Proceeding of IC-ITS 2015 e-ISBN:978-967-0850-07-8
International Conference on Information Technology & Society
8-9 June 2015, Kuala Lumpur, MALAYSIA

Optimal Clustering Using Modified Fuzzy C-Means Clustering Algorithm

Ahamed Shafeeq B M

Department of Computer Science & Engineering, Manipal Institute Of Technology, ManipalUniversity, Manipal -576104, India

email: ahamed.shafeeq@manipal.edu

Abstract. Fuzzy clustering has been widely studied and applied in a variety of applications and areas. In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. The investigation is needed to reveal whether the optimal number of clusters can be found on the run based on the cluster quality measure. The silhouette coefficient is the one of measure used to measure the quality of clusters. In the practical scenario, it is very difficult to fix the number of clusters in advance. In this paper we propose anoptimal clustering of data with modified Fuzzy C-Means algorithm. The proposed method works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number (K=2) of clusters required. In the former case it works same as Fuzzy C-means algorithm. In the latter case the algorithm computes the quality of clusters for each set of clusters. The process is repeated by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality. It is observed that the modified Fuzzy C-means algorithm produces quality clusters compared to the Fuzzy C-means clustering. It assigns the data point to their appropriate class or cluster more effectively.

Keywords: Fuzzy C-means clustering, Cluster quality, Silhouette coefficient.

INTRODUCTION

A fundamental problem that frequently arises in a great variety of fields such as data mining andknowledge discovery, and pattern classification is the clustering problem [1]. Nowadays, there is an urgent need for good quality clustering algorithms to address the huge amount of data. The quality of information play a very crucial role in decision makingof policy has attracted a great deal of attention in the information industry and in society as a whole. There isvery large amount of data availability in real world and it is very difficult to excess the useful informationfrom this huge database and provide the information to which it is needed within time limit and in requiredpattern. So data mining

is the tool for extracting the information from huge database and present it in theform in which it is needed for each specific task. The use of data mining is very vast. Clustering is the one of the most commonly used data mining technique. Cluster analysis of data is an important task in knowledge discovery and data mining. Cluster analysisaims to group data on the basis of similarities and dissimilarities among the data elements. In classical cluster analysis, these classes are required to form a partition of data such that the degree of association is strong for data within blocks of partition and weak for data in different blocks [2]. It is very helpful inapplication like to know the trend of market, fraud detection, and shopping pattern of customers, productioncontrol and science exploration etc. The process canbe performed in a supervised, semi-supervised or unsupervised manner [3]. Different algorithms have beenproposed which take into account the nature of the data and the input parameters in order to cluster the data [4].
There are many fuzzy clustering methods beingintroduced [5]. Fuzzy *C*-means clustering algorithm is one ofmost important and popular fuzzy clustering algorithms. Atpresent, the FCM algorithm has been extensively used infeature analysis, pattern recognition, image processing,classifier design, etc. [6][7]. The FCM isbased on hard c-means (HCM) and has been constructed by fuzzification of HCM. Some FCMs are used in the field ofclustering. Each FCM corresponds with the way to fuzzifythe HCM. FCMs are pointed out that it is difficult to classify data with nonlinear borders becauseconventional FCMs use squared distance between each datumand each cluster center for their dissimilarity.

PRELIMINARIES

Fuzzy Logic (FL) was initiated in 1965 [8, 9], by Lotfi A. Zadeh, professor for computer science at the University of California in Berkeley. FL techniques have been used in image-understanding applications such as detection of edges, feature extraction, classification, and clustering. Fuzzy logic poses the ability to mimic the human mind to effectively employ modes of reasoning that are approximate rather than exact. In soft computing, tolerance and impression are explored in decision making. The exploration of the tolerance for imprecision and uncertainty underlies the remarkable human ability to understand distorted speech, decipher sloppy handwriting, comprehend distinctions of natural language, summarize text, recognize and classify images. *Fuzzy c-means* (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. The FCM algorithm was proposed by Dunn in 1974, and then was improved and extended by Bezdek later in 1981[1].This technique was originally introduced by Jim Bezdek as an improvement on earlier clustering methods. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters.

In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means Algorithm [1]. The FCM algorithm attempts to partition a finite collection of n elements $X=\{X_1 \ldots . X_n\}$ into a collection of c fuzzy clusters with respect to some given criterion.The Fuzzy *C*-Means is sensitive to initial conditions and the

algorithmusually leads to local minimum results. The global Fuzzy *C*-Means clustering algorithm(GFCM) which is an incremental approach to clustering and it doesnot depend on any initial conditions and the better clusteringresults are obtained through a deterministic global searchprocedure [10]. The effective clustering can be formed with the selection of best m value and the membership function [11].

