E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

A NEW METHOD TO REMOVE DEPENDENCE OF FUZZY C-MEANS CLUSTERING TECHNIQUE ON RANDOM INITIALIZATION

Samarjit Das[1], Hemanta K. Baruah[2]
[1]*Department of Computer Science &IT,* [2]*Vice-Chancellor*
[1]*Cotton College, Assam, India*
[2] *Bodoland University, Assam, India*
[1] ssaimm@rediffmail.com, [2] hemanta_bh@yahoo.com

ABSTARCT:

Fuzzy clustering techniques deal the situations where there is a possibility of belonging a single object to more than one cluster. Although Fuzzy C-Means clustering technique of Bezdek is widely studied and applied, its performance is highly dependent on the randomly initialized membership values of the objects used for choosing the initial centroids. This paper proposes a modified method to remove the effect of random initialization from Fuzzy C-Means clustering technique and to improve the overall performance of it. In our proposed method we have used the algorithm of Yuan et al to determine the initial centroids. These initial centroids are then used in the conventional Fuzzy C- Means clustering technique of Bezdek to obtain the final clusters. We have tried to compare the performance of our proposed method with that of conventional Fuzzy C-means clustering technique of Bezdek by using Partition Coefficient and Clustering Entropy as validity indices.

Keywords: *Fuzzy C-Means clustering technique, Random initialization, Partition Coefficient, Clustering Entropy.*

1. INTRODUCTION

In conventional hard clustering techniques a large dataset is partitioned into some smaller clusters where an object either belongs completely to a particular cluster or does not belong to it at all. With the advent of the concept of Fuzzy Set Theory (FST) of Zadeh (1965) which particularly deals the situations pertaining to non-probabilistic uncertainty, the conventional hard clustering techniques have unlocked a new way of clustering, known as fuzzy clustering, where a single object may belong exactly to one cluster or partially to more than one cluster depending on the membership value of that object. Baruah (2011a, 2011b) has proved that the membership value of a fuzzy number can be expressed as a difference between the membership function and a reference function and therefore the fuzzy membership value and the fuzzy membership function for the complement of a fuzzy set are not the same. In the literature the Fuzzy C-Means (FCM) clustering technique of Bezdek (1981) has been found to be very popular among the research community. Derrig and Ostaszewski (1995) have applied the FCM of Bezdek (1981) in their research work where they have explained a method of pattern recognition for risk and claim classification. Das and Baruah (2013a) have shown the application of the FCM of Bezdek (1981) on vehicular pollution, through which they have discussed the importance of application of a fuzzy clustering technique on a dataset describing vehicular pollution, instead of a hard clustering technique. Das and Baruah (2013b) have applied the FCM of Bezdek (1981) and Gustafson and Kessel (GK) clustering technique of Gustafson and Kessel (1979) on the same dataset to make a comparison between these two clustering techniques and found that the overall performance of FCM is better than that of GK. Although it is evident in the literature that the FCM performs better as compared to other fuzzy clustering techniques, the performance of FCM is highly dependent on the randomly initialized membership values of the objects used for selecting the initial centroids. Yuan *et al.* (2004) proposed a systematic method for finding the initial centroids where there is no scope of randomness and therefore the centroids obtained by this method are

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

found to be consistent. In our proposed work we use these centroids thus obtained as the initial centroids in FCM of Bezdek (1981) to remove the effect of random initialization from FCM and also to improve the overall performance of it. Using Partition Coefficient (PC) and Clustering Entropy (CE) as validity indices (see Bezdek (1981) and Bensaid *et al.*(1996)) we have tried to make a comparison of the performances of these two clustering techniques.

In section-2 we provide the steps of the algorithms used in our present work. Through section-3 we describe our proposed model. The results and analysis of our present work have been placed in section-4. Finally we put the conclusion in section-5.

