ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

An Improvement over Fuzzy c-Means: A Prototype based approach

K. Mrudula[1], E. Keshava Reddy [2]

Research Scholar, Department of Mathematics, Jawaharlal Nehru Technological University Anantapur, Ananthapuramu, A.P. India[1]

Professor, Department of Mathematics, Jawaharlal Nehru Technological University Anantapur, Ananthapuramu, A.P. India[2]

ABSTRACT: Fuzzy c-Means clustering method is a leading data clustering algorithm. It is an iterative approach, where in, each pattern is assigned to its nearest cluster centre with a membership value in each cluster. It requires to compute nct number of distance computations, where n is the size of the dataset, c is the number of clusters and t is the number of iterations. However, for large value of n, it may be very difficult to implement. The aim of this paper is to reduce the number of distance computations by selecting few prototypes (representative patterns), say s, from n and use these prototypes in the iterative process instead of the whole data. Hence the number of distance computations becomes sct where $s << n$. The novelty of this approach is that *each prototype is fuzzy* i.e., each prototype is considered as fuzzy representative pattern of some group of patterns in the given data. Experimentally it is proved that the proposed method take less number of distance computations with a negligible reduction in clustering accuracy.

KEYWORDS: Data clustering, Fuzzy c-Means, prototypes.

I. INTRODUCTION

Fuzzy clustering is an active area of research with highly advanced applications including Image Processing, Document clustering, information retrieval systems, artificial intelligence and robotics etc. [7, 12]. Since the introduction of fuzzy sets by Zadeh, the fuzzy computing techniques have been highly effective with rich mathematical foundation [12]. The fuzzy clustering techniques such as Fuzzy c-Means (FCM) have been extensively studied and used in various scientific applications. Fuzzy c-Means (FCM) has been considered as an extension of the hard c-means clustering method and it is proved that FCM can achieve improved clustering accuracy than conventional c-means in case of non-isotropic clusters check the proof [9]. Unlike the hard clustering methods, in fuzzy clustering, each pattern can belong to multiple clusters with a degree of membership (belongingness) in each cluster. It is an iterative method, where in each iteration; it updates the membership value of every pattern in each individual cluster and recomputes the cluster centre. The time complexity of FCM is $O(n)$, where n is the number of patterns in the given data. However, for large value of n, it may be very difficult to implement. Further, it requires nct number of distance computations in the entire iterative process, c is the number of clusters and t is the number of iterations.

The aim of this paper is to reduce the number of distance computations by selecting few prototypes (representative patterns), say s, from n and use these prototypes in the iterative process instead of the whole data. Thereby, the number of distance computations becomes sct where $s << n$. The novelty of the proposed approach is that each *prototype is fuzzy* i.e., each prototype is considered as fuzzy representative pattern of some group of patterns in the given data. Experimentally it is proved that the proposed method takes less number of distance computations with a negligible reduction in clustering accuracy.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

The remaining paper is organized as follows. Section II explains the related work to improve FCM algorithm in detail. Section III introduces the proposed Prototype based FCM (PFCM). Experimental results are presented in Section IV and the conclusions are provided in Section V.

II. RELATED WORK

This section outlines the Fuzzy c-Means (FCM) clustering algorithm for the sake of completeness and then presents some recent attempts presented in the literature to reduce the running time of FCM.

The objective of FCM algorithm is to partition n data items (patterns) into c distinct clusters. It was initially proposed by Dunn and then extended by Bezdek [7, 13]. The iterative process starts by selecting c patterns as initial cluster centers and assigns each pattern to different clusters with a designated membership in each cluster such that the sum of all memberships of each pattern in various clusters should be equal to one. The membership value of a pattern depends on its nearestness to the particular cluster centre in each cluster.

Let $D = \{p_1, p_2, p_3, ..., p_n\}$ be a dataset and $V = \{v_1, v_2, v_3, ..., v_c\}$ be the set of initial cluster centres. The FCM clustering algorithm determines a partition of D by minimizing the objective function J which is defined as follows:

$$J = \sum_{i=1}^{n}\sum_{j=1}^{c} \mu_{ij}^{m} \left\| p_i - v_j \right\|^2$$

where m is the fuzziness index, $m \in [1, \infty]$ and μ_{ij} represents the membership of i^{th} pattern to j^{th} cluster centre.

