The Comparative study and Performance of HCM and MFPCM Algorithms on Iris data set

VUDA SREENIVASARAO

Professor & Head CSE, IT Dept. St .Mary's College of Engg. & Technology, Hyderabad ,India.

Abstract: 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. 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 HCM (Hard C -mean) clustering algorithm and MFPCM (Modified Fuzzy Possibilistic C - mean) clustering algorithm. In this paper we compare the performance analysis of Hard C mean (HCM) clustering algorithm and compare it with Modified Fuzzy possibilistic C mean algorithm. In this we compared HCM and MFPCM algorithm on different data sets. We measure complexity of HCM and MFPCM at different data sets. HCM clustering is a clustering technique which is separated from Modified Fuzzy Possibililstic C mean that employs Possibililstic partitioning.

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

1. 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. *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.

2. Hard C- Mean clustering algorithm :

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. In other words we can say that the sum of membership grades of each data point in all clusters is equal to one and in HCM membership grade of a specific data point in a specific cluster is one and in all the remaining clusters its membership grade is zero. Also number of clusters that is can't be less than or equal to one and they can't be equal to or greater than number of data elements because if number of clusters is equal to one than all data elements will lie-in same cluster and if number of clusters is equal to number of data elements than each data elements will lie in its own separate cluster. That is each cluster is having only one data point in this special case. The steps of HCM algorithm given below.

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

$$U^{(0)} \in M_C$$

Then for 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).

$$X_{ik}^{(r+1)} = \begin{cases} 1, d_i^{(r)} = \min d_{jk}^{(r)}\ forallj \in c \\ 0, otherwise \end{cases}$$

4. if $\|U^{()r-1)}-U^{®}\|<=\delta$(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.

3. Modified Fuzzy Possibililstic 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 Modified Fuzzy Possibililstic C - Mean Algorithm given below:

1. the objective function of the Modified Fuzzy Possibililstic C - Mean Algorithm 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:

$$u_{ij}=\left[\sum_{k-1}^{c}\left(\frac{d?\mathbf{x}_{\mathrm{j}},v_{\mathrm{i}})}{d?\mathbf{x}_{\mathrm{j}},v_{\mathrm{k}})}\right)^{2m/(m-1)}\right]^{-1}$$

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

$$t_{ij}=\left[\sum_{k-1}^{n}\left(\frac{d?\mathbf{x}_{\mathrm{j}},v_{\mathrm{i}})}{d?\mathbf{x}_{\mathrm{j}},v_{\mathrm{k}})}\right)^{2\eta/(\eta-1)}\right]^{-1}$$

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

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

4. Results:

4.1 Time complexity of HCM and MFPCM by varying no. of Clusters on Iris Data:

The implementation of HCM & MFPCM is done on iris Data in MATLAB. The data t contains 3 classes of 150 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two, the latter are NOT linearly separable from each other. The data set contain four attribute which are given below

The time complexity of HCM [11] is O(ndc^{2}i) and time complexity of MFPCCM is O(ncdi). Now keeping no. of data points constant, lets assume n=100, d=3, i=20 and varying no. of clusters, we obtain the following table and graph. Where n= number of data point, c= number of cluster, d= dimension, i= number of iteration

Table 4.1 Time Complexity when Number of cluster varying

S.No.	Number of Cluster	HCM Time Complexity	MFPCM Time Complexity
1	1	1000	2000
2	2	9000	5000
3	3	24000	8500
4	4	47000	10500

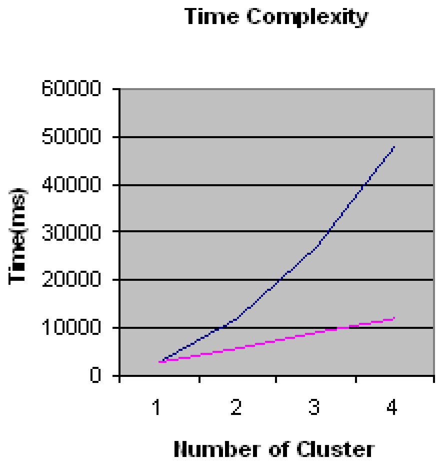

Figure 4.1 Time complexity of HCM and MFPCM by varying no. of Clusters

Now keeping no. of cluster constant, lets assume n=140, d=3, c=3 and varying no. of Iteration, we obtain the following table and graph.

Table4.2 Time Complexity when Number of Iterations varying

S.No.	Number of Iteration	HCM Time Complexity	MFPCM Time Complexity
1	5	6000	3000
2	10	10000	6000
3	15	16000	8000
4	22	25000	12000

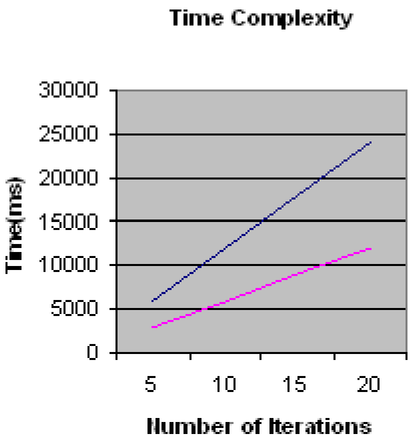

Figure 4.2 Time complexity of HCM and MFPCM by varying no. of Iterations

4.2. Comparison of space complexity of HCM and MFPCM :

The space complexity of HCM is O(nd+nc) and MFPCM is O(cd). Now keeping no. of data points constant, lets assume n=140, d=3 and varying no. of clusters we obtain the following graph.