2. ALGORITHMS

The basic task of a clustering technique is to divide n patterns, where n is a natural number , represented by vectors in a p-dimensional Euclidean space, into c, 2≤ c <n , categorically homogeneous subsets which are called clusters. Let the data set be X= {x_1, x_2, ……., x_n }, where x_k ={ x_{k1}, x_{k2}, ……., x_{kp} }, k= 1,2,3,……..,n. Each x_k is called a feature vector and x_{kj} where j=1,2,…..p is the j^{th} feature of the k^{th} feature vector. A partition of the dataset X into clusters is described by the membership functions of the elements of the cluster. Let S_1, S_2,……..,S_c denote the clusters with corresponding membership functions μ_{S_1} , μ_{S_2} , …, μ_{S_c} . A c x n matrix containing the membership values of the objects in the clusters

$\tilde{U} = [\mu_{S_i}(x_k)]_{i=1,2,.....,c, k=1,2,.....,n}$ is a fuzzy c- partition if it satisfies the following conditions

$$\sum_{i=1}^{c} \mu_{s_i}(x_k) = 1 \quad \text{for each k= 1,2,….,n.} \tag{1}$$

$$0 \le \sum_{k=1}^{n} \mu_{s_i}(x_k) \le n \quad \text{for each i= 1,2,….,c.} \tag{2}$$

Condition (1) says that each feature vector x_k has its total membership value 1 divided among all clusters. Condition (2) states that the sum of membership degrees of feature vectors in a given cluster does not exceed the total number of feature vectors. In our proposed model we have used the algorithm of Yuan *et al*. (2004) as a preprocessor to the FCM algorithm of Bezdek (1981). In sections 2.1 and 2.2 we provide the steps of FCM algorithm of Bezdek (1981) and the algorithm of Yuan *et al*. (2004) respectively.

2.1. *FCM Algorithm of Bezdek*

Step 1: Choose the number of clusters, c, 2≤c<n, where n is the total number of feature vectors. Choose m, 1≤ m <α. Define the vector norm ‖ ‖(generally defined by the Euclidean distance) i.e.

$$\| x_k - v_i \| = \sqrt{\sum_{j=1}^{p} (x_{kj} - v_{ij})^2} \tag{3}$$

where x_{kj} is the j^{th} feature of the k^{th} feature vector, for k=1,2,……,n; j=1,2,….,p and v_{ij}, j-dimensional centre of the i^{th} cluster for i=1,2,……,c; j=1,2,….,p; n, p and c denote the total number of feature vector , no. of features in each feature vector and total number of clusters respectively.

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

Choose the initial fuzzy partition (by putting some random values)

$$U^{(0)} = [\mu_{s_i}{}^{(0)}(x_k)]_{1\le i\le c, 1\le k\le n}$$

Choose a parameter $\in$>0 (this will tell us when to stop the iteration). Set the iteration counting parameter l equal to 0.

Step 2: Calculate the fuzzy cluster centers $\{v_i^{(l)}\}_{i=1,2,.....,c}$ given by the following formula

$$v_i^{(l)} = \frac{\sum_{k=1}^{n} (\mu_{s_i}{}^{(l)}(x_k))^m x_k}{\sum_{k=1}^{n} (\mu_{s_i}{}^{(l)}(x_k))^m} \quad (4)$$

for i = 1, 2 , ….. c; k= 1, 2, , …..n.

Step 3: Calculate the new partition matrix (i.e. membership matrix)

$$U^{(l+1)} = [\mu_{s_i}{}^{(l+1)}(x_k)]_{1\le i\le c, 1\le k\le n} \quad ,$$

where

$$\mu_{s_i}{}^{(l+1)}(x_k) = \frac{1}{\sum_{j=1}^{c} \left(\frac{\| x_k - v_i^{(l)} \|}{\| x_k - v_j^{(l)} \|}\right)^{\frac{2}{m-1}}} \quad (5)$$

for i=1,2,……..,c and k=1,2,……..,n.

