Dealing with Overlapping of Clusters by using Fuzzy K-Means Clustering

Chetana T. Baviskar¹, Sandip S. Patil²

¹Research Scholar Department of Computer Engineering, SSBT’s COET Bambhori, Jalgaon, India
²Associate Professor Department of Computer Engineering, SSBT’s COET Bambhori, Jalgaon, India

Abstract: Clustering is a process of finding classes of a data set with most similarity in the same group and most dissimilarity between different groups. Therefore, a cluster is usually represented as either grouping of similar data points around a centre. Clustering divides a given data set into a set of clusters. Clustering approaches can be classified into two categories namely: Hard clustering and Soft clustering. The hard clustering restricts that the data object in the given data belongs to exactly one cluster. The problem with hard clustering is that the different initial partitions can result in different final clusters. Soft clustering which also known as fuzzy clustering forms clusters such that data object can belong to more than one cluster to some degree known as membership levels. But sometimes the resulting membership values do not always correspond well to the degrees of belonging of the data which may results in to overlapping clusters. So the proposed fuzzy clustering approach assigns membership such as to deal with overlapping of clusters.

Keywords— Hard Clustering, Soft Clustering, Fuzzy Clustering, Fuzzy Set, Membership, Fuzzy Partition, Overlapped clusters.

I. INTRODUCTION

Clustering can be used to quantize the available data, to extract a set of cluster prototypes for the compact representation of the dataset, into homogeneous subsets [1]. Clustering approaches can be classified into two categories namely- Hard clustering and Soft clustering (Fuzzy Clustering). The hard clustering technique [3] forms uniquely defined cluster by grouping related attributes. In hard clustering given data's are divided into distinct clusters. That is, it can be see clearly whether an object belongs to a cluster or not. In other words, it means that every element must be uniquely classified as belonging to the cluster or not.

Clustering implies that the data set to be partitioned, to have an inherent grouping to some extent. Otherwise if the data is uniformly distributed, trying to find clusters of data will fail, or will lead to artificially introduced partitions [4]. However, such a partition is insufficient to represent many real situations. Fuzzy clustering is one method which can deal with the uncertainty situation of real data. In fuzzy clustering clusters are constructed with uncertain boundaries, so that to allows one object belongs to some overlapping clusters to some degree. In fuzzy clustering a cluster is associated with each element having a set of membership levels [5]. Membership levels are then used for assigning elements to one or more clusters. In other words, the essence of fuzzy clustering is to consider not only the belonging status to the clusters, but also to consider to what degree do the objects belong to the clusters. In fuzzy set theory an element can belong to a set in a degree k (0 ≤ k ≤ 1), in contrast to classical set theory where an element must definitely belong or not to a set.

Fuzzy clustering is a partition based clustering scheme [6] and is particularly useful when there are no apparent clear groupings in the data set. Partitioning schemes provide automatic detection of cluster boundaries and in case of fuzzy clustering, these cluster boundaries overlap. Every individual data entity belongs to not one but all of the clusters with varying degrees of membership.

Fuzzy clustering has a number of useful properties as follows:

- It provides membership values which are useful for interpretation.
- It is flexible with respect to distance used.
- If some of the membership values are known this can be incorporated into the numerical optimisation.

The main difference between fuzzy clustering and other clustering techniques is that it generates fuzzy partitions of the data instead of hard partitions. Therefore, data items may belong to several clusters, having in each cluster different membership values.

II. RELATED WORK

Clustering and classification are two main domains of data mining. The clustering is unsupervised learning and classification is supervised learning [1]. Clustering is divided into two categories namely hard clustering and soft clustering (fuzzy clustering).

