ISSN 2250-3153

Fuzzy C- Means Algorithm- A Review

R.Suganya, R.Shanthi

Department of CS, Dr.SNS.Rajalakshmi College of Arts & Science

Abstract- Clustering is a task of assigning a set of objects into groups called clusters. In general the clustering algorithms can be classified into two categories. One is hard clustering; another one is soft (fuzzy) clustering. Hard clustering, the data's are divided into distinct clusters, where each data element belongs to exactly one cluster. In soft clustering, data elements belong to more than one cluster, and associated with each element is a set of membership levels. In this paper we represent a survey on fuzzy c means clustering algorithm. These algorithms have recently been shown to produce good results in a wide variety of real world applications.

Index Terms- Soft clustering, hard clustering, FCM.

I. INTRODUCTION

Fast and robust clustering algorithms play an important role in extracting useful information in large databases. The aim of cluster analysis is to partition a set of N object into C clusters such that objects within cluster should be similar to each other and objects in different clusters are should be dissimilar with each other[1]. 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.

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a dataset, where the objects inside each cluster show a certain degree of similarity. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Cluster analysis is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization. It will often necessary to modify preprocessing and parameter until the result achieves the desired properties.

In Clustering, one of the most widely used algorithms is fuzzy clustering algorithms. Fuzzy set theory was first proposed by Zadeh in 1965 & it gave an idea of uncertainty of belonging which was described by a membership function. The use of fuzzy set provides imprecise class membership function. Applications of fuzzy set theory in cluster analysis were early proposed in the work of Bellman, Zadeh, and Ruspini This paper opens door step of fuzzy clustering [2]. Integration of fuzzy logic with data mining techniques has become one of the key constituents of soft computing in handling challenges posed by massive collections of natural data. The central idea in fuzzy clustering is the non-unique partitioning of the data into a collection of clusters. The data points are assigned membership values for each of the clusters and fuzzy clustering algorithm allow the clusters to grow into their natural shapes [3]. The fuzzy clustering algorithms can be divided into two types 1) Classical fuzzy clustering algorithms 2) Shape based fuzzy clustering algorithms. Classical fuzzy clustering algorithms can be divided into three types.1) The Fuzzy C-Means algorithm 2) The Gustafson-Kessel algorithm 3) The Gath-Geva algorithm. Shape based fuzzy clustering algorithm can be divided into 1) Circular shape based clustering algorithm 2) Elliptical shape based clustering algorithm 3) Generic shape based clustering algorithm. In this paper, represent a review on fuzzy c means, and extended version of fcm such as pcm, fpcm and their advantages and disadvantages of real time applications.

II. FUZZY C MEANS ALGORITHM

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. Fuzzy c-means algorithm is most widely used. Fuzzy c-means clustering was first reported in the literature for a special case (m=2) by Joe Dunn in 1974. The general case (for any m greater than 1) was developed by Jim Bezdek in his PhD thesis at Cornell University in 1973. It can be improved by Bezdek in 1981. The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

Algorithm

1. Initialize U=[u_{ij}] matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}}$$

3. Update$U^{(k)}$,$U^{(k+1)}$
4. $$d_{ij} = \sqrt{\sum_{i=1}^{n} (x_{i-} c_i)}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$$

5. $if \parallel U(k+1) - U(k) \parallel < \epsilon$ $then$ STOP; otherwise return to step 2.

Here m is any real number greater than 1,
u_{ij} is the degree of membership of x_i in the cluster j,
x_i is the ith of d-dimensional measured data,
c_j is the d-dimension center of the cluster,

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula.

Advantages
1) Unsupervised
2) Converges

Limitations:
1) Long computational time
2) Sensitivity to the initial guess (speed, local minima)
3) Sensitivity to noise and One expects low (or even no) membership degree for outliers (noisy points).

III. Possibilistic C-Means (PCM)

To overcome difficulties of the fcm, Krishnapuram and keller proposed a new clustering model named Possibilistic c-Means (PCM).

Algorithm

Fix the number of clusters C; fix m,1<m<∞;\
Set iteration counter l=1;
Intialize the possiblistic C-parttion $U^{(0)}$;
Estimate η_i
Repeat
Update the prototypes using $U^{(l)}$, as indicated bel ow;
Compute $U^{(l+1)}$
Increment l;
Until ($\| U^{(l-1)}-U^{(l)}\|<\varepsilon$);

{ The remaining part of algorithm is optional and to be used only when the actul shape of the generated possibility distribution is important }

Set iteration counter l=1;
Reestimate η_i
Repeat protypes using $U^{(l)}$, as indicated below;
Compute $U^{(l+1)}$
Increment l;
Until ($\| U^{(l-1)}-U^{(l)}\|<\varepsilon$);

η_i –determines distance at which the membership value of a point in a cluster becomes 0.5.

$$\eta_i = K\frac{\sum_{j=1}^{N} u_{ij}^{m} d_{ij}^{2}}{\sum_{j=1}^{N} u_{ij}^{m}}$$