After each iteration, the membership and cluster centres are updated as follows:

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

where $1 << i << c$ and $1 << j << n$

$$v_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} x_j}{\sum_{j=1}^{n} \mu_{ij}^{m}}; \text{ where} 1 << i << c$$

It has been studied that FCM often struck at local optimum. The global minimum value of J depends on the selection of initial cluster centres [2]. Further, it requires nct number of distance computations in the entire iterative process where t is the number of iterations.

Some improvements have been proposed in order to reduce the time complexity of hard c-means clustering method viz., reduce the number of distance computations in each iteration keeping the important points in the buffer (primary memory) while discarding unimportant points [3], reduce the number of scans over the data i.e., single pass *k*-means [8], bootstrap averaging [5], etc. More recently some prototype based approaches become popular, which produce quality clustering results similar to the conventional hard c-means method with a greater reduction in running time [10]. Some remarkable improvements over FCM for large datasets were proposed in [7, 13, 14, 15], which have motivated the present work.

The scope of this proposed work is to reduce the running time of FCM by reducing the number of distance computations. In this work we introduce the ***fuzzy prototypes*** that can be used in the iterative process instead of the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

whole data. To the best of our knowledge, the idea of finding ***fuzzy prototypes*** is not well explored in the literature. The proposed prototype based FCM (PFCM) is presented in the following section.

III. PROPOSED PROTOTYPE BASED FUZZY C-MEANS (PFCM)

The proposed prototype based Fuzzy c-Means (PCFM) runs in two phases. In the initial phase, it determines a set of fuzzy prototypes from n patterns in a single scan over the given data [11]. Each prototype, denoted by l_j is the representative of a set of patterns which are at a distance less than ε. It is to be noticed that, each pattern p_k may belong to one or more prototypes l_j. The membership of p_k in l_j is denoted by $\mu_{l_j p_k}$ and it is computed using the formula.

$$\mu_{l_j p_k} = \frac{1}{\sum_{i=1}^{s} \left(\frac{d(l_j, p_k)}{d(l_i, p_k)} \right)^{2/(m-1)}}$$

where m is the fuzziness index.

The sum of memberships of each pattern in various prototypes is equal to 1 i.e., $\sum_{j=1}^{s} \mu_{l_j p_k} = 1$. Here the number of prototypes, say s, depends on the value of ε. Larger the value of ε; lesser the number of prototypes and vice versa. The algorithm to find the ***fuzzy prototypes*** in presented in **Algorithm 1**. Let the set of prototypes after the first phase be $L = \{l_1, l_2, l_3, ..., l_s\}$. During the first phase, the following matrix M_S which is of size $n \times s$, is computed and stored in the primary memory.

$$M_S = \begin{bmatrix} \mu_{l_1 p_1} & \mu_{l_2 p_1} & . & . & . & \mu_{l_s p_1} \\ \mu_{l_1 p_2} & \mu_{l_2 p_2} & . & . & . & \mu_{l_s p_2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \mu_{l_1 p_n} & \mu_{l_2 p_n} & . & . & . & \mu_{l_s p_n} \end{bmatrix}_{n \times s}$$

In the matrix M_S the sum of elements in each row is equal to 1.

In the second phase, we run the FCM algorithm by selecting c initial cluster centers randomly from D. Here, in place of the entire dataset D, the set of prototypes L are used in the iterative process. Note that $|L| << |D|$. Each prototype l_j may belong to one or more clusters C_k s. The membership of l_j in C_k is denoted by $\mu_{C_k l_j}$ and it is computed using the formula

$$\mu_{C_k l_j} = \frac{1}{\sum_{i=1}^{c} \left(\frac{d(C_k, l_j)}{d(C_i, l_j)} \right)^{2/(m-1)}}$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

The sum of memberships of each prototype in various clusters is equal to 1. i.e., $\sum_{j=1}^{c} \mu_{C_j l_i} = 1$. During the second phase, the following matrix M_D of size $s \times c$ can be computed.

$$M_D = \begin{bmatrix} \mu_{C_1 l_1} & \mu_{C_2 l_1} & . & . & . & \mu_{C_c l_1} \\ \mu_{C_1 l_2} & \mu_{C_2 l_2} & . & . & . & \mu_{C_c l_2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \mu_{C_1 l_s} & \mu_{C_2 l_s} & . & . & . & \mu_{C_c l_s} \end{bmatrix}_{s \times c}$$

In the matrix M_D, the sum of elements in each row is equal to 1. In each iteration of PFCM, the approximate fuzzy membership of each pattern p_i in the cluster C_k can be computed using the following formula:

$$\mu^*_{C_k p_i} = \frac{\mu_{l_j p_i} + \mu_{C_k l_j}}{2}$$

Where l_j is the prototype, in which the membership of p_i is maximum among all prototypes.

