High Dimensional Data Using Fuzzy C – Means Clustering For Sensitive Distance Metric

G.Shanthini
Kovai Kalaimagal College of Arts and Science, Coimbatore – 109, India.

P.Ponsekar
Assistant Professor, Department of Computer Science,
Kovai Kalaimagal College of Arts and Science, Coimbatore – 109, India.

S.Vidhya
Assistant Professor, Department of Computer Science,
Kovai Kalaimagal College of Arts and Science, Coimbatore – 109, India.

Abstract – The major objective of clustering is to discover collection of comparable objects based on similarity metric. On the other hand, a similarity metric is generally specified by the user according to the requirements for obtaining better results.There are several approaches available for clustering objects. In the proposed approach an effective fuzzy clustering technique is used. 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. Penalized and Compensated terms are embedded with the Modified fuzzy positivistic clustering method's objective function to construct the Penalized based FPCM (PFPCM). In order to improve the clustering accuracy, third proposed approach uses the Improved Penalized Fuzzy C-Means (IPFCM). The penalty term takes the spatial dependence of the objects into consideration, which is inspired by theNeighborhood Expectation Maximization (NEM) algorithm and is modified according to the criterion of FCM. In this approach, penalized constraint is improved by using NEM algorithm and it is combined with compensated constraints. The proposed Improved Penalized for Fuzzy C-Means (IPFCM) clustering algorithm, uses improved penalized constraints which will help in better calculation of distance between the clusters and increasing the accuracy of clustering.

Index Terms – Cluster, Classifier, Distance Metrics, Data set, Partition.

1. INTRODUCTION

FCM algorithm is a distinctive clustering algorithm, which has been exploited in extensive range of engineering and scientific disciplines, for instance, medicine imaging, pattern detection, data mining and bioinformatics. In view of the fact, the initially developed FCM makes use of the squared-norm to determine the similarity between prototypes and data points, and it performs well only in the case of clustering spherical clusters. Furthermore, several algorithms are developed by numerous authors based on the FCM with the aim of clustering more general dataset. The FCM is responsive to noise.

In order to classify a data point, Pal deducted a technique that the data point must have their cluster centroid nearer, and it is the major role of membership. Also for the centroid estimation, the typicality is used for alleviating the unwanted effect of outliers. To enhance the PFCM approach MPFCM has been presented. This novel technique aims to give good results relating to the previous algorithms by modifying the Objective function used in PFCM.

The existing approach uses the probabilistic constraint to enable the memberships of a training sample across clusters that sum up to 1, which means the different grades of a training sample are shared by distinct clusters, but not as degrees of typicality. In contrast, each component generated by the Penalized Fuzzy C-Means (PFCM) corresponds to a dense region in the dataset. Each cluster is independent of the other clusters in the PFCM strategy.

If a training sample has been classified to a suitable cluster, then membership is a better constraint for which the training sample is closest to this cluster. In other words, typicality is an important factor for unburdening the undesirable effects of outliers to compute the cluster centers. To enhance the above mentioned existing approach in MPFCM, penalized and compensated constraints are incorporated. Hence, Yang have added the penalized term into fuzzy C-Means to construct the Penalized Fuzzy C-Means (PFCM) algorithm.

2. AN EFFICIENT FUZZY C-MEANS CLUSTERING FOR SENSITIVE DISTANCE METRIC

2.1. Fuzzy Clustering Algorithm

The fuzzified version of the K-Means algorithm is the Fuzzy C-Means (FCM). It is a method of clustering which allows one piece of data to belong to two or more clusters. This method

was developed by Dunn in 1973 this is frequently used in pattern recognition. The 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 J_{FCM}:

$$J_{FCM}(V,U,X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\, d^2(x_j,v_i), 1 < m < +\infty$$

In the equation $X = \{x_1, x_2, \ldots, x_n\} \subseteq R^p$ is the dataset in the p-dimensional vector space, where the number of data items is represented as p, c is the number of clusters with $2 \le c \le n-1$. $V = \{v_1, v_2, \ldots, v_c\}$ is the c centers or prototypes of the clusters, v_i represents the p-dimension center of the cluster$i$, and $d^2(x_j, v_i)$ represents a distance measure between object x_j and cluster center v_i. $U = \{u_{ij}\}$ 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] of p-dimensional measured data. The fuzzy partition matrix satisfies:

$$0 < \sum_{j=1}^{n} u_{ij} < n, \forall i \in \{1, \ldots, c\} \quad (2)$$

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1, \ldots, n\} \quad (3)$$

(2)

