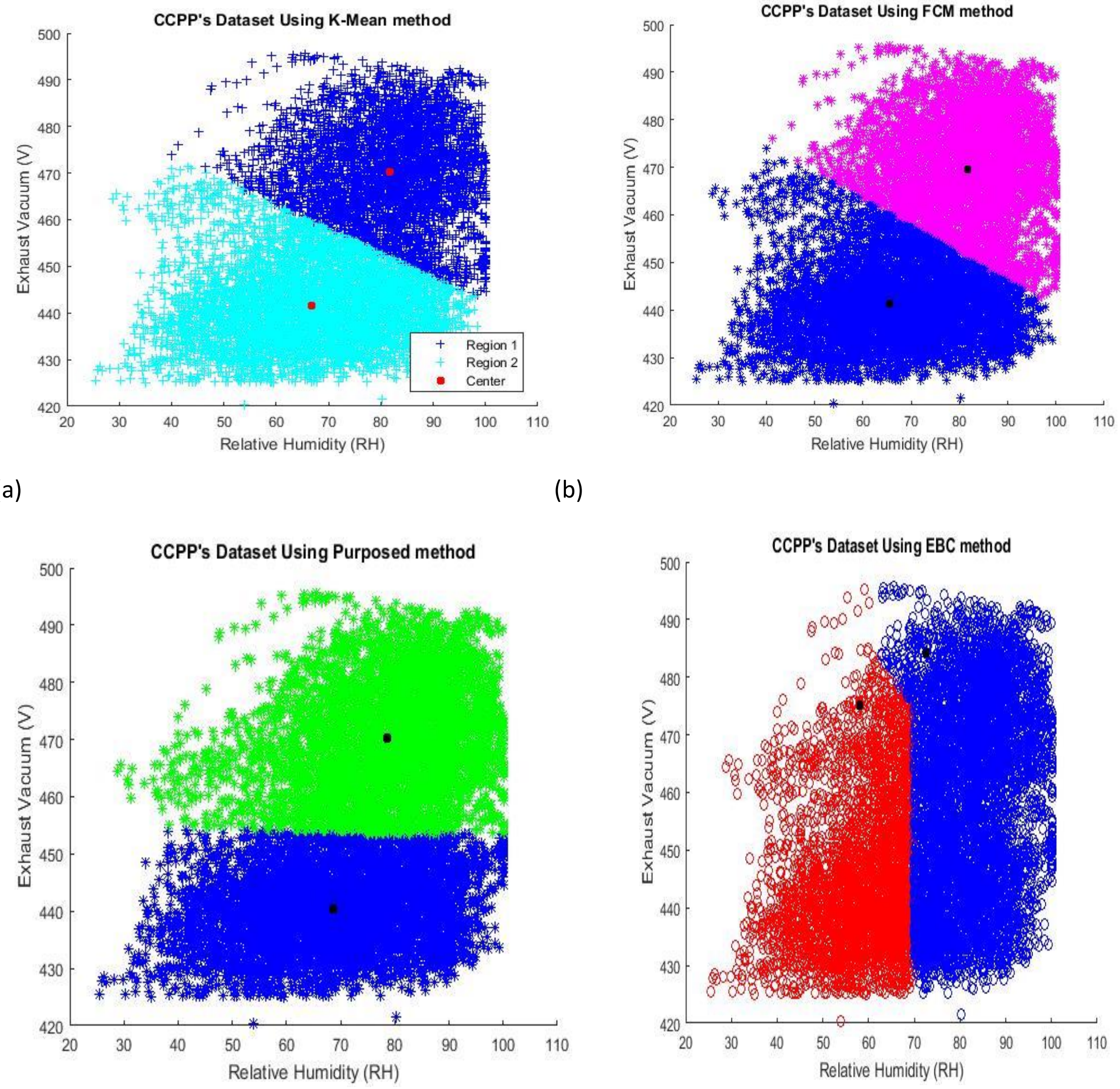
Figure 5.3 Clustering results on a Glass Dataset; (a) using k-means algorithm; (b) using FCM algorithm; (c) using EFC algorithm; (d) using Purposed method

