A Novel Penalized and Compensated Constraints based Modified Fuzzy Possibilistic C-Means for Data Clustering

Duraisamy K.¹ Haridass K.²

¹Research Scholar ²Assistant Professor& Head of Department
1Department of Computer Science ²Department of Computer Application
1,2NGM College Pollachi India

Abstract— A cluster is a group of objects which are similar to each other within a cluster and are dissimilar to the objects of other clusters. The similarity is typically calculated on the basis of distance between two objects or clusters. Two or more objects present inside a cluster and only if those objects are close to each other based on the distance between them. The major objective of clustering is to discover collection of comparable objects based on similarity metric. Fuzzy Possibilistic C-Means (FPCM) is the effective clustering algorithm available to cluster unlabeled data that produces both membership and typicality values during clustering process. In this approach, the efficiency of the Fuzzy Possibilistic C-means clustering approach is enhanced by using the penalized and compensated constraints based FPCM (PCFPCM). The proposed PCFPCM differs from the conventional clustering techniques by imposing the possibilistic reasoning strategy on fuzzy clustering with penalized and compensated constraints for updating the grades of membership and typicality. The performance of the proposed approaches is evaluated on the University of California, Irvine (UCI) machine repository datasets such as Iris, Wine, Lung Cancer and Lymphograma. The parameters used for the evaluation is Clustering accuracy, Mean Squared Error (MSE), Execution Time and Convergence behavior.

Key words: Unsupervised Learning, Fuzzy C-Mean, Fuzzy Possibility C-Means, Penalized and Compensated constraints based FPCM

I. INTRODUCTION

Clustering (also known as unsupervised learning) is the task of recognizing a finite group of categories (or clusters) to illustrate the data. Therefore, similar objects are clustered to the similar category and dissimilar objects to different clusters. Clustering is also known as unsupervised learning since the data objects are pointed to a collection of clusters which can be interpreted as classes additionally. Clustering is the process of assembling the data records into significant subclasses (clusters) in a way that increases the relationship within clusters and reduces the similarity among two different clusters. Other names for clustering are unsupervised learning (machine learning) and segmentation. Clustering is used to get an overview over a given data set. A set of clusters is often enough to get insight into the data distribution within a data set. Another important use of clustering algorithms is the preprocessing for some other data mining algorithm.

Fuzzy clustering methods allow the objects to belong to several clusters simultaneously, with different degrees of membership. Fuzzy clustering is a powerful unsupervised method for the analysis of data and construction of models. In many situations, fuzzy clustering is more natural than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The discrete nature of the hard partitioning also causes difficulties with algorithms based on analytic functional, since these functional are not differentiable. The concept of fuzzy partition is essential for cluster analysis, and consequently also for the identification techniques that are based on fuzzy clustering. Fuzzy and possibilistic partitions can be seen as a generalization of hard partition which is formulated in terms of classical subsets.

The remainder of this is organized as follows. Section 2 summarizes the concepts and literature survey. Section 3 discusses the proposed method, and section 4 provides the experiments with high accuracy. Finally, Section 5 presents the conclusions of the work.

II. LITERATURE SURVEY

A.M. Fahim et al., (2006) proposed an enhanced method for assigning data points to the suitable clusters. In the original K-Means algorithm in each iteration the distance is calculated between each data element to all centroids and the required computational time of this algorithm is depends on the number of data elements, number of clusters and number of iterations, so it is computationally expensive.

Likas et al., (2003) put forth a global K-Means clustering algorithm. The technique was an incremental move towards to clustering that dynamically includes one cluster center at a particular time in the course of a deterministic global exploration procedure comprises of N (with N being the size of the data set) executions of the K-Means algorithm from appropriate initial positions. Baolin Yi et al., (2010) proposed a new method to find the initial center and improve the sensitivity to the initial centers of K-Means algorithm. Barakbah et al., (2009) proposes a new approach to optimizing the designation of initial centroids for K-Means clustering. Celikyilmaz et al., (2008) proposed a new fuzzy system modeling approach based on improved fuzzy functions to model systems with continuous output variable.

