Comparative Analysis of Fuzzy C- Mean and Modified Fuzzy Possibilistic C -Mean Algorithms in Data Mining

Vuda. Sreenivasarao[1], Dr. S. Vidyavathi[2]
[1,2] CSIT Department, JNT University, Hyderabad. India.
E-mail: vudasrinivasarao@gmail.com, vidyasom@yahoo.co.in

Abstract : *Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Clustering is a primary data description method in data mining which group's most similar data. The data clustering is an important problem in a wide variety of fields. Including data mining, pattern recognition, and bioinformatics. It aims to organize a collection of data items into clusters, such that items within a cluster are more similar to each other than they are items in the other clusters. There are various algorithms used to solve this problem In this paper, we use FCM (Fuzzy C mean) clustering algorithm and MFPCM (Modified Fuzzy Possibililstic C mean) clustering algorithm. In this paper we compare the performance analysis of Fuzzy C mean (FCM) clustering algorithm and compare it with Modified Fuzzy possibililstic C mean algorithm. In this we compared FCM and MFPCM algorithm on different data sets. We measure complexity of FCM and MFPCM at different data sets. FCM clustering is a clustering technique which is separated from Modified Fuzzy Possibililstic C mean that employs Possibililstic partitioning. The FCM employs fuzzy portioning such that a point can belong to all groups with different membership grades between 0 and 1.*

Keywords : *Data clustering Algorithm, Portioning, Data Mining, Fuzzy C Mean, Modified Fuzzy Possibililstic C mean.*

I. Introduction

Data analysis is considered as a very important science in the real world. Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Data mining is a computational intelligence discipline that contributes tools for data analysis, discovery of new knowledge, and autonomous decision making. The task of processing large volume of data has accelerated the interest in this field. As mentioned in Mosley (2005) data mining is the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Data mining discovers description through clustering visualization, association, sequential analysis. Clustering is a primary data description method in data mining which group's most similar data. Data clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. 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 put each point of the data set to exactly one cluster. Since 1965, Zadeh proposed fuzzy sets in order to come closer of the physical world. Zadeh introduced the idea of partial memberships described by membership functions. Clustering algorithm partitions an unlabelled set of data into groups according to the similarity. Compared with the data classification, the data clustering is an unsupervised learning process, it does not need a labeled data set as training data, but the performance of the data clustering algorithm is often much poorer. Although the data classification has better performance, it needs a labeled data set as training data and labeled data for the classification is often very difficult and expensive to obtain. So there are many algorithms are proposed to improve the clustering performance.

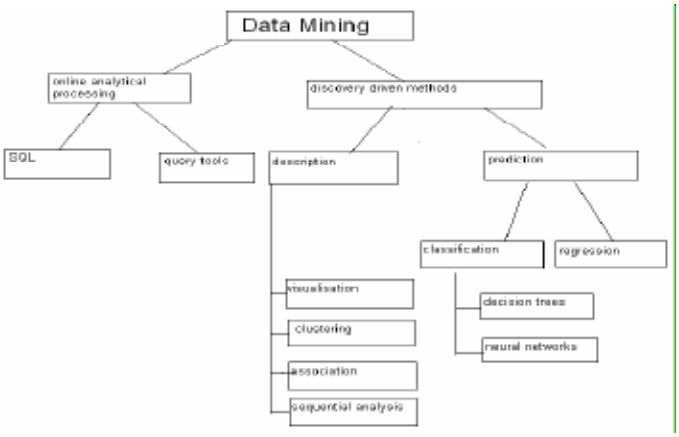

Fig 1.Data Mining Technique.

So it becomes important to have an overview of the concept of clustering. As shown in the fig. 1.1 clustering is one of the techniques of data mining. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait.
Clustering technique is used for combining observed objects into clusters (groups), which satisfy two main criteria:
Each group or cluster should be homogeneous objects that belong to the same group are similar to each other.
Each group of cluster should be different from other clusters, that is, objects that belong to one cluster should be different from the objects of other clusters.
Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects, which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset.