A. Hard Clustering

Hard clustering is also known as crisp clustering. It allocate each data pattern (data object) of given input to a single cluster. Thus in hard clustering, each data pattern (data object) belongs to only one cluster. Farley and Raftery in [7], suggest dividing the clustering methods into two main groups: partitioning and hierarchical methods. Han and Kamber in [1], suggest categorizing the methods into additional three main categories: density-based methods, model-based clustering and grid-based methods.
Partitioning Clustering

Partitioning clustering [8], directly divides data objects into some pre-specified number of clusters. The checking for all possible clusters is computationally impractical, certain greedy heuristics are used in the form of iterative optimization of cluster. The partitioning clustering consists of several approaches such as K-means Clustering, K-medoid Clustering, Relocation Algorithms, Probabilistic Clustering.

K-means clustering [9], is a method commonly used to automatically partition a data set into clusters (K). Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters. The distances used in clustering in most of the times do not actually represent the spatial distances.

In the K-medoid clustering [1], a cluster is represented by one of its points called medoid. A medoid is the centrally located data point. When medoids are selected, clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid. Every time a new medoid is selected, the distance between each object and its newly selected cluster centre has to be recomputed.

The Relocation algorithms [2] iteratively reallocate points between the k clusters. The points are reassigned based on the local search algorithm. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The three changeable elements of the general relocation algorithm are initialization, reassignments of the data points into clusters and update of the cluster parameters.

In the probabilistic approach [10], data is considered to be a sample independently drawn from a mixture model of several probability distributions. So the clusters are associated with the corresponding distributions parameters such as mean and variance.

Hierarchical Clustering

A hierarchical clustering [11], creates a hierarchical decomposition of the given set of data objects. It builds a cluster hierarchy, a tree of cluster, also known as a dendrogram. By cutting the tree at a desired level, a clustering of the data points into disjoint groups is obtained. A hierarchical clustering is used to find data on different levels of dissimilarity. A hierarchical clustering can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster.

Density-Based Algorithms

Density-based algorithms [2], are capable of discovering clusters of arbitrary shapes. These algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighbourhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighbourhood exceeds some parameter. In Density-Based connectivity clustering technique density and connectivity both measured in terms of local distribution of nearest neighbours. In Density Functions Clustering the density function is used to compute the density. Overall density is modeled as the sum of the density functions of all objects.

Model-Based Clustering

Model-based methods [1], hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number of clusters based on standard statistics, taking noise or outliers into account and thus yielding robust clustering methods.

Grid-Based Clustering

Grid-based clustering [1], quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.

B. Fuzzy Clustering

Fuzzy clustering [12], is the synthesis between the fuzzy logic and clustering which is the requirement of modern computing. The aim of fuzzy clustering is to model the ambiguity within the unlabeled data objects efficiently. Every data object is assigned a membership to represent the degree of belonging to certain class. The requirement that each object is assigned to only one cluster is relaxed to weaker requirement in which the object can belong to all of the clusters with a certain degree of membership. Thus it assigns degrees of membership in several clusters to each input pattern. Fuzzy clustering is categorized in three categories [5]: Fuzzy relation based clustering, fuzzy rule based clustering, objective function based clustering.

Fuzzy Relation Based Clustering

Fuzzy relation based clustering is based on fuzzy relations [5]. Liang et. al., in [13], introduced cluster analysis based on fuzzy equivalence relation. The approach used is the distance measure between two
trapezoidal fuzzy numbers is used to aggregate subjects linguistic assessments. The distance measure is used to characterize the interobjects similarity. The linguistic assessment is for attributes ratings to obtain the compatibility relation. Then a fuzzy equivalence relation based on fuzzy compatibility relation is constructed. The algorithm then uses a cluster validity index $A$ to determine the best number of clusters. Then a suitable $\lambda$ cut value is taken based on the fuzzy compatibility relation. The cluster analysis is done through the $\lambda$ cut value and a cluster validity index. This algorithm is finds out the best number of clusters and the objects belonging to the clusters. The cluster validity index measure is defined as shown in Equation 2.1:

$$A = \frac{R}{n \times (d_{\text{min}})^2}$$  \... (2.1)