Table4.3 Space Complexity when Number of Clusters varying

S.No.	Number of Cluster	HCM Space Complexity	MFPCM Space Complexity
1	10	500	4
2	15	700	8
3	20	900	12
4	25	1100	16

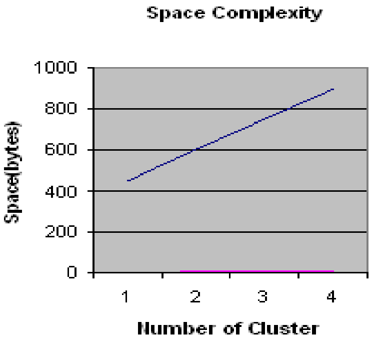

Figure4.3 space complexities of HCM and MFPCM by varying number of clusters

4.3 Complexity Analysis of HCM Algorithm :

The asymptotic efficiency of the algorithm has following notations:

i number HCM 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(ndc^2i), where empirically I grows very slowly with n,c and d.

The memory complexity of HCM 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, HCM 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

4.4 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

Table 4.4. Comparative Analysis of Complexities of HCM and MFPCM

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

5. Conclusion:

In partitioning based clustering algorithms, the number of final cluster (k) needs to be defined beforehand. Also, algorithms have problems like susceptible to local optima, sensitive to outliers, memory space and unknown number of iteration steps required to cluster. The time complexity of the MFPCM is O(ncdi) The memory complexity of MFPCM is cd and the input output complexity will be O(ndi). 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. The time complexity of the Hard C Mean algorithm is O(ndc^{2}i). The memory complexity of MFPCM is O(nd + nc),and the disk input output complexity will be O(ndi)

6. References:

[1] ude hemanth.D, D.Selvathi and J.Anitha,"Effective Fuzzy Clustering Algorithm for Abnormal MR Brain Image Segmentation",Page umber 609-614, International/Advance Computing Conference (IACC 2009),IEEE,2009.

[2] Sorin Istrail, "An Overview of Clustering Methods", With Applications to Bioinformatics.

[3] Wei Wang, Chunheng Wang, Xia Cui, Ai Wang, "*A Clustering Algorithm Combine the FCM algorithm with Supervised Learning Normal Mixture Model*"**,** IEEE 2008.

[4] Deepak Agrawal "*Web Data Clustering using FCM and Proximity Hints from Text as well as Hyperlink-structure*"**,** IEEE 2008.

[5] M. Brej and M. Sonka, "*Object localization and border detection criteria design in edge-based image segmentation automated learning from examples*"**,** IEEE Transactions on Medical imaging, vol. 19, pp. 973-985, 2000.

[6] S. Chen and D. Zhang, *"Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure"***,** IEEE Transactions on Systems, Man and Cybernetics, vol. 34, pp. 1907-1916, 1998.

[7] O. Sojodishijani, V. Rostami and A. R. Ramli, "*Real time color image segmentation with non-symmetric Gaussian membership functions",* Fifth International Conference on Computer Graphics, Imaging and Visualization, pp. 165-170, 2008.

[8] M. S. Yanp, K.L. Wu and J. Yub, *"A novel fuzzy clustering algorithm",* IEEE International Symposium on Computational Intelligence in Robotics and Automation, vol. 2, pp. 647- 652, 2003.

[9] L. Hui, *"Method of image segmentation on high-resolution image and classification for land covers",* Fourth International Conference on Natural Computation, vol. 5, pp. 563-566, 2008.

[10] D. L. Pham, *"Spatial models for fuzzy clustering",* Laboratory of Personality and Cognition, Gerontology Research Center, 2001.

[11] R. J. Almeida and J. M. C. Sousa**,** *"Comparison of fuzzy clustering algorithms for Classification",* International Symposium on Evolving Fuzzy Systems, pp. 112-117, 2006.

[12] Mohamed Fadhel Saad and Adel M.Alimi " Modified Fuzzy Poossibilistic C-means" ,International multi conference of Engineers and computer scientists -2009 Vol1

[13] Vuda sreenivasarao and Dr.S.Vidyavathi**,** "Comparative investigations and performance analysis of FCM and MFPCM algorithms on IRIS data", India Journal of computer science and engineering, Volume 1, Issue 2, July 2010, pp 145-151.

[14] Vuda sreenivasarao and Dr.S.Vidyavathi, "Comparative analysis of Fuzzy C-Mean and modified fuzzy possibilistic C-Mean algorithms in Data mining", International Journal of Computer Science and Technology , Volume 1, Issue no 1, Sep 2010, pp 100-102.

7. Author profile:

VUDA SREENIVASARAO received his M.Tech degree in Computer Science & Engg from the Satyabama University, in 2007.Currently working as Professor & Head in the Department of CSE & IT at St.Mary's college of Engineering & Technology, Hyderabad, India.. He is currently pursuing the PhD degree in CSE Depart in Singhania University, Rajasthan. His main research interests are Data Mining, Network Security, and Artificial Intelligence. He has got 10years of teaching experience .He has published 21 research papers in various international journals. He is a life member of various professional societies like MIACSIT, MISTE. MIAENG.