Chen Zhang et al., (2009) presented a new clustering method based on K-Means that have avoided alternative randomness of initial center. This approach focused on K-Means algorithm to the initial value of the dependence of K selected from the aspects of the algorithm is improved. Chunhui et al., (2008) presented a Similarity based Fuzzy and Possibilistic C-Means algorithm called SFCM. It is derived from original fuzzy and FPCM which was proposed by Bezdek.

Fang Yuan et al., (2004) investigated the standard K-Means clustering algorithm in this work and give our
improved version by selecting better initial centroids that the algorithm begins with.

Filippone et al., (2010) investigated a kernel extension of the classic possibilistic C-Means. In this extension, the author implicitly mapped input patterns into a possibly high-dimensional space by means of positive semidefinite kernels. However, it is not good for the image with noise and it also takes more time for execution. A new modified FPCM clustering algorithm is proposed by Ganesan et al., (2010) for color image segmentation of any type of color images. This new proposed clustering algorithm exhibits the robustness to noise, and also faster as compared to the traditional one.

Jiang-She Zhang et al., (2004) modified and improve these algorithms to overcome their shortcoming. A fast PCM clustering algorithm is proposed by Kai Li et al., (2003). Ojeda-Magafia et al., (2006) proposed a new technique to use the Gustafson-Kessel (GK) algorithm within the FPCM, such that the cluster distributions have a better adaptation with the natural distribution of the data. Xiao-Hong et al., (2005) presented a novel approach on Possibilistic Fuzzy C-Means Clustering Model Using Kernel Methods. The author insisted that fuzzy clustering method is based on kernel methods. This technique is said to be Kernel Possibilistic Fuzzy C-Means model (KPFPCM).

III. PROPOSED METHODOLOGY

This work presents a clustering algorithm called Fuzzy Possibilistic C-Means that merges the characteristics of both Fuzzy and Possibilistic C-Means. To enhance the FPCM approach MFPCM is presented. This novel technique aims to give good results relating to the previous algorithms by modifying the objective function used in FPCM. The objective function is based by adding new weight of data points in relation to every cluster and modifying the exponent of the distance between a point and a class.

A. Fuzzy Possibilistic Clustering Algorithm

Data analysis is considered as a very important science in the real world. Cluster analysis is a technique for classifying data; it is a method for finding clusters of a data set with most similarity in the same cluster and most dissimilarity between different clusters. The conventional clustering methods puiteach point of the data set to exactly one cluster. A fuzzy version of clustering appeared; it is Fuzzy C-Means with a weighting exponent \( m > 1 \), that uses the probabilistic constraint that the memberships of a data point across classes sum to one. The FCM is sensitive to noise. To mitigate such an effect, Krishnapuram and Keller throw away the constraint of memberships in FCM and propose the Possibilistic C-Means (PCM) algorithm. This work deduced that to classify a data point, cluster centroid has to be closest to the data point, and it is the role of membership. Also for estimating the centroids, the typicality is used for alleviating the undesirable effect of outliers. So Pal defines a clustering algorithm called Fuzzy Possibilistic C-Means that combines the characteristics of both fuzzy and possibilistic \( \epsilon \)-means. The proposed algorithm called Modified Fuzzy Possibilistic C-Means (MFPCM) aims to give good results relating to the previous algorithms by modifying the Objective function used in FPCM.

B. 3.2 Modified Fuzzy Possibilistic C-Means Technique (MFPCM)

The selection of suitable objective function is the major factor for the success of the cluster technique and to achieve enhanced clustering. Hence the clustering optimization is based on objective function to be used for clustering. To obtain an appropriate objective function, the following set of necessities is considered:

- The distance between clusters and the data points allocated to them must be reduced
- The distance between clusters must to be reduced
- The desirability between data and clusters is modeled by the objective function. Also it provides a new technique called Modified Suppressed Fuzzy C-Means, which considerably improves the function of FCM because of a prototype-driven learning of parameter \( \alpha \). The learning procedure of \( \alpha \) is dependent on an exponential separation strength between clusters and is updated at every iteration.