Where,

$A$ : Cluster validity index  
$R$ : Normalized compatibility relation  
n: Number of objects

- **Fuzzy Rule Based Clustering**
  The fuzzy rules are generated for the given input data set and by generated fuzzy rules the inference is drawn to form the clusters [5]. Eghbal G. Mansoori in [14], introduced Fuzzy Rule-Based Clustering Algorithm (FRBC). The algorithm automatically explore the potential clusters in the data patterns with no prior knowledge. The approach used in the algorithm is generation of fuzzy rules with the help of genetic algorithm and then using these rules the clusters are formed. Each rule is used as the cluster formation measure. The obtained clusters are specified with some interpretable fuzzy rules, which make the clusters understandable for humans.

- **Objective Function Based Fuzzy Clustering**
  Objective function is used as a dissimilarity measure in fuzzy clustering. Thanh Le et al., in [15], have proposed Fuzzy $C$-means algorithm. Fuzzy $C$-means (FCM) algorithm allows data objects to belong to multiple clusters based on the degree of membership. In real-world datasets, data objects are not equally distributed among clusters and the clusters will differ in size. It offers the opportunity to deal with the data that belong to more than one cluster at a time. It assigns memberships to a data object which are inversely related to the relative distance of the data object to cluster prototypes. However the resulting membership values of fuzzy $c$-means do not always correspond well to the degrees of belonging of the data and it may be inaccurate in a noisy environment.

  Georg Peters, in [16], have proposed Rough $k$-means clustering algorithm. Incorporating rough set into $K$-Means clustering requires the additional concept of lower and upper bounds. Calculation of centroid of clusters need to be modified to include the effects of lower as well as upper bounds. Selection of the initial parameters in rough $k$-means is more delicate. The rough $k$-means produces crisp intervals, the lower and upper approximations for the clusters while fuzzy clustering assigns the data objects to fuzzy clusters. Therefore small changes in the initial parameters can result in the assignment of a data object to different approximations while in fuzzy clustering the memberships only gradually change.

### III. Proposed Work

The approach used in proposed system is first the preprocessing of data is done. Then the preprocessed data is given as an input to the conventional hard clustering algorithm and proposed fuzzy clustering algorithm. Finally, the cluster formation is done.

- **A. Problem Statement**
  Clustering is nothing but partition of given data sets into clusters such that there is a larger similarity between data in the same group as compared to those among other groups. Clustering forms the group of similar objects. Traditional clustering methods states whether an object belongs to a cluster or not. Cluster result is sensitive to the selection of the initial cluster centroids and make it insufficient to represent natural data which is often uncertain [2]. Therefore fuzzy clustering method is offered to construct clusters with uncertain boundaries. It assigns the membership values to each data item showing that the degree to which data item belongs to a cluster. Fuzzyness allows one object belongs to overlapping clusters to some degree. So integration of clustering and fuzzyness is proposed to deal with problems where points are somewhat in between centres and hence reduces the ambiguity in the assignment of objects to clusters.

- **B. Objectives**
  - To initialize proper cluster centre.
  - To deal with cluster overlapping.

- **C. Assumptions**
  The proposed system requires some assumptions described as follows:

  - **Hard Partition**
    The structure of the partition matrix $U = [\mu_{ik}]$

    $$M = [\mu_{1,1} \mu_{1,2} \ldots \mu_{1,c} \mu_{2,1} \mu_{2,2} \ldots \mu_{2,c} \ldots \mu_{n,1} \mu_{n,2} \ldots \mu_{n,c}]$$

  Hard Partition [17] can be defined as a family of subsets. The objective of clustering is to partition the data set $X$ into $c$ clusters. Assume that is $c$ known, based on prior knowledge, using classical sets, a hard partition of $X$ can be defined as:

  $$\{A_i | 1 \leq i \leq c \subseteq P(X)\}$$

  Its properties are as follows:
Equation (3.1) states that the union subsets $A_i$ contains all the data. The subsets must be disjoint, as stated by Equation (3.2) and none of them is empty nor contains all the data in $X$ shown in Equation (3.3).