Advantage
Clustering noisy data samples

Disadvantages
1) Very sensitive to good initialization
2) Coincident clusters may result
Because the columns and rows of the typicality matrix are independent of each other
Sometimes this could be advantageous (start with a large value of c and get less distinct clusters)

IV. Fuzzy Possibilistic C Means Algorithm (FPCM)

To overcome difficulties of the pcm, Pal defines a clustering technique that integrates the features of both Fuzzy a Possibilistic c-means called Fuzzy Possibilistic c-Means (FPCM). Membership and Typicality's are very significant for the accurate characteristic of data substructure in clustering difficulty. An objective function in the fpcm depending on both membership and typicality's are represented as::

Memberships and topicalities is represented as:

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

Which of the following constraints

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

$$\sum_{i=1}^{c} t_{ij} = 1, \forall i \in \{1, \ldots, c\}$$

FPCM generates Memberships and possibilities at the same time, together with the usual point prototypes or cluster center for each cluster.

Advantage
1) Ignores the noise sensitivity deficiency of FCM
2) Overcomes the coincident clusters problem of PCM.

Disadvantages
1) The row sum constraints must be equal to one

V. Possibilistic Fuzzy C Means Algorithm (PFCM)

In fpcm, the constraint corresponding to the sum of all typicality values of all data to a cluster must be equal to one cause problems particularly for a big data set. In order to avoid this problem pal et al propose a new algorithm called Possibilistic Fuzzy c means algorithm (pfcm). The objective function is defined by

$$J_{PFCM}(U,T,V;Z) = \sum_{i=1}^{c}\sum_{j=1}^{n}(au_{ik}^{m} + bt_{ik}^{\eta}) \times \|z_k - v_i\|^{2} + \sum_{i=1}^{c}\delta_i \sum_{k=1}^{N}(1 - t_{ik})^{\eta}$$

Subject to
$\sum_{i=1}^{c}\mu_{ik} = 1\ \forall\, k, 0 < \mu_{ik}, t_{ik} < 1, a > 0, b < 0, m > 1, n > 1$
, a&b define the relative importance between the membership degrees and typicality values. The objective function J_{MN} can be

minimized if $D_{ikA} = ||z_k - v_i||_A > 0$, for every i and k,m,$\eta$>1 as well as z contains a minimum of c different data with these conditions we have (U,T^T,V)$^{\varepsilon M_{fc} \times M_{pc} \times \Re^p}$.The membership degree calculated with

$$J_{FC}(Z,U,V) = \sum_{i=1}^{c}\sum_{j=1}^{m}(u_{ij})^m ||z_k - v_i||^2$$

t_{ik}=1/1+($b(||z_k - v_i|)^2/\delta_i)^{(1/(\eta-1)}$, 1<i<c

$$v_i = (\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^\eta)Zk)/(\sum_{k=1}^{N}(au_{ik}^m + bt_{ik}^\eta)), \quad 1 < i < c$$

Advantage:

1) Ignores the noise sensitivity deficiency of FCM
2) Overcomes the coincident clusters problem of PCM.
3) Eliminates the row sum constraints of FPCM

VI. CONCLUSION AND FUTURE WORK

FCM algorithm is a distinctive clustering algorithm, 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. During the survey, we also find some points that can be further improvement in the future using advanced clustering technique to achieve more efficient accuracy in the result and reduce the time taken for data and/or information retrieval from large dataset.

REFERENCES

[1] M.S.Yang," A Survey of fuzzy clustering" Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16, 1993.

[2] A. vathy-Fogarassy, B.Feil, J.Abonyi"Minimal Spanning Tree based Fuzzy clustering" Proceedings of World academy of Sc., Eng & Technology, vol-8, Oct-2005, 7-12.

[3] Pal N.R, Pal K, Keller J.M. and Bezdek J.C, "A Possibilistic Fuzzy c-Means Clustering Algorithm", IEEE Transactions on Fuzzy Systems, Vol. 13, No. 4, Pp. 517–530, 2005.

[4] R. Krishnapuram amd J.M. Keller, "A possibilistic approach to clustering", IEEE Trans. Fuzzy Systems, Vol. 1, Pp. 98-110, 1993.

[5] Vuda Sreenivasarao and Dr.S. Vidyavathi, "Comparative Analysis of Fuzzy C- Mean and Modified Fuzzy Possibilistic C -Mean Algorithms in Data Mining", IJCST Vol. 1, No. 1, Pp. 104-106, 2010.

[6] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57

[7] J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York.

[8] Mohamed Fadhel Saad and Adel M. Alimi," Modified Fuzzy Possibilistic C-means," Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 -20, 2009, Hong Kong.

AUTHORS

First Author – R.Suganya M.Sc., M.Phil., Assistant Professor, Dr.SNS.Rajalakshmi College of Arts & Science, Chinnavedampatti, Coimbatore., Email-id: mail2cinna.rs@gmail.com

Second Author – R.Shanthi, Research Scholar, Dr.SNS.Rajalakshmi College of Arts & Science, Chinnavedampatti, Coimbatore. , Email-id: shanthisangi@gmail.com