If $x_k = v_i^{(l)}$, formula (5) cannot be used. In this case the membership function is

$$\mu_{s_i}{}^{(l+1)}(x_k) = \{^{1 if k=i}_{0 if k\ne i, i=1,2,....,c}$$

Step 4: Calculate $\Delta = \| U^{(l+1)} - U^{(l)} \|$ (6)

If $\Delta > \in$, repeat steps 2, 3 and 4. Otherwise, stop at some iteration count l^*.

2.2. *Algorithm of Yuan et al.*

Step 1: Set m =1;

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

Step 2: Compute the distance between each data point and all other data points in the set X;
Step 3:Find the closet pair of data points from the set X and form a data point set A_m ($1 \le m \le c$, c is the number of clusters) which contains these two data points, delete these two data points from the set X;
Step 4: Find the data point in X that is the closet to the data point set A_m, add it to A_m and delete it from X;
Step 5: Repeat step 4 until the number of data points in A_m reaches 0.75*(n/c); (where .75 is a multiplication factor (MF))
Step 6: If m<c, then m = m+1, find another pair of data points from X between which the distance is shortest, form another data point set A_m and delete them from X, go to step 4;
Step 7: For each data point set A_m ($1 \le m \le c$) find the arithmetic mean of the vectors of data points in A_m, these means will be the initial centroids.

3. OUR PRESENT WORK

As the initial centroids in FCM of Bezdek (1981) are obtained based on the randomly initialized membership values of the objects, therefore the final clusters thus obtained are also not fixed. In other words there is significant variation in the performance of FCM clustering technique while executed different times. In the algorithm of Yuan *et al*. (2004) a systematic method is used to find the initial centroids where there is no random initialization. In our proposed model we use these initial centroids thus obtained as inputs of FCM of Bezdek (1981). In this way we have tried to remove the effect of random initialization from FCM clustering technique and also to improve its overall performance. We explain our proposed model with a flowchart (see Fig.1). We have applied both FCM clustering technique and our method ten (10) times each on the same dataset (see Table1) and tried to make a comparison of the performances of these two clustering techniques. We have used two validity measures of Bezdek (1981) and Bensaid *et al*. (1996) and the number of iterations to obtain the performances of these two clustering techniques. The mathematical formulae of these two validity measures have been given in the following.

(a) Partition Coefficient (PC): measures the overlapping between clusters.

$$PC(c) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^2$$

(b) Clustering Entropy (CE): measures the fuzziness of the cluster partition

$$CE(c) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\log(\mu_{ij})$$

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

The proposed model of our present work has been given in the following.