In terms of membership functions, a partition can be conveniently represented by the partition matrix $U = [\mu_{ik}]_{c \times N}$. The $i^{th}$ row of this matrix contains values of the membership function $\mu_i$ of the $i^{th}$ subset $A_i$ of $X$. The elements of $U$ must satisfy the following conditions:

$$\mu_{ik} \in [0,1], 1 \leq i \leq c, 1 \leq k \leq N \quad \text{....(3.4)}$$

$$\sum_{i=1}^{c} \mu_{ik} = 1, 1 \leq k \leq N \quad \text{....(3.5)}$$

$$0 < \sum_{k=1}^{N} \mu_{ik} < N, 1 \leq i \leq c \quad \text{....(3.6)}$$

### Fuzzy Partition

Generalization of the hard partition to the fuzzy case follows directly by allowing $\mu_{ik}$ to attain real values in $[0,1]$. A more general form of fuzzy partition [17], the possibilistic partition, can be obtained by relaxing the constraint in Equation (3.5). This constraint, however, cannot be completely removed, in order to ensure that each point is assigned to at least one of the fuzzy subsets with a membership greater than zero. Equation (3.5) can be replaced by a less restrictive constraint $\forall k, \exists i, \mu_{ik} > 0$. The conditions for a possibilistic fuzzy partition matrix are:

$$\mu_{ik} \in [0,1], 1 \leq i \leq c, 1 \leq k \leq N \quad \text{....(3.7)}$$

$$\exists i, \mu_{ik} > 0, \forall k \quad \text{....(3.8)}$$

$$0 < \sum_{k=1}^{N} \mu_{ik} < N, 1 \leq i \leq c \quad \text{....(3.9)}$$

### Euclidean Distance

The proposed system uses the euclidean distance measure [1], which calculates the distance between the two $n$-dimensional data objects.

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{in} - x_{jn})^2} \quad \text{....(3.10)}$$

Where,

- $i = (x_{i1}, x_{i2}, ..., x_{in})$ $n$-dimensional data object.
- $j = (x_{j1}, x_{j2}, ..., x_{jn})$ $n$-dimensional data object.

### D. Architecture

The architecture of the proposed system is shown in Fig 1. Input to the proposed system is the Data set. Pre-processing of the input data set includes the analysis of the raw data in the input data set. Raw data is highly concerned with noise, missing values and inconsistency and the quality of data affects the data.
mining results. In order to improve the quality of data and consequently of the mining results, raw data is pre-processed so as to improve the efficiency and ease of mining process. After data pre-processing each data object is given as an input to the hard clustering and fuzzy clustering.

In hard clustering, the data objects are assigned to the clusters on the basis of binary membership values i.e., either 0 or 1. The hard clustering takes the input parameter k and partitions a set of m objects into k clusters. The technique work by computing the distance between a data object and the cluster center by using euclidian distance. The distance is used to add an object into one of the clusters so that intra-cluster similarity is high but inter-cluster similarity is low.

In fuzzy clustering, a data object will have an associated degree of membership for each cluster in the range [0 – 1], indicating the strength of its association in that cluster. It iteratively update the membership values of a data object with the predefined number of clusters. Thus, a data object can be the member of all clusters with the corresponding membership values. The process of calculation of cluster centres and the assignment of points to these centres is repeated until the cluster centres stabilize.