A. System architecture

The overview of the architecture of the system can be seen in Fig. 2. The proposed architecture will adopt the traditional architecture of a data mining system. Data from multiple channels is collected on the operational data store for fast transaction and up to date data that can be used for the

front office. Then, periodically, the data is extracted, cleans, transformed and imported into the data warehouse. The data will then will be send to the appropriate data marts for departmental use. Then, according to the needs of the user, either the enterprise data or the departmental data is sent to the OLAP tier for processing. The results is then stored and then sent to the decision makers through the use of thin clients. The overview of this architecture is seen in Fig. 2 . The proposed system is pretty good in theory as it provides compartmentalization of data and collection of data from multiple channels. The architecture is simple and sticks to the basis of founded work and should provide a good base for the system.

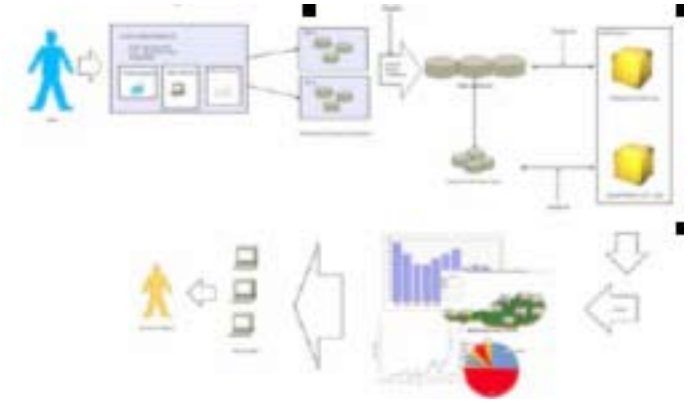

Fig 2: the over view of proposed system Architecture

II. Fuzzy C- Mean Algorithm

Fuzzy C Mean (FCM) is a data clustering technique in which a data set is grouped into n clusters with every data point in the dataset belonging to every cluster will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster.
The steps of FCM algorithm given below. :
Fix c and c is (2<=c<n) and select a value for parameter m'. Initialize the partition matrix U(0). Each step in this algorithm will be labeled as r, where r =0, 1, 2...

1. Calculate the c center vector {vij} for each step

$$Vij = \frac{\sum_{k=1}^{n} u_{ik}^{m'} \times x_{jk}}{\sum_{k=1}^{n} u_{ik}^{m'}}$$

2. Calculate the distance matrix D[c,n].

$$D\ ij = \left[\sum_{j=1}^{m} \left[x_{jk} - v_j \right]^2 \right]^{[1/2]}$$

3. Update the partition matrix for the rth step, U® as follow:

$$u_{ik}^{r-1} = 1$$

$$\sum_{j=1}^{c} \left[\frac{d_{ik}^{r}}{d_{jk}^{r}} \right]^{2/[m'-1]}$$

if $||U^{(k+1)}-U^{(k)}||<\delta$ then STOP: otherwise return to step 2 by iteratively updating the cluster centers and the membership grades for data point. FCM iteratively moves the cluster centers to the "right" location with in a dataset.

III. Modified Fuzzy Possibilistic C - Mean Algorithm

The FPCM algorithm attempts to partition a finite collection of elements X={x1, x2, x3.....xn} into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centers V, such that V=Vi,i=1,2,3.....,c And a partition matrix U such that U=Uij, i=1,2,3,.....c, j=1,2,.....n Where uij is a numerical value in [0, 1] that tells the degree to which the elements Xj belongs to the i-th cluster. Defines a family of fuzzy sets {Ai, i=1,2,3.....c} as a fuzzy c partition on a universe of data points X
1. Fuzzy set allows for degree of membership
2. A single point can have partial membership in more than one class.
3. There can be no empty classes and no class that contains no data points.