It can be easily verified that the value of $\mu^*_{C_k p_i}$ is very close to the exact fuzzy membership $\mu_{C_k p_i}$.

$$\mu_{C_k p_i} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d(C_k, p_i)}{d(C_j, p_i)} \right)^{2/(m-1)}}$$

For large datasets, where n is very large, computing the approximate $\mu^*_{C_k p_i}$ takes less computations compare to $\mu_{C_k p_i}$ in each iteration of FCM. Hence, the proposed prototype based FCM (PFCM) converges more quickly than the conventional FCM. Interestingly, using this PFCM, it is possible to achieve the clustering quality which is very close to that obtained using the conventional FCM. The small reduction in the clustering quality is accepted for large data applications as long as there is great reduction in running time.

IV. EXPERIMENTAL STUDY

This section presents a detailed experimental study on some benchmark datasets collected from UCI Machine Learning Repository [1]. The details of the datasets are given in Table 1. Experiments are conducted on a machine with corei3 processor and 8GB RAM. To run FCM algorithm, the initial centers are selected randomly from the dataset as per the number of clusters *c* specified for each dataset.

Dataset	Number of Patterns (*n*)	Number of dimensions (*d*)	Number of Clusters (*c*)
Iris	150	3	4
Pendigits	10992	16	10
OCR	10003	192	10
LIR	20000	16	26
Shuttle	58000	9	7

Table 1: Properties of the benchmark datasets collected from UCI Machine Learning Repository

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

Algorithm 1: Generating Fuzzy Prototypes (D, ε)

$L = \phi$;

for each $p_i \in D$ ***do***

Find a $l_j \in L$ such that $\|l_j - p_i\| \leq \varepsilon$

if there is no such l_j or when $L = \phi$ ***then***

$$L = L \cup \{p_i\};$$

end if

end for

for each $p_i \in D$ ***do***

Compute the membership of p_i in each leader l_j using the following formula

$$\mu_{l_j p_i} = \frac{1}{\sum_{k=1}^{s} \left(\frac{d(l_j, p_i)}{d(l_k, p_i)} \right)^{2/(m-1)}}$$

end for

Output:

L and the matrix M_S.

To run the proposed prototype based FCM, the value of the ε on every dataset is fixed as the average distance in 5% of randomly selected patterns from the data. The fuzziness index *m* is also fixed that *m*=2. The clustering quality is assessed using Rand-Index [4]. The Clustering Accuracy (CA) of FCM and PFCM are compared in Fig1. From Fig 1, it can be observed that PCFM has very less reduction in the CA is 5% (on average), when compared to conventional FCM, on the selected benchmark datasets as shown in Table 1.

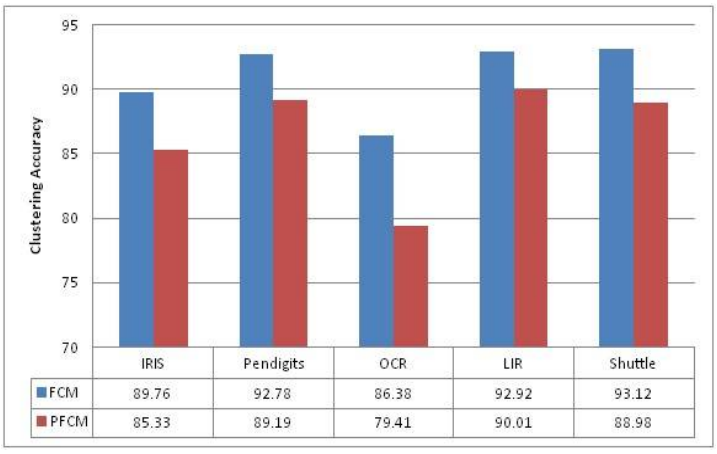


Fig.1. Comparing Clustering Accuracy (CA)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