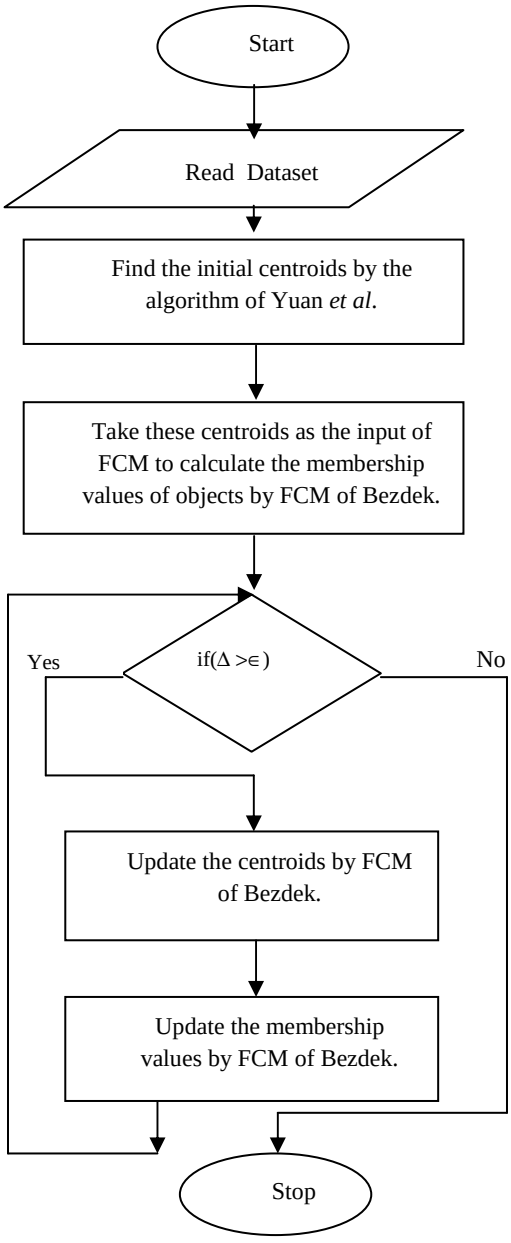

Figure 1. Flowchart of our proposed model.

The dataset of our present work consists of fifty(50) Feature Vectors (FV) each of which is of dimension three(03) namely Intelligent Quotient (IQ) , Achievement Motivation(AM) and Social Adjustment (SA). The numerical values of our dataset have been given in the following table.

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

Table 1. Data set of individual differences of fifty (50) feature vectors with dimension (feature) three (03).

FV	IQ	AM	SA	FV	IQ	AM	SA
1	91	18	55	**26**	110	18	55
2	85	16	40	**27**	100	16	40
3	120	19	74	**28**	100	18	75
4	90	18	75	**29**	70	14	30
5	92	17	74	**30**	105	17	55
6	82	17	55	**31**	79	14	35
7	95	19	75	**32**	80	15	34
8	89	18	74	**33**	125	20	75
9	96	19	75	**34**	100	19	75
10	90	17	55	**35**	125	19	85
11	97	16	54	**36**	80	18	60
12	125	21	74	**37**	85	18	70
13	100	19	75	**38**	145	25	90
14	90	17	54	**39**	80	18	74
15	100	18	84	**40**	92	17	55
16	95	19	75	**41**	120	18	70
17	130	23	85	**42**	145	30	80
18	130	19	75	**43**	95	18	50
19	90	17	55	**44**	80	16	36
20	91	17	56	**45**	90	17	55
21	140	22	82	**46**	115	23	84
22	92	18	75	**47**	100	18	80
23	101	18	55	**48**	80	14	35
24	85	16	54	**49**	105	19	75
25	97	19	54	**50**	120	21	74

4. RESULTS AND ANALYSIS

In this section we provide the results and analysis of our present work. Before making a comparison of the performance of FCM of Bezdek (1981) with that of our proposed model we have tried to optimize the performance level of our proposed model by taking the best choice of the multiplication factor (MF) (see step 5 of the algorithm of Yuan *et al*. (2004) in section 2.2). Fig. 2 shows that the performance of our proposed model is optimized when MF=.65.

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

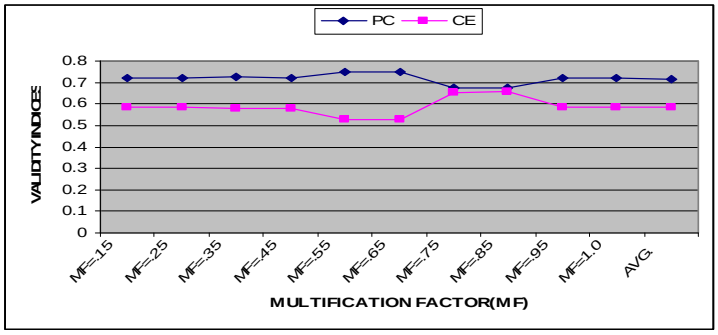

Figure 2. Performance levels of our proposed model with different values of MF.

