A New Preprocessor to Fuzzy c-Means Algorithm

Raveen S., P.S.V.S. Sai Prasad, and Raghavendra Rao Chillarige

School of Computer and Information Sciences
University of Hyderabad
Hyderabad, Telangana - 500046, India
raveenneevar@gmail.com, {saics,crrcs}@uohyd.ernet.in

Abstract. The fuzzy clustering scenario resulting from Fuzzy c-Means Algorithm (FCM) is highly sensitive to input parameters, number of clusters c and randomly initialized fuzzy membership matrix. Traditionally, the optimal fuzzy clustering scenario is arrived by exhaustive, repetitive invocations of FCM for different c values. In this paper, a new preprocessor based on Simplified Fuzzy Min-Max Neural Network (FMMNN) is proposed for FCM. The new preprocessor results in an estimation of the number of clusters in a given dataset and an initialization for fuzzy membership matrix. FCM algorithm with embeded preprocessor is named as FCMPre. Comparative experimental results of FCMPre with FCM based on benchmark datasets, empirically established that FCMPre discovers optimal (or near optimal) fuzzy clustering scenarios without exhaustive invocations of FCM along with obtaining significant computational gains.

Keywords: Soft Clustering, Fuzzy Clustering, Fuzzy c-Means Algorithm, Fuzzy Min-Max Neural Network.

1 Introduction

Clustering is a most important technique in data mining, which is used to discover groups and identify interesing distributions and patterns in the underlying data. Clustering divides data into different groups such that the data items belonging to the same groups are more similar than data items belonging to different groups. Clustering is widely used in various of applications. Some of the areas of application are business, biology, spatial data analysis, web mining etc. Clustering schema can be effectively used for data reduction, hypothesis generation, hypothesis testing and prediction based on groups. The clustering schema can be classified into different types based on different criteria. Based on the criteria of presence or absence of overlapping between clusters, clustering schema is classified into soft clustering, hard clustering respectively.

Soft clustering gives apt knowledge representation of real world scenarios as a data item belongs to more than one cluster (group) in several real world applications. Fuzzy clustering is an important soft clustering methodology based on the principles of fuzzy sets. In fuzzy clustering, a data item's belonging to a cluster is defined by a fuzzy membership function.

M.N. Murty et al. (Eds.): MIWAI 2014, LNAI 8875, pp. 124–135, 2014.

Fuzzy c-Means algorithm (FCM), is an important and widely used fuzzy clustering algorithm, was proposed by J. C. Bezdek in 1984 [16]. FCM is highly sensitive to input parameters, number of clusters c and randomly initialized fuzzy membership matrix. Hence in practice, repeated experiments need to be conducted using FCM for varying number of clusters, for different random initialization of fuzzy membership matrix to arrive at an optimal fuzzy clustering scenario. In this work, a novel preprocessor for FCM is proposed based on Simplified Fuzzy Min-Max Neural Network Algorithm (SFMMNN). The proposed preprocessor results in an estimation of the number of clusters c in a given dataset and an initialization of the fuzzy membership matrix. FCM algorithm with embedded preprocessor is named as FCMPre. The performance of FCMPre is analyzed based on comparative experimental analysis with FCM conducted on benchmark datasets.

Section 2 details the literature review of applications of FCM and improvements to FCM. FCM algorithm is described in Section 3. The proposed preprocessor to FCM and FCMPre algorithm is given in Section 4. Comparative experimental results and analysis is provided in Section 5.

2 Related Work

After its invention in 1984, FCM is applied in different application domains such as image processing, biology, bussiness etc. Several improvements to FCM are proposed in the literature. A review of applications and improvements to FCM is presented in this section.

W. E. Philips proposed application of Fuzzy c-Means segmentation technique for tissue differentiation in {MR} images of a hemorrhagic glioblastoma multiforme [8]. Demble proposed Fuzzy c-Means method for clustering microarray data [10]. A.B. Goktepe proposed Soil clustering by Fuzzy c-Means algorithm [6]. Keh-Shih Chuang proposed Fuzzy c-Means clustering with spatial information for image segmentation [7]. Li Xiang proposed an application of Fuzzy c-Means Clustering in Data Analysis of Metabolomics [11].

