A Robust Clustering Approach Based on KNN and Modified C-Means Algorithm

Amir Aliabadian

Abstract: Cluster analysis is used for clustering a data set into groups of similar individuals. It is an approach towards to unsupervised learning and is one of the major techniques in pattern recognition.FCM algorithm needs the number of classes and initial values of center for each cluster. These values are determined randomly, so it may cause target function converges to several local center. so many iterative stages are needed, until FCM can reach to global center for each cluster. In this paper, we suggest robust hybrid algorithm in which, we have real unsupervised learning algorithm, no need to initial center value and the number of clusters. The First layer in this algorithm finds initial clustering center by K-nearest neighbor (K-NN) rules based on unsupervised learning approach. In the second layer, we applied FCM only one time for having optimal clustering. It is done by means of Fuzzy clustering validation criterion, unlike FCM that needs iterative process. We applied new algorithm to several set of standard databases (IRIS).results show that this algorithm is more accurate than FCM both in estimation of optimal number of clusters and correctness of devotion of data to their real clusters.

Key word: Cluster analysis . FCM algorithm . K-nearest neighbor . target function.

I. INTRODUCTION

According to the features of fuzzy theory there is no exact boundary between clusters .the membership degree of each sample is shown by a number between 0 and 1.in fuzzy clustering, each sample belongs to several clusters with different membership degrees. Several parameters in the FCM algorithm play a role in convergence of the classes` members into their centers. This issue may lead the algorithm converges to a wrong point. To overcome this problem various algorithms have been proposed for clustering, that most of them are developed editions out of primary FCM method. These methods are based on optimizing a specific target function. Optimizing process involves in finding a local minimum indicating the optimum point for the clusters` center. Since this process may lead to several local minima, the global minimum should be looked for, by running the process with different initial points. As discussed above, most of the clustering algorithms are dependent on the initial value of some parameters such as the number of clusters, centre of clusters and etc. Which users should specify. This issues the main drawback of these methods. In order to overcome this problem the proposed algorithm in this paper offers an approach by which the fuzzy clustering can be done without any assumption on the initial values. This method is a two-layer algorithm of KNN-FCM that prevents from so many iteration stages. In the first layer we do the initial clustering by K-nearest neighbor (K-NN) rules based on unsupervised learning approach.

Manuscript received April, 2014.

Amir Aliabadian,Department of Faculty Member of Electrical and Computer engineering Department, Shomal University,Iran.

In the second layer, we applied only one FCM iteration for having optimal clustering. We applied new algorithm to six sets of standard databases. Results show that this algorithm is more accurate than FCM in clustering.

Hard Clustering: Suppose X={x1, x2…xn} is a finite set composed of n data vectors that belongs to space RPand their corresponding classes are not specified. A hard clustering from X are gained by classifying these samples into C (C>1) separate clusters. For xk∈ X, employing the following function does classifying process:

$$u_i : X \to \{0\ ,\ 1\} \quad , \quad x_k \to u_i(x_k) = u_{ik} =$$

$$\begin{cases} 1 & x_k \in i\text{th cluster} \\ 0 & \text{otherwise} \end{cases} \quad , \quad 1 \le i \le c \quad , \quad 1 \le k \le n \text{.....(1)}$$

This function attributes membership degree of each sample and specifies those samples belonging to each class. Therefore, hard clustering algorithms are not appropriate for incomplete, obscure information. In next section the fuzzy clustering and it's features will be explained.

Fuzzy Clustering: A fuzzy clustering in C partitions from X is defined by the membership matrix $U=[u_{ik}]_{c.n}$in which $u_{ik} = u_i(x_k)$ is the membership degree of x_k in the i^{th} cluster (u_i). Also U elements have the following features:

$$\begin{cases} u_{ik} \in [0,1] \quad , \quad \forall i,k \\ \sum_i u_{ik} = 1 \quad , \quad \forall k \\ 0 < \sum_i u_{ik} < n \quad , \quad \forall i \end{cases} \qquad (2)$$

In the next section one of the common algorithms for clustering called FCM and its validity will be discussed.