**Hard K-Means Clustering**

A hard partition of the data set X can be represented by a matrix \( U = [\mu_{ij}]_{kn} \), where \( \mu_{ij} \) denotes the membership with which \( j^{th} \) data object belongs to the \( i^{th} \) cluster, for \( 1 \leq i \leq k; 1 \leq j \leq n \). The matrix \( U \) is called the crisp partition matrix. It satisfies the following constraints shown in Equation (3.11)

\[
\mu_{ik} \in [0,1], 1 \leq j \leq n, 1 \leq i \leq k \quad ...(3.11)
\]

In the classical k-means (hard) clustering an object is assigned to exactly one cluster. As outcome of clustering process, it is possible that all the data objects may get assigned to a few clusters only while other clusters may be empty. Such an output is undesirable. In order to avoid such situations the following constraint is used as shown in Equation (3.12):

\[
0 < \sum_{j=1}^{n} \mu_{ij} < n, 1 \leq i \leq k \quad ...(3.12)
\]

The constraint shown in Equation (3.12) ensures that the number of clusters is at least two and none of the clusters is empty.

\[
\sum_{i=1}^{k} \mu_{ij} = 1, 1 \leq j \leq n \quad ...(3.13)
\]

The above constraint shown in Equation (3.13) expresses the fact that the sum of memberships of a data object over the set of clusters must be equal to 1. In the context of hard k-means it means that \( \mu_{ij} \) takes the value 1 for the specific cluster to which the \( j^{th} \) object is assigned and zero for all other clusters.

The K-means clustering, or Hard C-means clustering, is an algorithm based on finding data clusters in a data set such that a cost function of dissimilarity (or distance) measure is minimized. The dissimilarity measure is chosen as the Euclidean distance.

A set of n vectors \( X_j, j = 1,...,n \), are to be partitioned into c groups \( G_l, l = 1,...,c \). The cost function, based on the Euclidean distance between a vector \( X_k \) in group j and the corresponding cluster center \( C_l \), can be defined by Equation (3.14):

\[
J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left( \sum_{k \in G_l} \|X_k - C_l\|^2 \right) \quad ...(3.14)
\]

Where,

\[
J_i = \sum_{k, X_k \in G_l} \|X_k - C_l\|^2 \quad \text{is the cost function within} \ i.
\]

The partitioned groups are defined by \( c \times n \) binary membership matrix \( U \), where the element \( \mu_{ij} \) is 1 if the \( j^{th} \) data point \( X_j \) belongs to group i, and 0 otherwise. Once the cluster centers \( C_l \) are fixed, the minimizing \( \mu_{ij} \) for Equation (3.14) can be derived as follows:

\[
\mu_{ij} = \begin{cases} 1 & \text{if } \|X_j - C_i\|^2 \leq \|X_k - C_k\|^2, \text{for each } k \neq i \\ 0 & \text{Otherwise} \end{cases} \quad ...(3.15)
\]

Which means that \( X_j \) belongs to group i if \( C_i \) is the closest center among all centers.

On the other hand, if the membership matrix is fixed, i.e. if \( \mu_{ij} \) is fixed, then the optimal center \( C_l \) that minimize Equation (3.14) is the mean of all vectors in group l.

\[
C_l = \frac{1}{|G_l|} \sum_{k, X_k \in G_l} X_k \quad ...(3.16)
\]

Where,

\[
|G_l| \quad \text{is the size of } |G_l| \\
|G_l| = \sum_{i=1}^{n} \mu_{ij}
\]

**Fuzzy K-Means Clustering**

In fuzzy clustering, an object can belong to several clusters simultaneously, with different degrees of membership. However, Fuzzy K-Means still uses a cost function that is to be minimized while trying to partition the data set.

A fuzzy partition of the data set X can be represented by a matrix \( U = [\mu_{ij}]_{kn} \), where \( \mu_{ij} \) denotes the membership with which \( j^{th} \) data object belongs to the \( i^{th} \) cluster, for \( 1 \leq i \leq k; 1 \leq j \leq n \). The matrix \( U \) is called the fuzzy partition matrix. It satisfies the following constraints shown in Equation (3.17)
\( \mu_{ik} \in [0,1], 1 \leq j \leq n, 1 \leq i \leq k \) … (3.17)