There are several approaches that have been proposed to improve the performance and efficiency of FCM. Hathaway proposed extension of Fuzzy c-Means Algorithm for Incomplete Datasets [9] containing missing values. Tai Wai Cheng proposed Fast Fuzzy clustering Algorithm (FFCM) [14] by incorporating a new procedure for initialization of cluster centroids replacing random initialization in FCM. Eschirich proposed Fast accurate fuzzy clustering through data reduction [12] by aggregation of similar data items into a single unit. Xizhao Wang improved Fuzzy c-Means clustering based on feature-weight learning [13]. This method uses proper learning techniques to find relevant features and avoids those features causing ambiguity. Al-Zoubi proposed a fast Fuzzy Clustering Algorithm [15]. The approach involves improving the performance of Fuzzy c-Means algorithm by eliminating unnecessary distance calculation. This can be done as comparing membership value with a threshold value and if it is less than the threshold value, the distance calculation is eliminated. The choice of number of clusters (c) for FCM is traditionally determined by optimizing validity index on results of several experiments conducted with diverse c values. Nguyen proposed a dynamic approach, in which adjustement to initial c value takes place along with cluster formation [18].

In this paper, an alternative approach for determining number of clusters and initial centroid vectors is proposed by using a preprocessor. The relevance of proposed preprocessor is empirically established by comparison with traditional approach. The next section describes FCM in detail.

3 Fuzzy c Means

Fuzzy clustering is proposed by Dunn in 1974 [3] and is extended by J.C. Bezdek in 1984 [16] as Fuzzy c-means Algorithm. The basic idea of fuzzy clustering is to partition given dataset $X = \{X_1, X_2, ...X_n\}$ into c number of fuzzy partitions. The fuzzy membership matrix U represents the resulting fuzzy partition. $U = [u_{ij}]_{N\times c}$ is a matrix with dimensions $N \times c$, where u_{ij} represents membership value of i^{th} data object into j^{th} cluster. The objective function for finding optimal clustering schema for FCM is based on the least mean square method. The objective function is described as

$$J_m(U, v) = \Sigma_{k=1}^{N} \Sigma_{i=1}^{c} (u_{ik})^m \|X_k - v_i\|_A^2 \tag{1}$$

The variables used in objective function are
dataset $X = \{X_1, X_2, X_3...., X_N\}$ where $\forall X_i \in R^p$ (p : number of dimensions),
c : number of clusters; where $2 \le c \le N$,
m : weighting exponent which is called as fuzzifier ($1 < m < \infty$,)
U : fuzzy membership matrix (Fuzzy c-partition)
$v = (v_1, v_2, v_3, ...v_c)$: cluster centroids,
$v_i = (v_{i1}, v_{i2}, ..., v_{ip})$: i^{th} cluster centroid,
$\|\|_A$: induced A-norm on R^p

The FCM algorithm minimizes U and v in each iteration such that $J_m(U, v)$ is minimized. It uses an alternate optimization technique such that in each iteration of FCM, v is changed keeping U as fixed and then U is changed keeping v as fixed. The FCM algorithm is described in [16] and is given here for completeness.

Fuzzy c-Means Algorithm
Input:- Set of data vectors : $X = \{X_1, X_2, ...X_N\}$, c : number of clusters, m : fuzzifier, $\|\|_A$: distance norm.
Output:-U : Fuzzy c-partition of X.
Method
(1) Initialize membership matrix $U^{(0)}$.
(2) Calculate centroid vector $v^{(i)}, k = 1, 2, ..., c$ using the equation.

$$v^{(i)} = \frac{\Sigma_{k=1}^{N} (u_{ik})^m x_k}{\Sigma_{k=1}^{N} (u_{ik})^m} \tag{2}$$

(3) Calculate updated fuzzy membership matrix $U^q = [u^q]$ using the equation.

$$u_{ik}^q = (\Sigma_{j=1}^c (\frac{d_{ik}}{d_{jk}})^{\frac{2}{(m-1)}})^{-1} \tag{3}$$

where d_{ik} represents the distance measure of k^{th} object to the centre vector of i^{th} cluster.
(4) Compare $U^{(q)}$ to $U^{(q-1)}$. $if \|U^{(q)} - U^{(q-1)}\| < \epsilon$,stop. otherwise set $q = q+1$ and continue to step (2) .