Where m is a weighting exponent parameter on each fuzzy membership and establishes the amount of fuzziness of the resulting classification; it is a fixed number greater than one. Under the constraint of U the objective function J_{FCM} can be minimized. Specifically, taking of J_{FCM} with respect to $u_{ij}$and v_i and zeroing them respectively is necessary but not sufficient conditions for J_{FCM} to be at its local extreme will be as the following:

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d^2(x_j,v_i)}{d^2(x_j,v_k)}\right)^{2/(m-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \le n. \quad (4)$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} x_k}{\sum_{k=1}^{n} u_{ik}^{m}}, 1 \le i \le c. \quad (5)$$

The characteristics of both Fuzzy and Possibility C-Means are combined. Memberships and typicality's are important for the correct feature of data substructure in clustering problem. Thus, an objective function in the PFCM depending on both memberships and typicality's can be represented as below:

$$J_{FPCM}(U,T,V) = \sum_{i=1}^{c}\sum_{j=1}^{n} (u_{ij}^{m} + t_{ij}^{\eta})\, d^2(x_j,v_i) \quad (4)$$

With the following constraints:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1, \ldots, n\}$$

$$\sum_{j=1}^{n} t_{ij} = 1, \forall i \in \{1, \ldots, c\} \quad (10)$$

(5)

A solution of the objective function can be obtained through an iterative process where the degrees of membership, typicality and the cluster centers are updated with the equations as follows.

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d^2(x_j,v_i)}{d^2(x_j,v_k)}\right)^{2/(m-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \le n.$$

$$t_{ij} = \left[\sum_{k=1}^{n}\left(\frac{d^2(x_j,v_i)}{d^2(x_j,v_k)}\right)^{2/(\eta-1)}\right]^{-1}, 1 \le i \le c, 1 \le j \le n.$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik}^{m} + t_{ik}^{\eta}) X_K}{\sum_{k=1}^{n}(u_{ik}^{m} + t_{ik}^{\eta})}, 1 \le i \le c. \quad (6)$$

PFCM constructs memberships and possibilities simultaneously, along with the usual point prototypes or cluster centers for each cluster. Hybridization of PCM and FCM is the PFCM that often avoids various problems of PCM and FCM and PFCM. PFCM solves the noise sensitivity defect of FCM, overcomes the coincident clusters problem of PCM. But the estimation of centroids is influenced by the noise data. Hence, Modified PFCM is built to overcome this difficulty.

2.2. Modified Penalized Fuzzy C-Means Technique (MPFCM)

A new algorithm was given by Wen-Liang Hung called Modified Suppressed Fuzzy C-Means (MS-FCM), which significantly improves the performance of FCM due to a prototype-driven learning parameter α [84]. Exponential separation strength between clusters is the base for the learning process of α and is updated at each of the iteration. The parameter α can be computed as

$$\alpha = exp\left[-\min_{i \neq k} \frac{||v_i - v_k||^2}{\beta}\right] \quad (13)$$

(6)

In the above equation β is a normalized term so that β is chosen as a sample variance. That is, β is defined:

$$\beta = \frac{\sum_{j=1}^{n} ||x_j - \bar{x}||^2}{n}$$

Where $\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n}$ (7)

But the remark which must be mentioned here is the common value used for this parameter by all the data at each iteration, which may lead to error. A new parameter is added with this which suppresses this common value of α and replaces it by a new parameter like a weight to each vector. Or every point of the dataset has a weight in relation to every cluster. Consequently this weight permits to have a better classification especially in the case of noise data. The following equation is used to calculate the weight.

$$w_{ji} = exp\left[-\frac{\|x_j - v_i\|^2}{\left[\sum_{j=1}^{n}\|x_j - \bar{v}\|^2\right] * c/n}\right] \quad (14)$$

(8)

In (14) w_{ji} represents weight of the point j in relation to the class$i$. This weight is used to modify the fuzzy and typical partition. The objective function is composed of two expressions: the first is the fuzzy function and uses a fuzziness weighting exponent, the second is possibilistic function and uses a typical weighting exponent; but the two coefficients in the objective function are only used as exhibitor of membership and typicality.