The maximum total membership possible in a cluster for all the \( n \) data points is \( n \). It is the ideal possible membership for all the data points in a cluster as shown in Equation (3.18).

\[ \sum_{i=1}^{n} \mu_{ij}(x_i) = n \quad \ldots (3.18) \]

The cost function shown in Equation (3.19) for Fuzzy K-Means is a generalization of Equation (3.14):

\[ J(U, C_1, \ldots, C_c) = \sum_{i=1}^{c} I_i = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^m d_{ij} \ldots (3.19) \]

Where,

\( \mu_{ij} \): is membership of data point,

\( C_i \): is cluster centre of fuzzy group \( i \).

\( d_{ij} = \| C_i - x_j \| \): is Euclidean distance between \( i^{th} \) cluster centre and \( j^{th} \) data point.

\( m \): is fuzzification parameter or weighting exponent

The necessary conditions for Equation (3.19) to reach its minimum are given in Equation (3.20) and Equation (3.21):

\[ C_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m x_j}{\sum_{j=1}^{n} \mu_{ij}^m} \quad \ldots (3.20) \]

And

\[ \mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^{2/m-1}} \quad \ldots (3.21) \]

The fuzziness coefficient \( m \), where \( 1 < m < \infty \), measures the tolerance of the required clustering. This value determines how much the clusters can overlap with one another. The higher the value of \( m \), the larger the overlap between clusters. In other words, the higher the fuzziness coefficient the algorithm uses, a larger number of data points will fall inside a fuzzy band where the degree of membership is neither 0 nor 1, but somewhere in between.

With the liberalized constrain given in Equation (3.18), the maximum membership value generated by the new fuzzy membership function is not limited to one. When conventional fuzzy membership distributions are required, (within a range of zero to one) these values can be generated by the process of normalization. Normalization finds the maximum memberships among all clusters and rescales the memberships from this maximum values using Equation (3.22).

\[ \mu_{ik}^{\text{norm}}(x_i) = \frac{\mu_{ik}^{\text{old}}(x_i)}{\max_{k} \mu_{ik}^{\text{old}}} \quad i = 1 \text{ to } n; \quad k = 1 \text{ to } p \]

Where,

\( \mu_{ik}^{\text{norm}}(X_i) \): is normalized membership of \( X_i \) in the \( k^{th} \) cluster.

\( \mu_{ik}^{\text{old}}(X_i) \): is the old membership.

\( p \): is the no. of specified clusters.

\( n \): is the no. of data points.

max(): returns maximum membership value in \( k^{th} \) cluster.

**Termination Condition**

The required accuracy of the degree of membership determines the number of iterations completed by the FKM algorithm. This measure of accuracy is calculated using the degree of membership from one iteration to the next, taking the largest of these values across all data points considering all of the clusters. The measure of accuracy between iteration \( k \) and \( k + 1 \) with \( \varepsilon \), is calculated as follows:

\[ \varepsilon = \delta_{ij}^k \left( \delta_{ij}^{k+1} - \delta_{ij}^k \right) \quad \ldots (3.23) \]

**IV. Conclusions**

Partitions generated in conventional crisp (hard) clustering algorithms assigns each object to exactly one cluster. As objects are located between clusters, it cannot be assigned to strictly one cluster. Membership degrees between zero and one are used in fuzzy clustering instead of crisp assignments of the data to clusters. As one object belongs to more than one cluster the overlapping of cluster occurs. In these cases, fuzzy clustering methods provide a much more adequate tool for representing data structures. There is no explicit membership function, but only a membership matrix which contains membership values for the specific data points upon which the clustering algorithm has been applied. Once, the membership values for the data set have been obtained, it is possible to produce an approximation of the corresponding membership functions for each cluster. In real applications there is very often no sharp boundary between clusters so that fuzzy clustering is often better suited for the data. It seems that non-zero membership degree property of proposed fuzzy k-means clustering has smoothing effect in the handling of overlapped clusters.

**REFERENCES**