There are different validation techniques available in the literature [16] to evaluate the validity of the fuzzy clustering schema. Popular validity indices are Partition Coefficient (PC), Partition Entropy (PE). Both PC and PE suffer from the monotonic evolution tendency with c. The validity index MPC (Modification of the PC) [2] reduces the monotonic tendency. Hence MPC is used in the present work for validation of fuzzy clustering.

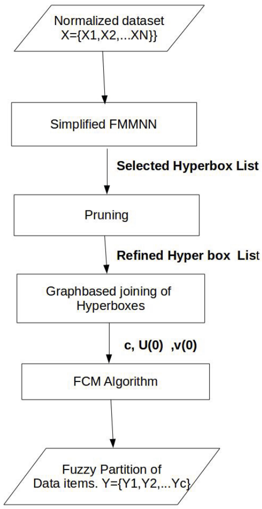

Fig. 1. Flow chart of FCMPre Algorithm

4 New Preprocessor to Fuzzy c-Means Algorithm

Fuzzy c-mean algorithm requires the number of clusters c as input parameter. Given a validation index V, the problem of obtaining optimal c-value is computationally highly intensive. The problem can be formally specified as follows. Assuming optimization requires maximization of V and N being the size of the dataset, the optimum c value is c^* where

$$V_{c^*} = max_{2 \leq c \leq (N-1)} V_c \tag{4}$$

The problem is further complicated owing to variability in behaviour of FCM algorithm due to random initialization of the fuzzy membership matrix. Usually in practice an experiment is repeated M times for a particular c value, for arriving at optimal fuzzy clustering. Hence for obtaining V^* in practice, requires $(N-2)*M$ invocations of FCM. In this section, a novel preprocessing technique is proposed for arriving at an estimation of c value and obtaining initial centroids which leads to a deterministic procedure for obtaining an initial fuzzy membership matrix so that an optimal or near optimal clustering configuration can be obtained without performing an exhaustive invocations of FCM.

The proposed approach involves applying a fast clustering algorithm for estimation of clusters followed by a graph based approach for merging of clusters (a decision guided by intended validation index). The resulting clusters determine the estimated c value and initial centroids which are adapted to FCM and if required iterations of FCM are continued till convergence.

The literature survey conducted has identified that Fuzzy Min-Max neural network (FMMNN), proposed by Simpson [17], is a fast algorithm for arriving at an estimation of clusters. In this work, a modification of FMMNN is used for the initial estimation of clusters in the proposed approach which is described in Section 4.1.

4.1 Fuzzy Min-Max Neural Network

Fuzzy min-max Neural Network for clustering is a single pass clustering algorithm proposed by Simpson [17]. FMMNN is a fast clustering algorithm for fuzzy clustering, which group the set of data vectors into different hyperboxes. Hyperboxes are the p-dimensional unit cubes which is associated with a membership function which describes to what membership degree the data vector belongs to a particular hyperbox fuzzy set. The membership degree ranges from 0 to 1 in which '0' means no membership and '1' means full membership. The FMMNN for clustering is discussed in Section 4.2.

4.2 Fuzzy Min-Max Clustering Algorithm

The first step of FMMNN is the initialization, which includes the formation of point hyperboxes (hyperbox with input data vector as min and max points) into an uncommitted set (UC) for all data vectors and a point hyperbox of the first data vector is introduced into a committed set (CS). After initialization, with each input pattern, a three step process (Expansion, Overlap test and Contraction) is applied.

In the literature, a simplified FMMNN for classification was proposed which uses only the initialization and expansion step of the original algorithm [4]. This results in overlapping hyperboxes in simplified FMMNN. Since the proposed preprocessor for FCM algorithm is meant for drawing an information about soft clustering and obtaining an initial fuzzy membership matrix for FCM, the overlapping of hyperboxes is allowed. Hence the concept of simplified FMMNN used for classification in [4] is adapted for proposed preprocessing strategy.

The Simplified Fuzzy Min Max Neural Network algorithm (SFMMNN) for clustering is given below.