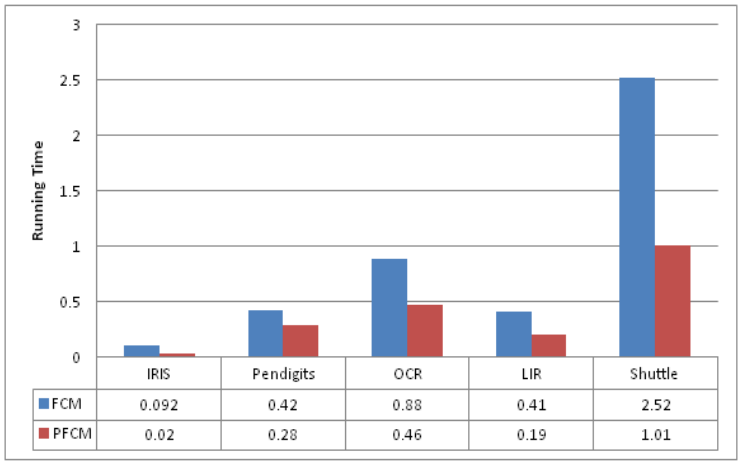
International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

Fig 2 shows that PFCM takes very less time when compared to Conventional FCM. From Fig 2, it is observed that there is 78.26% reduction in running time on Iris data. On pendigits the reduction in running time is 33.33. Similarly we can observe that 47.72%, 53.65% and 59.92% reduction in running time over OCR, LIR and Shuttle datasets respectively.



	IRIS	Pendigits	OCR	LIR	Shuttle
FCM	0.092	0.42	0.88	0.41	2.52
PFCM	0.02	0.28	0.46	0.19	1.01

Fig.2. Comparing Running Time

From the experimental study, it is proved that the proposed prototype based FCM result a great reduction in running time while there is small deviation in the clustering quality.

V. CONCLUSIONS

This paper presented a new prototype based Fuzzy c-Means (FCM) clustering method. The objective of this prototype based approach is to reduce the running time of FCM so that it can be used in case of very large datasets. The key idea is to identify some prototypes from the entire dataset and run FCM over these prototypes. Experimentally it is proved that the proposed method reduces the running time of FCM with a little compromise in the clustering accuracy.

REFERENCES

[1] Arthur Asuncion and David Newman. Uci machine learning repository, 2007
[2] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3):191–203, 1984.
[3] P. S. Bradley, U. Fayyad, and C. Raina. Scaling clustering algorithms to large databases. In Proceedings of Fourth International Conference on Knowledge Discovery and Data Mining, pages 9–15, AAAI Press, 1998.
[4] Roelof K Brouwer. Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. Journal of Intelligent Information Systems, 32(3):213–235, 2009.
[5] Ian Davidson and Ashwin Satyanarayana. Speeding up k-means clustering by bootstrap averaging. In IEEE data mining workshop on clustering large data sets, 2003.
[6] Gan, Guojun, Qiujun Lan, and C. Ma. "Scalable clustering by truncated fuzzy c-means." Big Data and Information Analytics (2016).
[7] Timothy C Havens, James C Bezdek, Christopher Leckie, Lawrence O Hall, and Marimuthu Palaniswami. Fuzzy c-means algorithms for very large data. IEEE Transactions on Fuzzy Systems, 20(6):1130–1146, 2012.
[8] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.
[9] Nikhil R Pal, Kuhu Pal, James M Keller, and James C Bezdek. A possibilistic fuzzy c-means clustering algorithm. IEEE Transactions on fuzzy systems, 13(4):517–530, 2005.
[10] T. Hitendra Sarma and P. Viswanath. Speeding-up the k-means clustering method: A prototype based approach. In Proc. 3rd Int. Conf. on Pattern Recognition and Machine Intelligence(PReMI)LNCS 5909, page 5661, Berlin Heidelberg, 2009. Springer-Verlag.
[11] T Hitendra Sarma, P Viswanath, and B Eswara Reddy. A hybrid approach to speed-up the k-means clustering method. International Journal of Machine Learning and Cybernetics, 4(2):107–117, 2013.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

[12] Sabu Sebastian and Ramakrishnan,T. V.(2011) Multi-fuzzy sets: an extension of fuzzy sets, Fuzzy Inf.Eng. 1, 35-43.
[13] Mei, Jian-Ping, et al. "Large scale document categorization with fuzzy clustering." *IEEE Transactions on Fuzzy Systems* (2016).
[14] Himmelspach, Ludmila, and Stefan Conrad. "A Possibilistic Multivariate Fuzzy c-Means Clustering Algorithm." International Conference on Scalable Uncertainty Management. Springer International Publishing, 2016.
[15] Mei, Jian-Ping, et al. "Large scale document categorization with fuzzy clustering." IEEE Transactions on Fuzzy Systems (2016).