We classified two attributes of Glass Identification dataset name as 1. Na: Sodium & 2. Mg: Magnesium with 214 number of instances. This dataset having six types of classes. Results obtained by different algorithms is shown in fig. 5.2 where six different colours represent all six clusters and black solid point as center of that respective cluster. Table of obtained centres and clusters for all algorithms is shown as:

Table 5.2.c Clustering Analysis on Glass Dataset

Algorithm	*Center 1*	*Center 2*	*Center 3*	*Center4*	*Center5*	*Center6*	*DataPointsinClusters*
K-Mean Method	13.10 8.42	11.56 14.46	14.55 8.86	13.34 11.56	14.33 6.61	13.52 9.75	[8,74,20,45,60,7]
FCM Method	12.99 8.88	11.86 14.31	13.31 11.05	13.18 8.14	14.66 6.56	14.38 9.05	[16,66,22,41,62,6]
EFC Method	13.25 8.54	12.44 10.78	13.56 10.17	13.33 9.87	13.13 7.89	14.56 9.24	[26,51,36,42,37,22]
Purposed Method	13.28 8.52	12.48 10.19	13.33 9.87	12.88 8.26	13.70 7.83	14.17 9.39	[15,61,36,45,48,9]

4. CCPP DataSet:



(a) (b)

(c) (d)

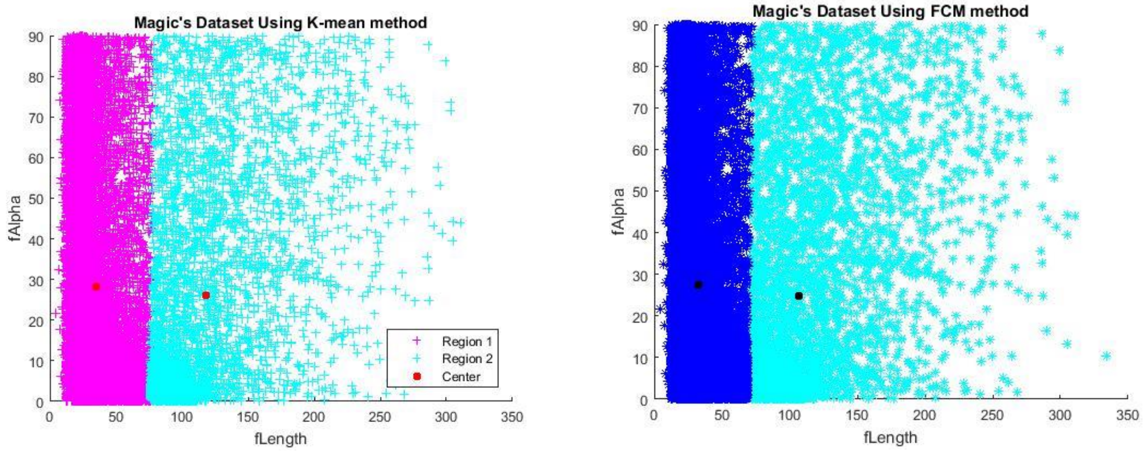
Figure 5.3 Clustering results on a Glass Dataset; (a) using k-means algorithm; (b) using FCM algorithm; (c) using EFC algorithm; (d) using Purposed method

In this, Relative Humidity and Exhaust Vacuum attributes of CCPP (Combined Cycle Power Point) are classified for visual representation. This dataset has 9568 number of instances which forms two classes. In fig 5.3, among all EFC gives worst results as compare to all others, then K-Mean, after that FCM and Purposed method gives best result in this case. Table of obtained Centres and number of data points in each class is mentioned below:

Table 5.2.d Clustering Analysis on CCPP Dataset

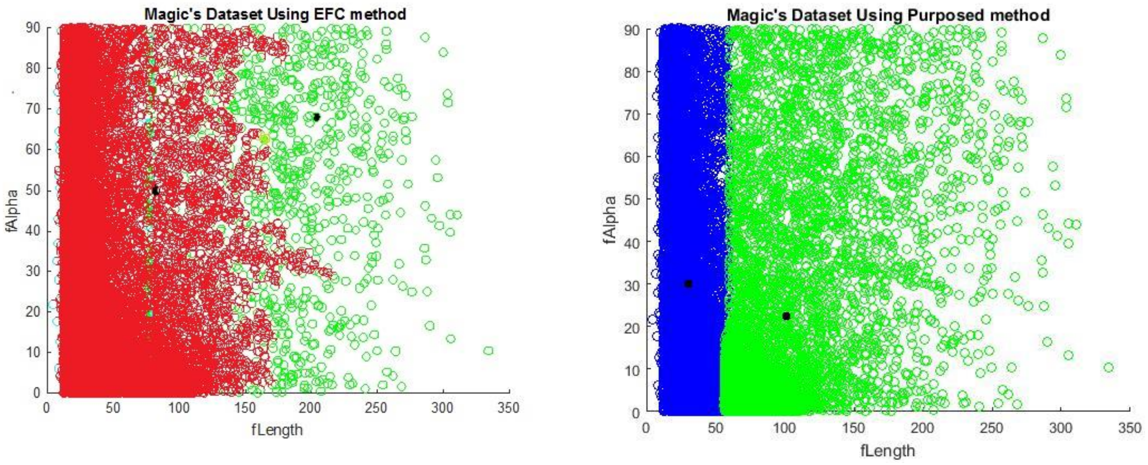
Algorithm	*Center1*	*Center2*	*DataPoints in Clusters*
K-Mean Method	66.72 441.65	81.72 470.36	[5331,4237]
FCM Method	65.60 441.27	81.78 459.60	[5182,4286]
EFC Method	58.10 475.13	72.47 484.20	[6026, 3542]
Purposed Method	68.60 441.64	78.78 470.12	[5171,4397]

5. Magic DataSet:



(a)

(b)



(c)

Figure 5.3 Clustering results on a Glass Dataset; (a) using k-means algorithm; (b) using FCM algorithm; (c) using EFC algorithm; (d) using Purposed method

MAGIC Gamma Telescope Dataset consists 19020 number of instances with 11 attributes. Here we clustered 1. Flenght and 2. Falpha attributes which forms 2 classes. For purposed algorithm threshold value 0.8 is taken. Two classes are shown with two different colour and black solid point is for center of the cluster. Obtained centres with their respective clusters from different methods are described in below table:

Table 5.2.e Clustering Analysis on Magic Dataset

Algorithm	*Center 1*	*Center 2*	*DataPoints in Clusters*
K-Mean Method	118.166 35.038	25.999 28.107	[4167,14853]
FCM Method	106.569 32.922	24.683 27.360	[4818,14202]
EFC Method	201.87 71.89	83.54 53.32	[2087,16933
Purposed Method	101.349 22.5187	30.192 30.102	[6163, 12857]

5.3 Error rate of clustering algorithms on the test datasets

Error rate is one of the external quality measures for clustering. It determines the falsity rate based on the misplaced data points with in a cluster. ER is calculated as:

$$ER = \frac{no.\,of\ \ misplaced\ odjects}{total\ no.\,objects\ within\ dataset} \times 100$$

As standard number of classes with their objects are given for each dataset, we can check this error rate by comparing standard results with our obtained results by different clustering methods. Table of each dataset with their clusters is as shown below:

Table 5.3.a List of distributed data points in a cluster

DataSets	*EFC*	*K-Mean*	*FCM*	*Purposed Method*
Iris [50,50,50]	[69,48,32]	[59,50,41]	[58,50,42]	[60,50,40]
Wine [59,71,48]	[43,89,46]	[62,69,47]	[61,71,46]	[61,69,48]
Glass [70,76,17,13,9,29]	[42,51,37,26,36,22]	[35,124,17,7,5,26]	[60,66,27,18,7,36]	[41,64,31,30,14,34]
Magic [12332,6688]	[17027, 1993]	[15180, 3840]	[11645, 7375]	[12857,6163]

Based on above table the results of error rate for tested datasets are as follows:

Table 5.3.b Error rates of clustering algorithms

DataSets	*EFC (%)*	*K-Mean (%)*	*FCM (%)*	*Purposed Method (%)*
Iris	12	6	5.33	6
Wine	10.67	3.37	2.24	2.24
Glass	28.03	22.42	11.21	19.15
Magic	24.76	14.97	3.61	2.76

From the above results, we can say that the proposed algorithm is best among all other test algorithms. Although results of FCM and purposed method are almost same but the point is, in FCM method we have to give the number of desired clusters manually. But as in our new method there is no need of giving cluster center explicitly. This method is design to keep in mind the problems of standard clustering algorithm like k-mean and FCM i.e. these methods can’t produce initial cluster centres and due to reason of initialization of center points or membership matrix (in case of FCM) these methods may leads the results in local minima. In other words, the EBAC algorithm converges to global optimum while the other algorithms may get stuck in local optimum solutions. Entropy based Fuzzy Clustering is also provides automatic centres but there are lot of drawbacks in that algorithm as described earlier.

5.4 The centroids obtained by the EBAC algorithm on the test datasets

Description of best centroids obtained by the Entropy Based Automatic Clustering for all test dataset is shown in below tables:

Table 5.4. a) Iris Dataset; Threshold Value: 0.84

Center1	Center2	Center3
14.7843137254902	4.14516129032258	20.6750000000000
46.0588235294118	18.7096774193548	57.1750000000000
28.2156862745098	32.3387096774194	30.6250000000000
60.3333333333333	50.6612903225807	68.1000000000000

Table 5.4.b Wine Dataset; Threshold Value: 0.87

Center 1	Center2	Center3
12.831020	13.567857	12.495967
2.6765306	2.0524285	2.4404838
2.3983673	2.4215714	2.2788709
20.681632	17.627142	20.714516
97.204081	107.94285	92.564516
1.9628571	2.6805142	2.1064516
1.2957142	2.6667142	1.8561290
0.4218367	0.3042857	0.3806451
1.3675510	1.8325714	1.4874193
5.7406122	5.5612857	4.0361290
0.8642857	1.0243714	0.9488709
2.1887755	2.9822857	2.5114516
651.97959	1078.0714	443.75806

Table 5.4.c Glass Dataset: Threshold Value: 0.932

Center1	Center2	Center3	Center4	Center5	Center6
1.517693	1.520226	1.518862	1.516224	1.521167	1.515856
13.06984	13.79774	13.17314	13.07452	13.81093	14.66066

3.475230	2.508064	0.918285	3.361190	2.965000	1.525333
1.303692	1.584516	1.779142	1.595714	0.906250	1.666666
72.78107	71.73580	73.18914	72.85952	72.18000	73.19000
0.566000	0.449354	0.372571	0.900238	0.103125	0.298666
8.587692	9.456129	9.996571	7.956666	9.839687	7.829333
0.002769	0.313870	0.404285	0.029761	0.059375	0.700666
0.060153	0.050000	0.080000	0.053571	0.059375	0.032000

Table 5.4.c CCPP Dataset: Threshold Value: 0.87

Center1	Center2
25.128326	12.478315
63.772300	41.909099
1011.1447	1016.0274
68.940947	79.028708
441.52237	471.18174

Table 5.4.d Magic Dataset: Threshold Value: 0.9

Center1	Center2
72.94	37.89
28.16	17.41
0.30	2.65
0.16	0.43
-19.22	0.24
19.22	6.77
0.39	0.13
21.08	30.58
256.48	152.87

We execute test datasets with their all attributes and above are the obtained centres with respect to datasets.

CHAPTER – 6

Conclusion & Future work

We dedicated our work in continuation of Entropy based Fuzzy Clustering algorithm for clustering large data sets. It puts to advantage; the initial parameter free nature of EBAC algorithm. Derived results also support the idea that purposed method is parameter free and is easy to implement and produces better results than conventional K-Means algorithm, FCM method without pre-knowledge of number of cluster centres on five benchmark datasets. Major thing of this algorithm is to handle threshold value for similarity measures in order to create centres. Although in all test cases its value lies between 0.8 to 0.9. This number indicates the closeness between data points and its centre point. Threshold value depends on the similarity measure and datasets itself. The future work will be to find out any automatic method to calculate threshold value.