In Fig. 3 we see that the value of the validity index PC of FCM varies significantly in ten (10) different executions. It is also seen in Fig. 3 that with the best choice of MF (when MF=.65) our proposed model results consistent and better performance (i.e. with higher values of PC) than FCM. A similar result is reflected in Fig.4 with the validity index CE. That is, our proposed model shows consistent and better performance (i.e. lower values of CE) in contrast to inconsistent performance of FCM in ten(10) different executions.

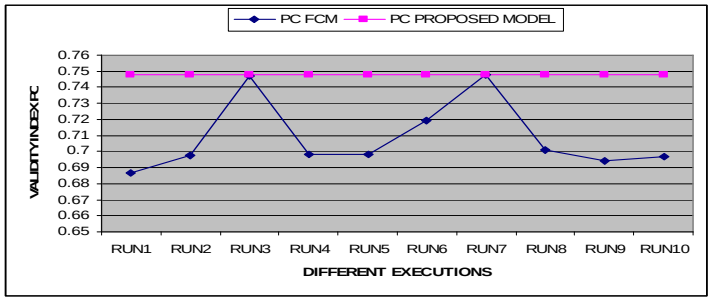

Figure 3. Measures of the validity index PC of FCM and that (optimized value only) of our proposed model in ten (10) different executions.

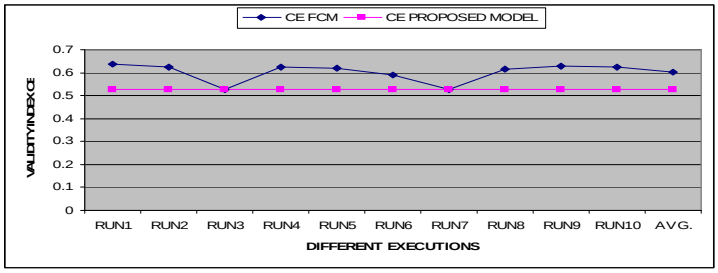

Figure 4. Measures of the validity index CE of FCM and that (optimized value only) of our proposed model in ten (10) different executions.

Fig. 5 shows that the average value of the validity index PC (for different values of MF) of our proposed model is more than that of FCM. This means that our model performs better than FCM even though the best

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

choice of MF in our proposed model is not taken. In Fig. 6 we see that the average performance of our proposed model is better (i.e. the value of CE is less) than that of FCM.

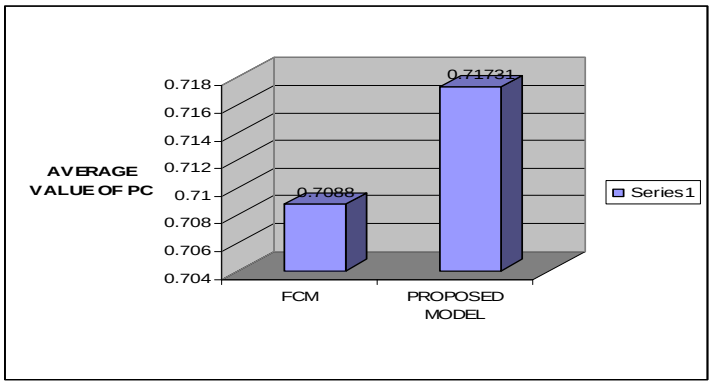

Figure 5. Average value of the validity index PC of FCM and that (for different values of MF) of our proposed model in ten (10) different executions.

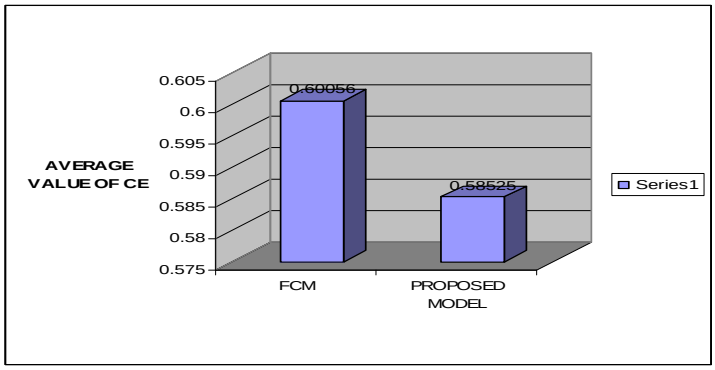

Figure 6. Average value of the validity index CE of FCM and that (for different values of MF) of our proposed model in ten (10) different executions.