FCM starts with an initial guess for the cluster centres, which are intended to mark the mean location of each cluster. The initial guess for these cluster centres is most likely incorrect. Additionally, FCM assigns every data point a membership grade for each cluster. By iteratively updating the cluster centres and the membership grades for each data point, FCM iteratively moves the cluster centres to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster centre weighted by that data point's membership grade.

Membership Matrix U (partition matrix)

In this subsection, the data for clustering and the membership by which the each datum belong to the each cluster are defined. The data set X = X_1, X_2…X_N is given. The membership by which X_i belongs to cluster j is denoted by N X C matrix u with elements U_{ij} (i∈ {1….N}, j∈ {1 …. C}). The constraint for U is: $\sum_{j=1}^{C} U_{ij}$

FUZZY C-MEANS CLUSTERING

The algorithm is based on the assumption that the desired number of cluster c is given and in addition, a particular distance, a real number m (1,∞), and a small positive number ε, serving as a stopping criterion, are chosen.

Fuzzy C-Means Algorithm: The algorithm for partitioning, where each cluster's center is represented bymean value of objects in the cluster.

Input: k: the number of clusters. D: a data set containing n objects. M=2.[2]
Output: A set of k clusters.
Method:
Algorithm steps:

1. Initialize U=[Uij] matrix, U(0)
2. At k-step calculated the centers vectors $C^{(k)}$=[C_j]

$$C_i = \frac{\sum_{i=1}^{N} U_{ij} X_i}{\sum_{i=1}^{N} U_{ij}^{m}}$$

3. Update $U^{(k)}$, $U^{(k+1)}$

$$Uij=\frac{1}{\sum_{k=1}^{C}\frac{\|X_i-C_i\|}{\|X_i-C_k\|}^{\frac{2}{m-1}}}$$

4. If $\| U^{(k+1)}-U^{(k)}\|<\varepsilon$ then stop, otherwise return to step 2

One of the main disadvantages of Fuzzy C-Means is the fact that you must specify the number of clusters as aninput to the algorithm. As designed, the algorithm is not capable of determining the appropriate number ofclusters and depends upon the user to identify this in advance.

MODIFIED FUZZY C-MEANS CLUSTERING

The Fuzzy C-Means algorithm finds the predefined number of clusters. In the practical scenario, it is very muchessential to find the number of clusters for unknown dataset on the runtime. The fixing of number of clustersmay lead to poor quality clustering. The proposed method finds the number of clusters on the run based onthe cluster quality output. This method works for both the cases i.e. for known number of clusters in advanceas well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or byinput the minimum number of clusters required. In the former case it works same as Fuzzy C-Means algorithm. Inthe latter case the algorithm computes the new clusters by incrementing the cluster counter by one in eachiteration until it satisfies the validity of cluster quality threshold. The modified algorithm is as follows:

Input: k: number of clusters (for dynamic clustering initialize k=2)
Fixed number of clusters = yes or no (Boolean).
D: a data set containing n objects. m=2 ,ϵ=0.001.
Output: A set of k clusters.

Method:

1. A) For unknown number of clusters -- Initialize K=2.
   B) For the fixed number of clusters -Take the input K from the user.
   Boolean: fixed_number_of_clusters= yes or No.
2. If fixed_number_of_clusters=yes then repeat step 4 to 7 and go to step 10
3. Else Repeat the steps 4 to 7 for k and k+1
4. Initialize U= [Uij] matrix, U(0).
5. At k-step calculated the centers vectors $C^{(k)}=[C_j]$

$$C_i=\frac{\sum_{i=1}^{N} U_{ij} X_i}{\sum_{i=1}^{N} U_{ij}^{m}} \quad (1)$$

6. Update U(k), U(k+1)

$$Uij=\frac{1}{\sum_{k=1}^{C}\frac{\|X_i-C_i\|}{\|X_i-C_k\|}^{\frac{2}{m-1}}} \quad (2)$$

7. If $\| U^{(k+1)} -U^{(k)}\| < \varepsilon$ then stop, otherwise return to step 5

8. Compute silhouette coefficient(k) and silhouette coefficient(k+1) from Eq.6
9. If silhouette coefficient(k) < silhouette coefficient(k+1)
   Then k= k + 1 go to step 4 else return optimal clusters (k) and go to step 10.

