International Journal of information Technology and Knowledge Management
July-December 2010, Volume 2

COMPARATIVE ANALYSIS OF FUZZY C MEAN AND HARD C MEAN ALGORITHM

Pawan Kumar [1], Mr. Pankaj Verma[2], Rakesh Shrma [3]

In this paper, we use FCM (Fuzzy C mean) clustering algorithm and HCM (Hard C Mean) algorithm. In this paper we compare the performance analysis of Fuzzy C mean (FCM) clustering algorithm and compare it with hard C mean algorithm. In this we compared FCM and HCM algorithm on different data sets. We measure complexity of FCM and HCM at different data sets. FCM clustering is a clustering technique which is separated from hard C mean that employs hard 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, Partioniong, Fuzzy Logic

1. INTRODUCTION

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.

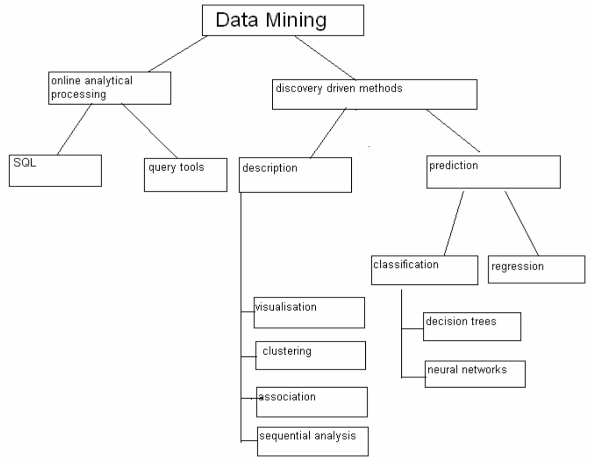

Fig 1.1 DATA MINING TECHNIQUE.

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 techniques fall into a group undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. Clustering technique is used for combining observed objects into clusters (groups), which satisfy tow main criteria:

- Each group or cluster is 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.

2. HCM AND FCM ALGORITHM

In this section we describe the HCM and FCM algorithm.

2.1. Hard C Mean Clustering

[1] M.TECH (IT) STUDENT OF M.M.UNIVERSITY MULLANA, HARYANA (INDIA)
[2] DEPTT. OF INFORMATION TECHNOLOGY, HCTM, KAITHAL, HARYANA (INDIA)
[3] DEPTT. OF INFORMATION TECHNOLOGY HCTM, KAITHAL, HARYANA (INDIA)

In non fuzzy or hard clustering, data is divided into crisp clusters, where each data point belongs to exactly one cluster.

- Used to classify data in crisp set
- Each data point will be assigned to only one cluster
- Clusters are also known as partitions
- U is a matrix with c rows and n columns
- The cardinality gives number of unique c partitions for n data points

In this clustering technique partial membership is not allowed .HCM is used to classify data in a crisp sense. By this we mean that each data point will be assigned to one and only one data cluster. In this sense, these clusters are also called as partitions that are partitions of the data. In case of hard c mean each data element can be a member of one and only one cluster at a time

2.1.1. Algorithm of HCM

```
1.    fix c(2<=c<n) and initialize the U matrix

          $U^{(0)}$  Є $M_c$

Then do r=0, 1, 2, 3……………

2. Calculate the center vectors{ $V^{®}$ with $U^{®}$ }

3. Update $U^{®}$ ; calculate the updated characteristic function(for a all i,k);

   1,$d_k^{®}$=min $d_k^{®for}$ for all jЄc

 {            $x_{ik}^{(r+1)}$=        0, otherwise

4. if ||$U^{()r-1)}$-$U^{®}$||<=δ(tolerance level

STOP: otherwise set r=r+1 and return to step 2.In step 4 the notation || || is any matrix norm such as the Euclidean norm.
```

Figure 2.1 HCM Algorithm

2.1.2. View of HCM Algorithm

In this algorithm first of all we fix the value of c between one and n and after that we calculate the initialization matrix for the given data elements which we have stored in the file data.txt, whose file pointer is finp.

Step.1. Calculation of Initialization Matrix:

For storing initialization matrix we use the file inint.txt whose file pointer is fi and we have stored each elements of initialization matrix in the file as a float data objects. First of all we have calculated the value of integer variable p whose value is equal to n/c and then from the first cluster to the second last cluster we move in such a way that in the first cluster we assign membership grade equal to one to the first p data elements and in the second cluster we assign membership grade equal grade to one to the first p data elements in the second cluster we assign membership grades equal to one to the next p elements and all the remaining elements in that cluster have membership grade equal to zero .

Step.2. Calculation of Center Vectors:

Formula for calculating center vectors

$$V_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^{m'} \times x_{kj}}{\sum_{k=1}^{n} u_{ik}^{m'}}$$

Through this formula we have calculated the center vectors of each clusters in file called as vect.txt. Whose file pointer is fv.

Step.3. Calculation for Distance Matrix:

Formula for calculating distance matrix is

$$d_{ik} = \left[\sum_{j=1}^{m}\left[x_{kj} - v_{ij}\right]^2\right]^{[1/2]}$$

Where x_{kj} is data element, d_{ik} is the distance matrix and v_{ij} is the element of the cluster center vector.