Fuzzy C-means Clustering and Its Validity: FCM algorithm is the most common fuzzy algorithm for analyzing clusters and, which is based on the optimization of the following target function:

$$\min[Jm(U,V) = \sum_i \sum_k u_{ik}^m d^2(x_k, V_i)] \qquad (3)$$

Where $U = [u_{ik}] \in R^{cn}$ is a fuzzy cluster consisted of C parts from n data elements obtained from data set $X = \{x_1,\ldots,x_n\} \in R^{cn}$ and the centers for C fuzzy clusters are defined as $V = (V_1,\ldots V_c) \in R_{cn}$. The parameter m>1 is a measure of the fuzziness degree. Here if m=1 the fuzzy algorithm will be transformed into HCM (hard clustering method).

The convergence of FCM algorithm in finding samples with C clusters (X(u*)) and agents of cluster centers (V*) has been well proven [1]. By solving the above optimization problem, the necessary conditions for minimizing the target function $J_m(u,v)$ will be obtained as below:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (d(x_k, V_i) / d(x_k, V_j))^{2/(m-1)}}, \quad (4)$$

$$V_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \quad (5)$$

Here, $d^2(x_k, V_i) = \| x_k - V_i \|^2$ is the squared Euclidean distance from x_k to the center (V_i) of the clusters. In the end, the optimal solution for (U*, V*) will be obtained by iterating over the above equations.

The function $J_m(u,v)$ might cont to undesirable points which results in different clustering. In order to avoid this issue, the FCM algorithm must be run for more than once and for different arbitrary starting center points and different values of m, C and then the average of the results can be used. Therefore performance of FCM deeply depends on the initial values of these parameters.

THE CREDIT OF FUZZY CLUSTERING: At last, C fuzzy partitions will be obtained by the FCM algorithm, which will determine the structure regulating the set of analyzed data. In each cluster, items with higher level of membership have the most similarity. For measuring the credit and the value of the performed clustering, we will use a common technique, called "the credit of clustering". Various functions are available for measuring this credit. These functions assign a number to each clustering method, which indicates the ability of each method for clustering. The common functions for the credit of clustering are clustering entropy (H), common factor of clustering (PC), the known function (UDF) and the criteria for clusters compaction and separation (CS). It has been demonstrated that all the above measures have almost similar behaviors in evaluating the fuzzy clustering [2]. In this paper, the method of standard PC function has been used:

$$PC(U, C) = \frac{\sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^2}{n} \quad (6)$$

If the samples of data set X become completely separated in the cluster ($u_{ik} \rightarrow 1$), then we will not have any common member in the clusters. In this case, PC will be close to 1 and the method will be close to hard clustering. On the other hand, the worst clustering happens in the case where we face an unclear condition, and every sample with the identical level of membership belongs to all clusters ($PC \rightarrow 1/c$, $u_{ik} \rightarrow 1/c$). Therefore, it is clear that the best clustering happens by maximizing the PC for $C=2 \ldots C_{max}$. In our new algorithm, these characteristics of credit functions are utilized for finding the optimized value for the number of clusters, hence knowing additional information is not necessary for the user.

THE PROPOSED HYBRID ALGORITHM: FCM algorithm based on initial values converges to the point at which the Jm function is minimum. But in practice the position of clusters' centers is not known, so that choosing different values for starting points may lead to different local minima. In our new algorithm, we have 2 layers. In the first layer in an unsupervised clustering the initial centers of the clusters are obtained by using KNN rule. This rule devotes each sample to the cluster that have the most neighbors out of K nearest neighbors with it. Therefore, the first layer of algorithm will deal with partitioning the space X into C parts in which in every part the samples are similar from the Euclidean distance perspective. This grouping will be obtained by using the first part of KNN algorithm. For $1 \le i \le C$, Ei will formulate every C: (y_1, K – NN of y_i,G_i) with a set of K nearest neighbors. This relation means that for every sample of Yi, as defined above, the Ei set is relevant to K nearest neighbors. Also G_i, the center of i_{th} cluster, is defined as below:

$$G_i = \frac{\sum_{x_k \in E_i} x_k}{K+1} \quad (7)$$

While running the algorithm, if i=1, then y_1 ($y_1 \in E_1$) will be the furthest sample from the general center (G_0) relative to all samples of $x_k \in X$. Here G_0 defined as below:

$$G_0 = \frac{\sum_{x_k \in X} x_k}{n} \quad (8)$$

If $2 \le i \le c$, then sample y_i will create E_i as below:

$$y_i \in E_i$$
$$1 \le \alpha \le i-1 \ , \ y_i \notin E_\alpha \quad (9)$$
$$y_i \neq KNN \text{ of } y_\alpha$$

Yi is the furthest sample from Gi.