10. STOP

Algorithm: Optimal clustering of data with modified Fuzzy C-Means Algorithm

The silhouette coefficientis intrinsic method to assess the clustering quality [1]. The quality of clustering is evaluated using Silhouette coefficient. Silhouette of object Oj determines how much Oj belongs to its cluster. The coefficient value is in between -1 and 1.Calculate a(O) as the average distance between Oand all other objects in the cluster to which o belongs. Similarly, b(O) is the minimum average distance from Oto all clusters to which Odoesnot belong.

Formally, suppose O$\in C_i$(1$\leq$ i $\leq$k) then

$$a(O)=\frac{\sum_{o' \in Ci, o \neq o'} dist(o,o')}{|C_i|-1} \quad (4)$$

And

$$b(O)= \min \left\{ \frac{\sum_{o' \in C_j} dist(o,o')}{|C_j|} \right\} \quad (5)$$

C_j:1$\leq j \leq k, j \neq$i

The silhouette coefficient of o is then defined as

$$S(O)=\frac{b(O)-a(O)}{max\{a(O),b(O)\}} \quad (6)$$

Therefore, whenthe silhouette coefficient value of Oapproaches 1, the cluster containing Ois compactand Ois far away from other clusters, which is the preferable case. To measure a cluster's fitness within a clustering, we can compute the average silhouettecoefficient value of all objects in the cluster. To measure the quality of a clustering,we can use the average silhouette coefficient value of all objects in the data set.

Silhouette Coefficient	Interpretation
0.71 – 1	Strong cluster
0.51 – 0.7	Reasonable cluster
0.26 – 0.5	Weak or artificial cluster
≤ 0.25	No cluster found

[Source: Efficient and Effective Clustering Methods for Spatial Data Mining- Raymond T. Ng, Jiawei Han]

EXPERIMENTAL RESULTS AND ANALYSIS

In this section, in order to verify the effectiveness of the proposed algorithm, we take some randomnumbers of 300, 500, 1000 and 2000 data points. The experimental results show that the proposed methodoutperforms basic Fuzzy C-Means algorithm in quality and optimality for the unknown data set. The experiment isconducted on synthetic data

set. The new algorithm works for fixed number of clusters as well as unknownnumber of clusters.

Table 1: Experimental Results

Data points	No. of clusters (FCM)	No. of clusters (OFCM)	Silhouette coefficient (FCM)	Silhouette coefficient (OFCM)	Time Taken (Sec). (FCM)	Time Taken(Sec.) (OFCM)
300(Figure 1)	5	2	0.7281	0.8150	0.007799	0.074820
500(Figure2)	4	3	0.7334	0.7966	0.010389	0.016548
1000(Figure 3)	7	2	0.7238	0.7986	0.026259	0.029370
2000(Figure4)	3	5	0.7324	0.7919	0.113786	0.119178