Through this formula we have calculated the distance matrix d[c] [n] for the given data elements and clusters and we have stored this distance matrix in the file called as dist.txt. Whose file pointer is fdi

Step.4. Calculating of new Membership Grade Matrix:

For storing new membership grades we have used the nmg.txt file whose file pointer is fn. Here, the membership grades are assigned in such a way that in each column of the new membership grade matrix the data element is assigned the membership grade equal to one only in that cluster to whom its corresponding distance is minimum and in all other remaining clusters its membership grade is zero.

2.2. Fuzzy c Mean Algorithm

FCM is an iterative algorithm. The aim of FCM is to find clusters centers (cancroids) that minimize a dissimilarity function.The FCM algorithm attempts to partition a finite collection of elements X={x_1, x2, x3………x_n} 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=v_i,i=1,2,3……………,c And a partition matrix U such that U=u_{ij},i=1,2,3,……………c, j=1,2,……………n

Where u_{ij} is a numerical value in [0, 1] that tells the degree to which the elements x_j 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

- Fuzzy set allows for degree of membership
- A single point can have partial membership in more than one class.
- There can be no empty classes and no class that contains no data points.

2.2.1. FCM Algorithm

1. Fix c (2<=c<n) and select a value for parameter m'. Initialize the partition matrix $^{(0)}$. Each step in this algorithm will be labeled r, where r=0, 1, 2…………

2 Calculate the c center {vi®} for each step

$$V_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^{m'} \times x_{kj}}{\sum_{k=1}^{n} u_{ik}^{m'}}$$

3 Calculate the distance matrix $D_{[c,n]}$.

$$D_{ij} = \left[\sum_{j=1}^{m} \left[x_{kj} - v_{ij} \right]^2 \right]^{[1/2]}$$

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

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

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

FIG2.2 FCM ALGORITHM

FCM iteratively moves the cluster centers to the “right” location with in a dataset.

2.2.2. View of FCM Algorithm:

In this algorithm first of all we fix the value of c between one and n and after that we calculate the initialization matrix for the given data elements which we have stored in the file data.txt, whose file pointer is finp.

Step.1. Calculation of Initialization Matrix:

Logic for calculating the initialization is same as that of HCM.

Step.2. Calculation of Center Vectors

Formula for calculating center vector is

$$V_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^{m'} \times x_{kj}}{\sum_{k=1}^{n} u_{ik}^{m'}}$$

Where x_{kj} is data element, u_{ik} is the element of membership grade matrix, v_{ij} is the element of the cluster center vector and m’ is the amount of fuzziness.

Through this formula we have calculated the center vectors of each cluster and we have stored the center vector of all the clusters in a file called vect.txt. Whose file pointer is fv.

Step.3. Calculation for Distance Matrix:

Formula for calculating distance matrix is

$$d_{ik} = \left[\sum_{j=1}^{m} \left[x_{kj} - v_{ij} \right]^2 \right]^{[1/2]}$$

Where x_{kj} is data element, d_{ik} is the distance matrix and v_{ij} is the element of the cluster center vector.

Through this formula we have calculated the distance matrix d[c][n] for the given data elements and clusters and we have stored this distance matrix in the file called as dist.txt. Whose file pointer is fdi. This is same as in HCM

Step.4. Calculation of new Membership Grade Matrix:

For storing new membership grades we have used the nmg.txt file whose file pointer is fn.

To calculate the new membership grade matrix the formula is:

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

Where r is iteration number, u_{ik} is the element of membership grade matrix, d_{ik} is the element of the distance matrix and m’ is the amount of fuzziness.

3. RESULTS

After implementation and study we get the following results.

3.1 Complexity Analysis of HCM 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 HCM is cd

I/O complexity of HCM is ndi

3.2. 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(ndc^{2}i), 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.

3.3. Comparative Analysis of Complexities of HCM and FCM

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

4. 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] Huang, E*xtensions to the k-means algorithms for clustering large data sets with*

categorical values. Data mining and knowledge discovery, 2:283-304, 1998.

[2] Teknomo, and kardi. "*K-Means clustering tutorial",* IEEE Press, 2003

[3] David Altman, *Efficient Fuzzy Clustering of Multi-spectral Images*, FUZZ-IEEE , 1999 [10] 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.

[4] Steven Eschrich, Jingwei Ke, Lawrence O. Hall and Dmitry B. Goldgof, Fast Accurate Fuzzy

Clustering through Data

Reduction,IEEE Transactions on

[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. Kamber K. Data Mining: Concepts and Techniques. Morgan Kaufman ,2000

[8] Steven Eschrich, Jingwei Ke, Lawrence O. Hall and Dmitry B. Goldgof, Fast Accurate Fuzzy Clustering through Data Reduction,

[9] B. Jeon, Y. Yung and K. Hong "Image segmentation by unsupervised sparse clustering, " pattern recognition letters 27science direct,(2006) 1650-1664