K ,is the number of the nearest neighbors to yi which can take an integer value between 1 and n-1. On the other hand, C, the number of Ei sets ($1 \le i \le C$), is influenced by K and the number of samples (n). Thus, when $k \rightarrow (n-1)$ then $c \rightarrow 1$ and if $k \rightarrow 1$ then $c \rightarrow n/2$. This lets us to express K as the following function:

$$K = Integer\left(\frac{n}{c} - 1\right) \quad (10)$$

Also, unclassified samples of x¢(E1…Ec) will be related to the nearest centers (G1…Gc), and then all these centers will be updated with new members. Therefore, at the end of this stage every sample will be assigned to the obtained sets of (E1,…,Ec).In other words, this process lets us divide the data set X to some groups (with C clusters) in order to define the initial cluster centers and also to become close to the best clustering.

In the second step of our algorithm we have employed FCM algorithm with one iteration and using the PC function to choose the number of clusters(c) automatically. To sum up this stage in brief, first we should run one FCM iteration, calculate $U^* = [u^*_{ik}]_{c.n}$

Using equation (4), obtain the centers of the fuzzy clusters ($V^* = (V^*_1, \ldots .V^*_c)$) using equation (5), calculate PC function using equation (6), choose the maximum value for PC and related c^*(the number of clusters) as the optimum value for it, and finally use the outcome along with the obtained c^* as our final clustering result.

RESULTS

For comparing our method with FCM and KNN hard algorithm, we used six standard datasets and applied our proposed method to them. Knowing the correct results a priori, one can easily calculate the error rate for each method. Table I contains the results of the calculated error rate for each method. In case of FCM the error rate value is the average value over twenty iterations with different starting points.

The S1-S4 sets contain 2,3,4 and 6 overlapping clusters respectively. In each cluster, there are twenty samples from R^2 space, and the overlap degree progressively changes between clusters of each set.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

The S5 and S6 sets are from IRIS data which contain 3 clusters. Each cluster contains fifty samples in R^4 space. IRIS data are samples taken from different flowers used to determine both the family and type of flowers. Each sample has four dimensions; each represents the length and width of leaflet and petal separately. The first cluster of the data is well separated from clusters 2 and 3 which have high interference.

Regarding that, this set contains real data in which the elements have high interference, utilizing a fuzzy measure seems to be appropriate for and compatible with the nature of the data. The S5 involves each three classes of IRIS while S6 contains only two clusters of 2 and 3. Using these sets of data we can analyze the behavior of the algorithm on two inseparable clusters. Clustering algorithms, particularly the fuzzy algorithms, use IRIS data sets for testing the operation. In the following, the operation of algorithms i.e. the number of iterations, the rate of wrong sample clustering and stability against changes in the weight of fuzziness (m), will be considered. The rate of wrong clustering will be calculated using the maximum value of membership function for each sample and changing it into a decisive result, which could be referred to as final defuzzification.

In the first analysis we performed FCM with twenty iterations with different initial centers and the stop measure was set to $\varepsilon = 0.001$ while the fuzziness coefficient was set to $m = 2$.

TABLE 1: COMPARISON OF OPERATION OF METHODS (M=2)

Datasets	The correct number of clusters	the number of clusters each method calculated			Number Of FCM iterations
		FCM	KNN	KNN-FCM	
S1	2	0	0	0	7
S2	3	5	5	3	8
S3	4	3	2	3	19
S4	6	7	10	7	16
IRIS	3	16	13	15	11
IRIS23	2	15	14	14	10