Algorithm Simplified FMMNN (SFMMNN)
Input:- Set of data vectors, $X = \{X_1, X_2, ..., X_N\}$
Output:- Committed Set of Hyperboxes, $CS = \{B_1, ...B_l\}$
Method

1. Initialize θ, γ . Initialize uncommitted set (UC) with point hyperboxes for each data vector $X_h \in X$. Remove the first hyperbox from UC and include in committed set CS .
2. $\forall X_h \in \{X_2, ...X_N\}$
   (a) Select $B_j \in CS$ which satisfy equation
   $\Sigma_{i=1}^{p}(max(w_{ji}, x_{hi}) - min(x_{hi}, v_{ji})) \leq p\theta$

   (b) If B_j exists
      i. Update the max boundaries of B_j, which satisfies the expansion criteria.
      $w_{ji}^{new} = max(w_{ji}^{old}, x_{hi})$
      ii. Update the min boundaries of B_j as
      $v_{ji}^{new} = min(v_{ji}^{old}, x_{hi})$
   else
      i. Remove $B_h \in UC$ and include in CS .
3. Return CS

4.3 Fuzzy Min-Max Preprocesser for Fuzzy c-Means

The first step of the preprocessor is the application of simplified FMMNN. The resulting selected hyperboxes may be very high in number and some may contain very few data vetors and can mislead the recommendation of c value. A post pruning process is done to remove the irrelevant hyperboxes from the list, which does not contain at least t% of total input data vectors.

SFMMNN under the constraint of expansion parameter θ may represent a large cluster with multiple hyperboxes. In order to obtain the appropriate recommendation of c value, a graph based combining process is proposed. The graph based combining process checks whether the individual hyperboxes can be combined to give better result. The union of small hyperboxes into larger one will be based on the comparison of validity index (V_m) obtained for the current set of hyperboxes with an intended set of hyperboxes resulting from the union of two hyperboxes under consideration. The procedure for graph based approach for combining hyperboxes is given below.

Graph based approach for combining Hyperboxes

- A graph $G = (V, E)$ constructed from refined list of hyperboxes considering each hyperbox as nodes and placing an edge between overlapping hyperboxes .
- Graph G is represented as adjacency list.
- $\forall v \in V$
  - $\forall u \in$ adjacencyList(v)
    1. Construct two fuzzy clustering scenarios.
       * **Scenario1:-** Fuzzy clustering resulting from the current set of hyperboxes.
       * **Scenario2:-** Fuzzy clustering resulting from collapsing u and v into single hyperbox along with remaining hyperboxes.
    2. Compute Validity index (V_m) for Scenario1 and Scenario2.
    3. if $V_m(Scenario2) \quad > \quad V_m(Scenario1)$
       * If true, then Collapse u and v into new hyperbox B_{new} and update the list.

After obtaining a selected list of hyperboxes using graph based combining, cluster centroids are calculated for these selected hyperboxes, based on the data vectors assigned to each hyperbox. The number of hyperboxes determines the recommended c value and their centroids form the initial centroid vectors of FCM algorithm. Based on centroid vectors, initial fuzzy membership matrix is computed. The execution of FCM is continued based on the newly obtained membership matrix.