I) Penalized and Compensated Constraints based Fuzzy Possibilistic C-Means (PCFPCM)

This work presents Penalized and compensated constraints that are embedded with the previously discussed Modified Fuzzy Possibilistic C-Means algorithm. The objective function of the MFPCM is given in equation (1.1).

\[
J_{MFPCM} = \sum_{i=1}^{c} \sum_{j=1}^{n} \left( u_{ij}^m w_{ij}^m d^m(x_j, v_i) \right) + \sum_{j=1}^{n} t_{ij}^m d^m(x_j, v_i)
\]

(1.1)

In the proposed approach the penalized and compensated terms are added to the objective function of MFPCM to construct the objective function of PCFPCM. The penalized constraint can be represented as follows

\[
\frac{1}{2} \sum_{x=1}^{n} \sum_{i=1}^{c} \left( u_{ij}^m / \alpha_i + t_{ij}^m / \beta_j \right)
\]

(1.2)

Where

\[
\alpha_i = \sum_{j=1}^{n} u_{ij}^m / \sum_{j=1}^{n} \sum_{i=1}^{c} w_{ij}^m, i = 1, 2, ..., c
\]

(1.3)

\[
\beta_j = \sum_{i=1}^{c} t_{ij}^m / \sum_{j=1}^{n} \sum_{i=1}^{c} t_{ij}^m, j = 1, 2, ..., n
\]

(1.4)

Where \( \alpha_i \) is a proportional constant of class \( i \); \( \beta_i \) is a proportional constant of training vector \( z_n \); and \( \nu \) (\( \nu \geq 0 \)); \( \tau \) (\( \tau \geq 0 \)) are also constants. In these functions, \( \alpha_i \) and \( \beta_j \) are defined in equations above. Membership \( u_{ij} \) and typicality \( t_{ij} \) for the penalized component is presented below.

\[
(u_{ij})_{p} = \left( \sum_{i=1}^{c} \left( \frac{\| z_i - \sigma_i \|^2 - \nu \ln \alpha_i^{1/(m-1)} }{ \| z_i - \sigma_i \|^2 - \nu \ln \alpha_i^{1/(m-1)} } \right) \right)^{-1}
\]

(1.5)

\[
(t_{ij})_{p} = \left( \sum_{i=1}^{c} \left( \frac{\| z_i - \sigma_i \|^2 - \nu \ln \beta_j^{1/(\eta-1)} }{ \| z_i - \sigma_i \|^2 - \nu \ln \beta_j^{1/(\eta-1)} } \right) \right)^{-1}
\]

(1.6)

\[
(j = 1, 2, ..., n, i = 1, 2, ..., c)
\]

(1.7)
In the previous expression \( \alpha_j = v_i = \frac{\sum_{j=1}^{c}(u_{ij}^0 + t_{ij}^0)^2}{\sum_{j=1}^{c}(u_{ij}^0 + t_{ij}^0)} \), \( 1 \leq i \leq c \), which is the centroid. The compensated constraints can be represented as follows:

\[
\frac{1}{2} \sum_{x=1}^{n} \sum_{i=1}^{c} (u_{ij}^0 \tan \alpha_i + t_{ij}^0 \tan \beta_j)
\]

(1.5)

Where membership \( u_{ij} \) and typicality \( t_{ij} \) for the compensation is presented below:

\[
(u_{ij})^c = \left( \sum_{j=1}^{c} \left( \frac{||z_i - \alpha_j||^2 - \tau \tanh(\alpha_i)}{||z_i - \alpha_j||^2 - \tau \tanh(\alpha_i)} \right)^{1/(m-1)} \right)^{-1}
\]

(1.6)

\[
(t_{ij})^c = \left( \sum_{j=1}^{c} \left( \frac{||z_i - \alpha_j||^2 - \tau \tanh(\beta_j)}{||z_i - \alpha_j||^2 - \tau \tanh(\beta_j)} \right)^{1/(\eta-1)} \right)^{-1}
\]