In Fig. 7 we see that the average number of iterations (for different values of MF) of our proposed model is less than that of FCM in ten(10) different executions.

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

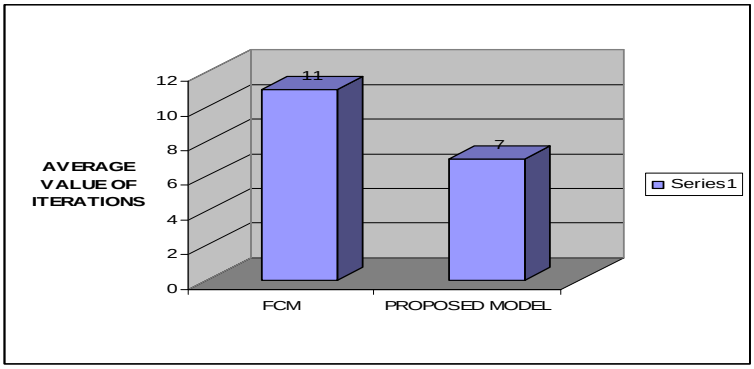

Figure 7. Average number of iterations of FCM and that (for different values of MF) of our proposed model in ten (10) different executions.

Thus we see in the results that with respect to the two validity indices (i.e. PC and CE) and the number of iterations our proposed model has a consistent and better performance.

5. CONCLUSIONS

Although FCM clustering technique is very popular among the research community, the major disadvantage of it is that its performance is very inconsistent due to the randomly initialized membership values of the feature vectors for selecting the initial centroids. Our proposed model which uses the algorithm of Yuan *et al*. as a preprocessor of FCM of Bezdek, can remove this inconsistency of FCM due to randomness by giving consistent and better performance. Although the average performance level of our proposed model is higher than that of FCM, it is advisable to optimize the performance level of our model with the best choice of the multiplication factor.

References

[1] Baruah, H.K.(2011a): Towards forming a field of fuzzy sets. International Journal of Energy, Information and Communications, **2**(1), pp. 16-20.
[2] Baruah, H.K.(2011b): The theory of fuzzy sets: beliefs and realities. International Journal of Energy, Information and Communications, **2**(2),pp. 1-22.
[3] Bensaid, A.M.; Hall, L.O.; Bezdek, J.C.(1996): Validity- guided (re) clustering with applications to image segmentation. IEEE Trans. on Fuzzy Object, **2**(2), pp.112-123.
[4] Bezdek, J.C.(1981). *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York.
[5] Das, S.; Baruah, H. K.(2013a): Application of Fuzzy C-Means Clustering Technique in Vehicular Pollution. Journal of Process Management – New Technologies,**1**(3), pp.96-107.
[6] Das, S.; Baruah, H. K.(2013b): A comparison of two fuzzy clustering techniques . Journal of Process Management – New Technologies,**1**(4), pp.1-15.
[7] Derrig, R. A.; Ostaszewski, K. M.(1995): Fuzzy techniques of pattern recognition in risk and claim classification. Journal of Risk and Insurance, **62**(3), pp.447-482.
[8] Gustafson, D.E.; Kessel ,W.C. (1979):Fuzzy clustering with a fuzzy covariance matrix, Proc. IEEE CDC, San Diego, CA, USA, pp.761–766.
[9] Yuan, F. ; Meng, Z.H.; Zhang, H.X.; Dong, C.R.(2004): A new algorithm to get the initial centroids, Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29.
[10] Zadeh, L. A.(1965): Fuzzy sets. Information and Control, **8**(3), pp. 338-353.