The steps of MFPCM algorithm given below:

1. The objective function of the MFPCM can be formulated as follows:

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

2. Calculate U = {μ ij} represents a fuzzy partition matrix, is defined as:

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

3. Calculate T = {t ij} represents a typical partition matrix, is defined as :

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

4. Calculate V = {v ij} represents c centers of the clusters, is defined as:

$$v_i = \frac{\sum_{j=1}^{n} \left(\mu_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta} \right) * x_j}{\sum_{j=1}^{n} \left(\mu_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta} \right)}$$

IV. Results

After implementation and study we get the following results.

A. Complexity Analysis of FCM Algorithm

The asymptotic efficiency of the algorithm has following

notations:

i number FCM over entire dataset.
n number of data points.
c number of clusters
d number of dimensions

The time complexity of the fuzzy c mean algorithm is O(ndc2i), where empirically I grows very slowly with n,c and d.

The memory complexity of FCM is O (nd + nc), where nf is the size of data set and nc the size of U matrix.
For data sets, which cannot be loaded into memory, FCM will have disk accesses every iteration. Thus the disk input output complexity will be O(ndi) It is likely that for those data sets the U matrix cannot be kept in memory too. Thus, it will increase the disk input/output complexity further

B. Complexity Analysis of MFPCM Algorithm

The asymptotic efficiency of the algorithm has following notations:

i number of k means passes over entire dataset.
n number of data points.
c number of clusters
d number of dimensions

The time complexity of the hard c mean algorithm is O (ncdi), where empirically I grows very slowly with n, c and d.
The memory complexity of MFPCM is cd
I/O complexity of MFPCM is ndi

C. Comparative Analysis of Complexities of FCM and MFPCM:

Algorithm	Time complexity	Space complexity	I/O complexity
FCM	$O(ndc^2i)$,	O(nd + nc)	O(ndi)
MFPCM	O(ncdi),	cd	ndi

V. Conclusion

Fuzzy clustering, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. They have been mainly used in discovering association rules and functional dependencies and image retrieval.

References

[1] Teknomo, and kardi. *"K-Means clustering tutorial"*, IEEE Press, 2003
[2] David Altman, Efficient Fuzzy Clustering of Multi-spectral Images, FUZZ-IEEE, 1999
[3] Steven Eschrich, Jingwei Ke, Lawrence O. Hall and Dmitry B. Goldgof, Fast Accurate Fuzzy Clustering through Data Reduction, IEEE Transactions on
[4] B. Jeon, Y. Yung and K. Hong *"Image segmentation by unsupervised sparse clustering, "* pattern recognition letters 27science direct,(2006) 1650-1664
[5] *Vicenc Torra, 2004"* Fuzzy C- means For fuzzy hierarchical
[6] E.H. Ruspini. A new approach to clustering. Information and control, 22-32.[65] K-means clustering Algorithm data mining tutorial started by KINGSLEYTAGBO at 12-14-2004
[7] J. Han and M. Cambers K. Data Mining: Concepts and Techniques. Morgan Kaufman, 2000.
[8] Steven Eschrich, Jingwei Ked, Lawrence O. Hall and Dmitry B. Goldgof, Fast Accurate Fuzzy Clustering through Data Reduction.
[9] Richard J. Hathaway and James C. Bezdek, Extending Fuzzy and Probabilistic Clustering to Very Large Data Sets, Journal of Computational Statistics and Data Analysis, 2006, accepted.

Vuda Sreenivasarao received the M.Tech degree in Computer Science & Engg from the Satyabama University, in 2007. He is research scholar in CSIT Department, JNT University Hyderabad Andhra Pradesh, India. His research interests include Network Security, Cryptography, and Data Mining & Artificial Intelligence.

Dr. S Vidyavathi received her PhD degree from IIT Mumbai; she is currently working as Associate Professor in CSIT Department, JNT University, and Andhra Pradesh, India.