(1.7)

To obtain an efficient clustering the penalization term must be removed (i.e. the noise is removed) and the compensation term must be added to the basic objective function of the existing MFPCM. This brings out the objective function of PCFPCM and it is given in (5.21):

\[
J_{PCFPCM} = \sum_{i=1}^{c} \sum_{j=1}^{n} \left( u_{ij}^0 w_i^m d_{m}^2(x_i, v_j) \right) + t_{ij}^0 w_i^m d_{m}^2(x_i, v_j) + \frac{1}{2} \sum_{x=1}^{n} \sum_{i=1}^{c} (u_{ij}^0 \ln \alpha_i + t_{ij}^0 \ln \beta_j) \]

(1.7)

The final objective function is presented in (1.7).

2) 3.2.1 Algorithm of PCFPCM

Step 1: FCM algorithm is an iterative clustering method that brings out an optimal c partition by minimizing the weighted within group sum of squared error objective function.

Step 2: X is the dataset in the p-dimensional vector space, the number of data items is represented as \( p, c \) is the number of clusters with \( 2 \leq c \leq n - 1 \).

Step 3: \( V \) is the c centers or prototypes of the clusters, \( v_i \) represents the p-dimension center of the cluster i, and \( d^2(x_i, v_j) \) represents a distance measure between object \( x_i \) and cluster centre \( v_j \).

Step 4: \( U = \{ u_{ij} \} \) represents a fuzzy partition matrix with \( u_{ij} = u_i(x_j) \) is the degree of membership of \( x_j \) in the i\(^{th} \) cluster; \( x_j \) is the j\(^{th} \) p-dimensional measured data.

Step 5: To recover this weakness of FCM, relaxed the constrained condition of the fuzzy c-partition to obtain a possibilistic type of membership function.

Step 6: The characteristics of both Fuzzy and Possibilistic C-Means are combined in the objective function of the FPCM. But the estimation of centroids is influenced by the noise data.

Step 7: Exponential separation strength between clusters is the base for the learning process of \( \alpha \) and is updated at each of the iteration.

Step 8: A new parameter is added with this which suppresses this common value of \( \alpha \) and replaces it by a new parameter like a weight to each vector. Add Weighting exponent as exhibitor of distance in the objective functions.

Step 9: Penalization term must be removed (i.e. the noise is removed) and the compensation term must be added to the basic objective function of the existing MFPCM.

IV. EXPERIMENTAL RESULTS

To evaluate the proposed penalized and compensated constraints based Fuzzy Possibilistic C-Means (PCFPCM) against Fuzzy Possibilistic C-Means (FPCM) and Modified Fuzzy Possibilistic C-Means (MFPCM), experiments were carried out using the similar experimental setup and parameters as discussed in this chapter. The experiment is done in MATLAB with such datasets as Iris, Wine, Lung Cancer and Lymphograma.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>FPCM</th>
<th>MFPCM</th>
<th>PCFPCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>68</td>
<td>8.00</td>
<td>0.49</td>
</tr>
<tr>
<td>Wine</td>
<td>73</td>
<td>5.50</td>
<td>0.48</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>75</td>
<td>4.80</td>
<td>0.45</td>
</tr>
<tr>
<td>Lymphograma</td>
<td>80</td>
<td>3.70</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Table 1: Accuracy, Execution Time and Mean Squared Error
Thus the proposed method has less error rate when compared with others.

V. CONCLUSION

This work provides a penalized constraint of FCM is improved by using NEM algorithm and it is combined with compensated constraints which is said to be Improved Penalized and Compensated constraints for Fuzzy Possibilistic C-Means (IPCFPCM) clustering algorithm. The usage of improved penalized constraints in MFPCM will help in better calculation of distance between the clusters and increasing the accuracy of clustering. Thus the proposed method may perform better performance with high accuracy.

REFERENCES