A new relation enables a more rapid decrease in the function and increase in the membership and the typicality when they tend toward 1 and decrease this degree when they tend toward 0. This relation is to add Weighting exponent as exhibitor of distance in the two objective functions. The objective function of the MPFCM can be formulated as follows:

$$J_{MPFCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij}^{m} w_{ij}^{m} d^{2m}(x_j, v_i) + t_{ij}^{\eta} w_{ij}^{\eta} d^{2\eta}(x_j, v_i))$$

$U = \{u_{ij}\}$ Represents a fuzzy partition matrix, defined as:

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)}\right)^{2m/(m-1)}\right]^{-1} \quad (9)$$

$T = \{t_{ij}\}$ Represents a typical partition matrix, defined as:

$$t_{ij} = \left[\sum_{k=1}^{n}\left(\frac{d^2(x_j, v_i)}{d^2(x_j, v_k)}\right)^{2\eta/(\eta-1)}\right]^{-1} \quad (17)$$

$V = \{v_i\}$ Represents c centers of the clusters, defined as:

$$v_i = \frac{\sum_{j=1}^{n}(u_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta}) * x_j}{\sum_{j=1}^{n}(u_{ik}^{m} w_{ji}^{m} + t_{ik}^{\eta} w_{ij}^{\eta})} \quad (10)$$

Where d^2describesthe Density-sensitive distance metric.

Let data points be the nodes of graph G = (V,E), and p $\in V^l$ be a path of length l = $|p|$ connecting the nodes p_1 and $p^{|p|}$, in which $(p_k, p_{k+1}) \in E$, $1 \leq k < |p|$, Let P_{ij} denote the set of all paths connecting nodes x_i and x_j . The density- sensitive distance metric between two points is defined to be

$|p|=1$

$$\boldsymbol{D_{ij=min_{p\in P_{ij}}} \sum_{k=1}^{|P|-1} L(P_k,P_k+1)}$$

Thus *Dij* satisfies the four conditions for a metric, i.e. $D_{ij} = D_{ji}, D_{ij} \geq 0; D_{ij} \leq D_{ik} + D_{kj}$ for all x_i, x_j, x_k; and $D_{ij} = 0\ iff\ x_i = x_j$.As a result,the density-sensitive distance metric can measure the geodesic distance along the manifold, which results in any two points in the same region of high density being connected by a lot of shorter edges while any two points in different regions of high density are connected by a longer edge through a region of low density.

3. PERFORMANCE COMPARISON

Lung Cancer Dataset

Clustering Accuracy

Accuracy of the clustering results is calculated for FCM, PFCM and the proposed MPFCM for lung cancer dataset. Table 6.9 shows comparison of the accuracy of clustering results for the proposed method with the FCM method and PFCM for lung cancer dataset.

Clustering Technique	Clustering Accuracy (%)
FCM	94.8
PFCM	95.9
MPFCM	98.4

Table 1 Comparison of Clustering Accuracy For Lung Cancer Dataset

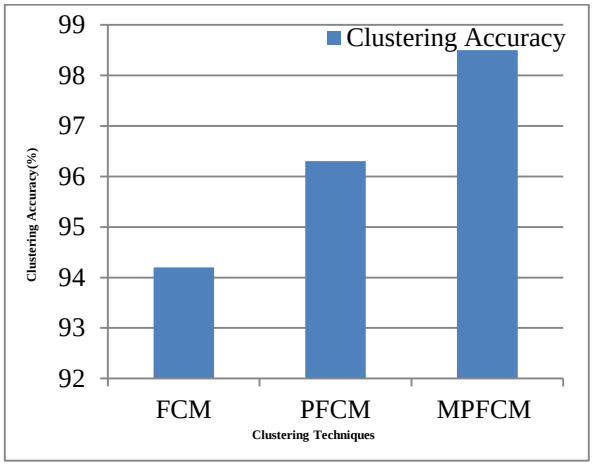

Figure 1: Comparison of Clustering Accuracy for Lung Cancer Dataset

It can be observed from the figure 1 that the accuracy of clustering result using FCM and PFCM method are 94.2% and 96.3% respectively, and that of the proposed MPFCM is 98.5% for lung cancer dataset.