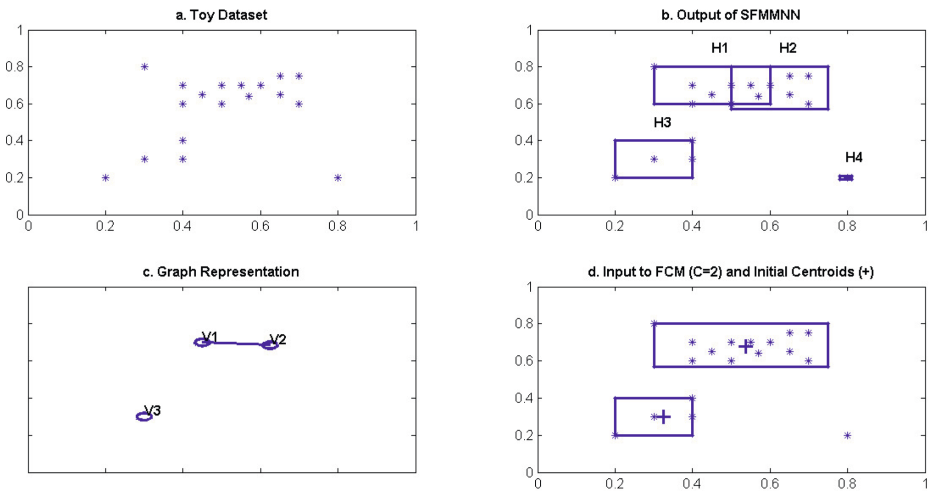

Fig. 2. Illustration of Preprocessor to FCM

Figure 2 illustrates the steps involved in the proposed preprocessor for FCM. A toy dataset with two features as given in Fig. 2-a is used for illustration. Applying SFMMNN, resulted in four hyperboxes (see Fig. 2-b). During pruning step hyperbox H4 is dropped as it contains very few objects ($< t\%$ where t = 5). Figure 2-c depicts the graph constructed after pruning phase. Edge is included between vertices V1, V2 as the corresponding hyperboxes H1, H2 are overlapping. During graph based approach for combining, V1 and V2 got collapsed into a single vertex. In the resulting scenario of two hyperboxes, the centroid vectors are computed as an average of objects falling absolutely into hyperboxes. Hence the proposed preprocessor resulted in recommending number of clusters $c = 2$ and initial centroid vectors as depicted in Fig. 2-d.

FCM algorithm with inclusion of proposed preprocessing step is named as FCMPre algorithm. The flow chart of FCMPre algorithm is given in Fig. 1

5 Experiments and Results

5.1 Experimental Design

Code for FCM and FCMPre are developed in Java environments. Experiments are conducted on benchmark datasets from UCI Machine Learning Repository. The objective of experimentation is for the empirical assesment of relevance of proposed preprocessor by comparing the results of FCMPre with optimal results obtained from FCM using exhaustive approach described in Section 4.

Table 1. Description of datasets used for Experimental study

Name	Number of data vectors	Dimension
Aggregation	788	2
Ecoli	336	7
Glass	214	7
Iris	150	4
Yeast	1484	8

The problem of obtaining optimal c^* as described in Section 4 using FCM needs applying FCM for a repeated number of times, for varying values of c from 2 to $N-1$. In our experimental result FCM is applied repeatedly for 10 times for varying c values from 2 to 25. The m value is fixed at 2.0 as it is the recommended value from literature [1]. The number of iterations and time taken to converge is computed in each invocation of FCM. The performance of FCMPre and FCM are validated on the basis of MPC Coefficient as the validation measure.

FCMPre algorithm requires SFMMNN algorithm as a preprocessing algorithm. SFMMNN requires specification of θ and γ. The recommended value of γ is 4.0 [5] and the same is used in experiments. The behaviour of SFMMNN is greatly influenced by value of θ. It is expected that different values of θ can result in different number of selected hyperboxes, which after pruning and graph based combining hyperboxes may result in different initial configuration (number of centers, initial centroid vectors) for FCM invocation as part of FCMPre algorithm. For a particular θ, γ values, repeated invocations of SFMMNN result in same initial configuration for FCM making repeated execution of FCMPre for a particular θ, γ not required. Hence arriving at optimal clustering configuration using FCMPre is taken as the best clustering configuration resulted from applying FCMPre for $\theta = 0.2, 0.3, 0.4, 0.5$.

From now onwards "FCM experiment" represents applying FCM repeatedly for 10 times for c values from 2 to 25. "FCMPre experiment" represents applying FCMPre for θ values 0.2, 0.3, 0.4, 0.5.

5.2 Results

In FCM experiments, the maximum limit for c is fixed at 25, because all these datasets contain at most 10 clusters as per the information given from UCI Machine Learning Repository. The maximum value of c for yeast dataset is fixed at 23 due to its behaviour of taking too many iterations to converge for high values of c and the validity measures obtained are inferior. For each dataset, for repeated experiments that are done for different values of c, the best validity measure obtained is taken into consideration for further analysis. In case, if two executions resulted in the same value of MPC Coefficient, the one with less number of iterations is taken as best configuration for that c value.

The FCMPre is executed on these five datasets for different values of θ. The range of θ varies from 0.2 to 0.5.

5.3 Analysis

The experimental results reestablished the sensitivity of FCM for different values of c, for different random initialization of fuzzy membership matrix even for the same c.