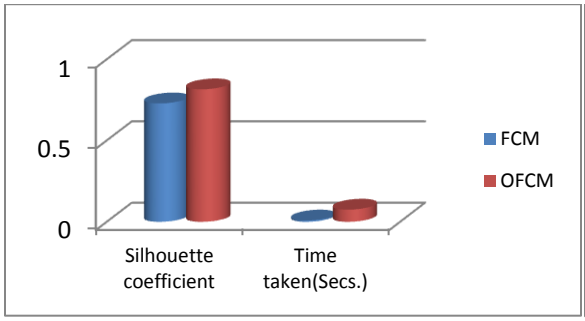

FIGURE 1:FCM VS OFCM

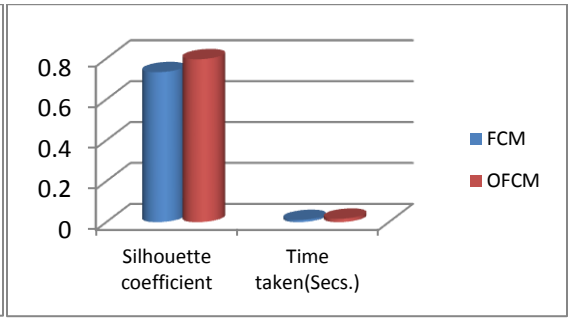

FIGURE 2:FCM VS OFCM

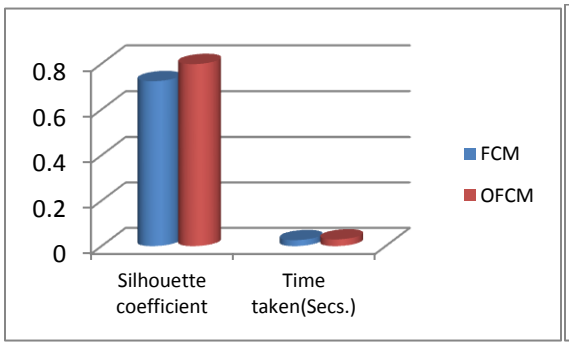

FIGURE 3:FCM Vs OFCM

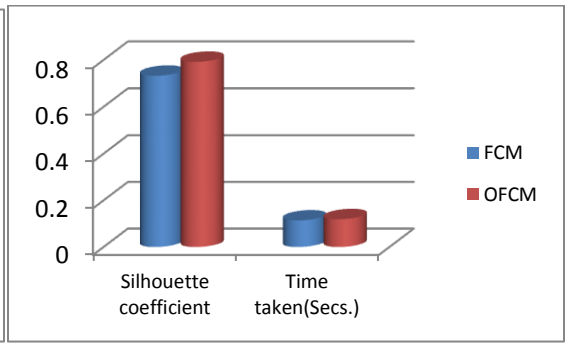

FIGURE4:FCM Vs OFCM

The above experimental results on synthetic data show that the proposed method gives optimal numberof clusters for the unknown data set. It is also observed that the time taken in the proposed method is almost same as Fuzzy C-Means algorithm for smaller data set. The algorithm is developed and tested for efficiency of different data points in MATLAB. The algorithm takes more computational time compared to the Fuzzy C-Meansalgorithm for large dataset in some cases. The algorithm works same as Fuzzy C-Means for the fixed number ofclusters. For the unknown data set it starts with the minimum number of cluster given by the user and afterthe completion of every set of

iteration, the algorithm checks for efficiency and it repeats by incrementingthe number of cluster by 1 until it reaches the termination condition.

CONCLUSION

Fuzzy *C*-Means is one of the algorithms for clustering based on optimizing an objective function and sensitive to initial conditions. This paper proposed an improved data clustering for the unknown nature of data set. The algorithm works well forthe unknown data set with better results than the basic Fuzzy C-Means clustering. The Fuzzy C-Means algorithm is well known forits simplicity and the modification is done in the proposed method with retention of simplicity. The Fuzzy C-Meansalgorithm takes number of clusters (K) as input from the user. The major problem in any clustering algorithm isfixing the number of clusters in advance. In the practical scenario, it is very difficult to fix the number ofcluster in advance. If the fixed number of cluster is very small then there is a chance of putting dissimilarobjects into same group and suppose the number of fixed cluster is large, and then the more similar objects willbe put into different groups. The proposed algorithm will overcome this problem by finding the optimalnumber of clusters on the run. The main drawback of the proposed approach is that it takes morecomputational time than the Fuzzy C-Means for larger data sets. Future work can focus on how to reduce the timecomplexity without compromising cluster quality and optimality. More experiments will be conducted withnatural datasets with different features.

REFERENCES

1. Bezdek, James C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. ISBN 0-306-40671-3.
2. George J Klir and Bo Yuan. (2012). Fuzzy sets and fuzzy logic theory and applications, pp 358-365, PHI.
3. Ahmed, Mohamed N.; Yamany, Sameh M.; Mohamed, Nevin; Farag, Aly A.; Moriarty, Thomas (2002). A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data. IEEE Transactions on Medical Imaging21 (3): 193–199. doi:10.1109/42.996338.
4. Jiawei Han and MichelineKamber (2006). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, second Edition, pp 443-490.
5. F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler(1999). “Fuzzy cluster analysis,” Wiley Press, New York.
6. M. C. Clark, L. O. Hall (1995). MRI segmentation using fuzzy clustering techniques: integrating knowledge, http*://www.csee.usf.edu/*.
7. Y. W. Lim, S. U. Lee(1990). On the Color Image Segmentation Algorithm Based on the Thresholding and the Fuzzy c-means Techniques, Pattern Recognition, 23(9): pp. 935-951.

8. L.A.Zadeh (1998). Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems", Soft computing, pp 23-25, Springer-Verlag.
9. L.A. Zadeh (2002).Toward a perception-based theory of probabilistic reasoning with imprecise probabilities, Journal of Statistical Planning and Inference, Elsevier, pp 233–264.
10. Weina Wang, Yunjie Zhang, Yi Li and Xiaona Zhang (2006). The Global Fuzzy C-Means Clustering Algorithm, Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China.
11. Lin Zhu, Fu-Lai Chung, and Shitong Wang (2009). Generalized Fuzzy C-Means Clustering Algorithm with Improved Fuzzy Partitions", IEEE Transactions on Systems, Man, andCybernatics—Part B: Cybernstics, Vol. 39, No. 3.