Lymphography Dataset

Clustering Accuracy

Clustering Technique	Clustering Accuracy (%)
FCM	94.2
PFCM	96.3
MPFCM	98.5

Table 2 Comparison of Clustering Accuracy in Lymphography Dataset

Accuracy of the clustering results is calculated for FCM, PFCM and the proposed MPFCM in lymphography dataset.

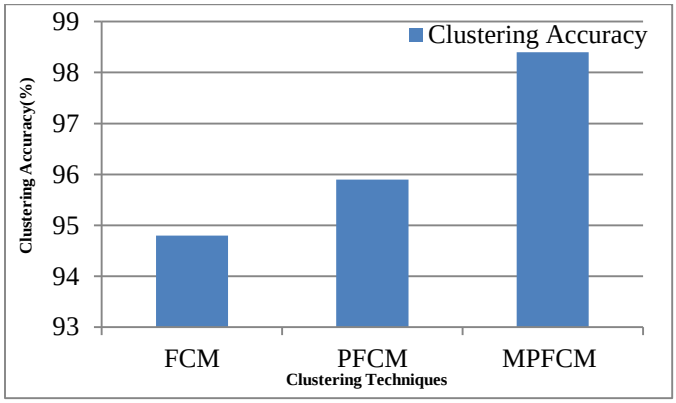

Figure 2 Comparison of Clustering Accuracy for Lymphography Dataset

It can be observed from the figure 2 that the accuracy of clustering result using FCM and PFCM method are 94.8% and 95.9% respectively, and that of the proposed MPFCM is 98.4% for lymphography dataset.

This research focuses on the effective clustering techniques for data clustering. Effective fuzzy clustering approaches are used in this research which improves the results of clustering.

In the approach, penalized FCM is improved by using NEM algorithm and it is combined with compensated constraints which is said to be Improved Penalized constraints for Fuzzy Possibilistic C-Means (IPFPCM) clustering algorithm. The usage of improved penalized constraints in MPFCM will help in better calculation of distance between the clusters and increasing the accuracy of clustering.

The performances of the proposed approaches are evaluated on UCI machine repository datasets namely Iris, Wine, Lung cancer and Lymphography. It is observed from the experimental results that the proposed MPFCM method outperforms the other proposed approaches in terms of accuracy, Mean Squared Error, Execution Time and Convergence Behavior. Thus, the proposed MPFCM approach is best suited for the data clustering applications.

The present research work can be applied to various specific applications in the field of data mining.

Clustering plays an outstanding role in data mining applications such as Scientific Data Exploration, Information Retrieval and Text Mining, Spatial Database Applications, Web Analysis, Marketing, Medical Diagnostics especially

Gene Classification, Computational Biology, Customer Relationship Management (CRM), etc.

REFERENCES

[1] A.K. Jain and R.C. Dubes, "Algorithms for Clustering Data" ,Prentice Hall, 1988.
[2] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, 1999.
[3] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of Royal Statistical Society, series B, Vol. 1, No. 39, Pp. 1–31, 1977.
[4] Bezdek J.C and Pal S.K, "Fuzzy Models for Pattern Recognition", IEEE Press, New York, 1992.
[5] D. Fisher, "Knowledge acquisition via incremental conceptual clustering" Machine Learning, Vol. 2, Pp. 139-172, 1987.
[6] Dong-Chul Park, "Intuitive Fuzzy C-Means Algorithm for MRI Segmentation", International Conference on Information Science and Applications (ICISA), Pp. 1–7, 2010.
[7] E.P. Xing, A.Y. Ng, M.I. Jordan and S. Russell, "Distance metric learning, with application to clustering with side-information", NIPS 15, Pp. 505–512, 2003.
[8] Filippone M, Masulli F and Rovetta S, "Applying the Possibilistic c-Means Algorithm in Kernel-Induced Spaces", IEEE Transactions on Fuzzy Systems, Vol. 18, No. 3, Pp. 572–584, 2010.
[9] Frank A and Asuncion A, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine", CA: University of California, School of Information and Computer Science, 2010.
[10] G. Karypis, E.-H Han and V. Kumar. "Chameleon: Hierarchical Clustering using Dynamic Modelling", IEEE Computer, Pp.