The primary objective of FCMPre is to obtain optimal fuzzy clustering configuration without conducting exhaustive FCM experiments. Figure 3 and Fig. 4 depict the MPC validation index obtained for FCM (best value for each c value), FCMPre ($\theta = 0.2, 0.3, 0.4, 0.5$) for e-coli and yeast dataset respectively. From the experimental results obtained, the primary inference is that except for yeast dataset, FCMPre experiment could achieve the optimal clustering configuration obtained by exhaustive FCM experiments. In Yeast dataset, FCMPre could obtain fourth best cluster configuration whose MPC value is close to the best MPC value obtained in FCM experiments.

The salient comparative results of FCM and FCMPre are summarized in Table 2. Table 2 reports computational time for " FCM experiment", "FCMPre experiment", the best θ value obtained by FCMPre experiment, best MPC values obtained by FCM, FCMPre experiments, best c values recommended by FCM ans FCMPre experiments. Based on Table 2, FCMPre resulted in obtaining the optimal configuration (best c , best MPC) as FCM for Aggregation, Ecoli, Glass, Iris datasets. A near optimal configuration is obtained by FCMPre for

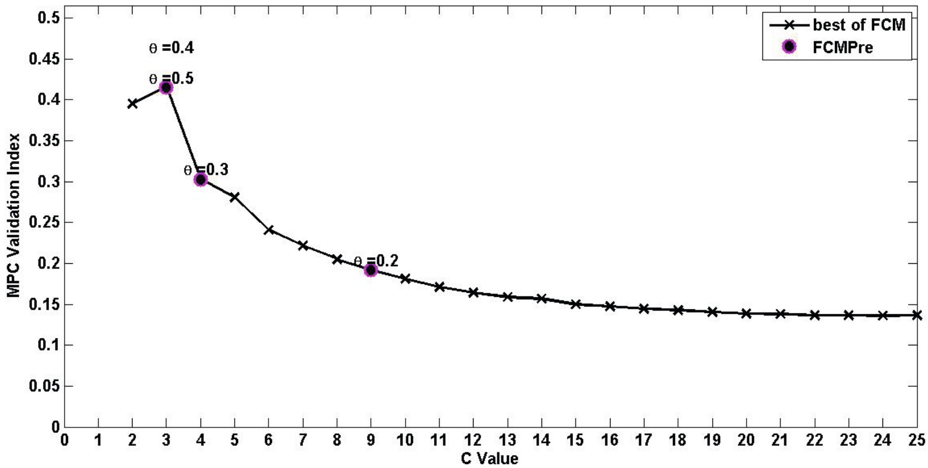

Fig. 3. Comparison of MPC validity measure for ecoli dataset

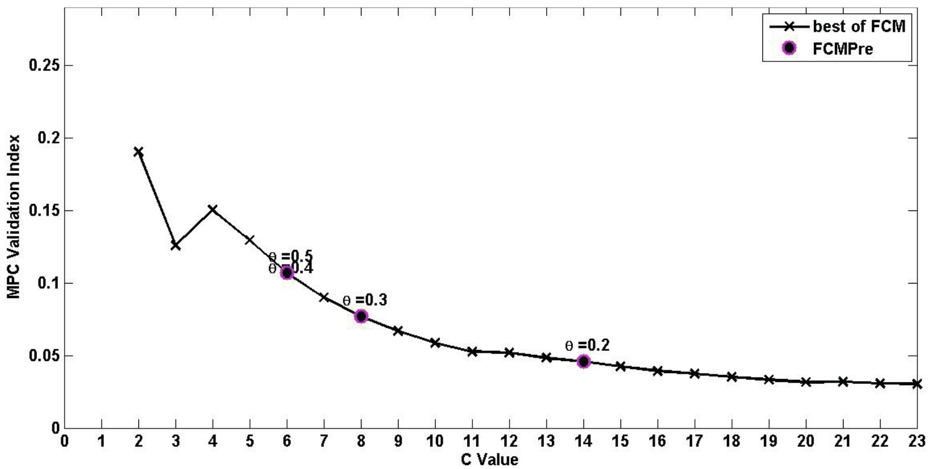

Fig. 4. Comparison of MPC validity measure for Yeast dataset

yeast dataset. These significant results are achieved by FCMPre with 99.5% and above computational gain over FCM.