BIBLIOGRAPHY

[1] D. W. Rui Xu, "Survey of Clustering Algorithms," *IEEE TRANSACTIONS ON NEURAL NETWORKS,* Vols. VOL. 16, NO. 3, pp. 645-678, MAY 2005.

[2] N. P. Jasmine Irani, "Clustering Techniques and the Similarity Measures used in Clustering: A Survey," *International Journal of Computer Applications,* vol. Volume 134 – No.7, January 2016.

[3] M. B. H. Abdelkarim Ben Ayed, "Survey on clustering methods : Towards fuzzy clustering for big data," in *International conference on Soft computing and Pattern recognition,IEEE,* 2014.

[4] D. M. U. B.G.Obula Reddy1, "Literature Survey On Clustering Techniques," *IOSR Journal of Computer Engineering (IOSRJCE),* vol. 3, no. 1, pp. 1-12, Aug, 2012.

[5] D. M. U. B.G.Obula Reddy1, "Literature Survey On Clustering Techniques," *IOSR Journal of Computer Engineering,* vol. 3, no. 1, pp. 1-50, July-Aug. 2012.

[6] L. S.-h. T. Y.-n. Tao Huang, "Research of Clustering Algorithm Based on K-means," *Computer technology and development,* vol. 7, 2011.

[7] H.-P. K. M. Ester, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD,* vol. 62, pp. 226-231, 1996.

[8] M. M. B. H.-P. K. J. S. M. Ankerst, "Optics: ordering points to identify the clustering structure," *ACM Sigmod Record,* vol. 28, pp. 49-60, 1999.

[9] M. E. H.-P. K. J. S. Xiaowei Xu, "A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases," in *Proceedings 14th International Conference on Data Engineering*, USA, Feb 1998.

[10] L. Y. Du HaiZhou, "An improved BIRCH clustering algorithm and application in thermal power," in *International Conference on Web Information Systems and Mining, IEEE,* 2010.

[11] D. R. R. Piyush Lathiya, "Improved CURE Clustering for Big Data using," in *International Conference on Inventive Computation Technologies (ICICT),* Coimbatore, India, January 2017.

[12] J. Y. M. W. Wang, "Sting: A statistical information grid approach to spatial data mining," *Proc. of the International Conference Very Large Data Bases (VLDB),* pp. 186-195, 1997.

[13] S. C. A. Z. Gholamhosein Sheikholeslami, "WaveCluster: a wavelet-based clustering approach for spatial data," *The VLDB Journal,* vol. 8, p. 289–304, July,1999.

[14] N. Grover, "A study of various Fuzzy Clustering Algorithms," *International Journal of Engineering Research,* vol. 3, no. 3, pp. 177-181, Mar,2014.

[15] G.-L. C. Shui-Li Chen1, "A Fuzzy Neural Network Model Based on Fuzzy Clustering and Its," in *Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering,* 2008.

[16] M. C. C. A. R. J. HARVEY, "A Genetic Algorithm Approach," *Computers and Mathematics with Applications,* vol. 99, pp. 99-108, 1999.

[17] S. M. Timothy C. Havens, "Fuzzy c-Means Algorithms for Very Large Data," *IEEE TRANSACTIONS ON FUZZY SYSTEMS,*, vol. 20, no. 6, pp. 1130-1146, 2012.

[18] M. D. S. T. H. L. J. Yao, "Entropy-based fuzzy clustering and fuzzy modeling," *Fuzzy Sets and Systems,* vol. 113, pp. 381-388, 2000.

[19] A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," *Information Sciences,* vol. 222, pp. 175-184, Feb,2013.

[20] F.-L. C. Jiefang Liu, "Black Hole Entropic Fuzzy Clustering," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS,* 2017.

[21] T. ZHANG, "BIRCH: A New Data Clustering Algorithm and Its," *Springer,* vol. 1, pp. 141-182, 1997.