Table 2. Salient comparative results of FCM and FCMPre experiments

Name	CT of FCM (sec)	CT of FCMPre (sec)	Best θ FCMPre	Best MPC (FCM)	Best MPC (FCMPre)	Best C FCM	Best C FCMPre
Aggregation	7057.07	2.89	0.4	0.5934	0.5934	4	4
Ecoli	2209.45	5.91	0.5/0.4	0.4159	0.4159	3	3
Glass	1121.1	0.377	0.3/0.4	0.6653	0.6653	2	2
Iris	263.03	0.392	0.5	0.7199	0.7199	2	2
Yeast	73756.6	358.9	0.4/0.5	0.1905	0.1073	2	6

6 Conclusion

In this paper, algorithm FCMPre is proposed for arriving at optimal or near optimal fuzzy clustering using the FCM algorithm avoiding exhaustive experiments for varying sizes of c. FCMPre experiment is computationally effective and obtained huge computational gain over FCM experiments. FCMPre experiments resulted in optimal or near optimal fuzzy clustering scenario compared to FCM experiments. The experimental results empirically validate the utility of FCMPre for obtaining reliable fuzzy clustering scenario.

References

1. Pal, N.R., Bezdek, J.C.: On Cluster Validity for the Fuzzy C-means Model. Trans. Fuz. Sys. 3, 370–379 (1995)
2. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. Fuzzy Sets and Systems 158, 2095–2117 (2007)
3. Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Cybernetics and Systems 3, 32–57 (1973)
4. Ma, D., Liu, J., Wang, Z.: The Pattern Classification Based on Fuzzy Min-max Neural Network with New Algorithm. In: Wang, J., Yen, G.G., Polycarpou, M.M. (eds.) ISNN 2012, Part II. LNCS, vol. 7368, pp. 1–9. Springer, Heidelberg (2012)
5. Simpson, P.K.: Fuzzy min-max neural networks. In: 1991 IEEE International Joint Conference on Neural Networks, vol. 2, pp. 1658–1659 (1991)
6. Goktepe, A.B., Altun, S., Sezer, A.: Soil clustering by Fuzzy c-Means algorithm 36, 691–697 (2005)
7. Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J., Chen, T.-J.: Fuzzy c-means clustering with spatial information for image segmentation. Computerized Medical Imaging and Graphics 30, 9–15 (2006)
8. Phillips II, W.E., Velthuizen, R.P., Phuphanich, S., Hall, L.O., Clarke, L.P., Silbiger, M.L.: Application of Fuzzy c-Means segmentation technique for tissue differentiation in MR images of a hemorrhagic glioblastoma multiforme. Magnetic Resonance Imaging 13, 277–290 (1995)
9. Hathaway, R.J., Bezdek, J.C.: Fuzzy C-means Clustering of Incomplete Data. Trans. Sys. Man Cyber. Part B 31, 735–744 (2001)
10. Dembl, D., Kastner, P.: Fuzzy C-means method for clustering microarray data. Bioinformatics 19, 973–980 (2003)

11. Li, X., Lu, X., Tian, J., Gao, P., Kong, H., Xu, G.: Application of Fuzzy c-Means Clustering in Data Analysis of Metabolomics. Analytical Chemistry 81, 4468–4475 (2009)
12. Eschrich, S., Ke, J., Hall, L.O., Goldgof, D.B.: Fast accurate fuzzy clustering through data reduction. IEEE Transactions on Fuzzy Systems 11, 262–270 (2003)
13. Wang, X., Wang, Y., Wang, L.: Improving Fuzzy c-Means clustering based on feature-weight learning. Pattern Recognition Letters 25, 1123–1132 (2004)
14. Cheng, T.W., Goldgof, D.B., Hall, L.O.: Fast fuzzy clustering. Fuzzy Sets and Systems 93, 49–56 (1998)
15. Al-Zoubi, M.B., Hudaib, A., Al-Shboul, B.: A Fast Fuzzy Clustering Algorithm. In: Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, vol. 3, pp. 28–32 (2007)
16. Bezdek, J., Ehrlich, R., Full, W.: FCM: The Fuzzy c-Means clustering algorithm. Computers & Geosciences 10, 191–203 (1984)
17. Simpson, P.K.: Fuzzy min-max neural networks - Part 2: Clustering. IEEE Transactions on Fuzzy Systems 1 (1993)
18. Nguyen, D.T., Doan, H.: An Approach to Determine the Number of Clusters for Clustering Algorithms. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part I. LNCS (LNAI), vol. 7653, pp. 485–494. Springer, Heidelberg (2012)