TABLE 2: CALCULATION OF NUMBER OF CLUSTERS EVALUATION

Datasets	c^*	FCM		Hybrid KNN-FCM	
		c^*	PCmax	c^*	PCmax
S1	2	2	0.9219	2	0.9229
S2	3	2	0.8528	3	0.8428
S3	4	3	0.9119	4	0.8747
S4	6	2	0.8519	4	0.8641
IRIS	3	2	0.8860	2	0.8319
IRIS23	2	2	0.7408	2	0.7483

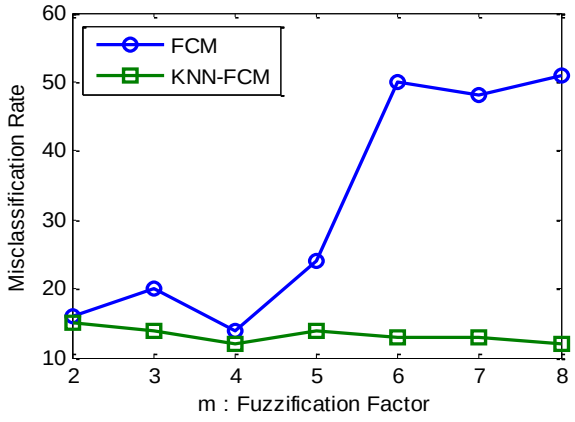


Fig. 1 Study on changes of m and its effects

$$u_{ik,\max} = \begin{cases} 1 & \text{if } uik \geq usk \ , \ 1 \leq s \leq c \ , \ s \neq i, \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

In Fig. 2 you can see the way of clustering for datasets and the accuracy of the proposed method, in this figure we have the X-Y space, in which there are several samples which should be clustered, as you see from (a) to (e), we obtained clustering for S1 to IRIS respectively.

Fuzzy Coefficient Changes' Effects: In order to examine the stability of both FCM and our method with respect to parameter m and also make a comparison between the m, we performed both methods with different values of m and collected the results in Fig. 1. The results show that the proposed method is more stable than FCM. As can bee seen from this figure, the FCM strongly depends on the exact value of m, i.e. Fuzziness coefficient, and the results change automatically by m variation. In contrast, our method benefits from high robustness against parameters 'variation, so user does not need to worry about specifying the parameters' value. In Fig. 1 we have the results for IRIS dataset that has the most varieties in results. This figure shows the misclassification rate with respect to m variations.

In fuzzy methods, analyzing the membership degree of samples can do measuring the quality of the method. in Fig. 3 the trend of membership function for different clusters has been shown. As you see, in this figure most of the samples have high degree of membership to their main clusters.

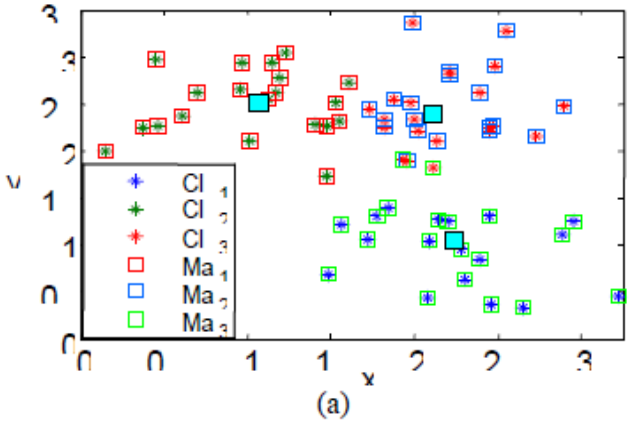
Comparison of the performance of fuzzy clustering methods in finding the number of optimized clusters automatically.

In this section, making use of the standard function, i.e. PC, the strength of FCM and Hybrid algorithms, in finding the number of optimized clusters and identical previous datasets, will be studied.

The division coefficient PC is a measure of the classification quality. Therefore, for different number of clusters (C=2,…,Cmax=√n), we will repeat the clustering, and make use of the PC measure. The effect of K on C parts obtained from the fuzzy clustering will be studied. The best clustering will be obtained by finding the highest value for PC.

As you see in table II, the Hybrid method obtained the number of clusters more accurate than FCM algorithm. In our proposed hybrid method PC_{max} is so greater than 1/C and this method has specified the correct number of clusters for all datasets except S4 and IRIS. This is due to the high interference in these clusters, e.g. S4 has 6 overlapping clusters.

Hybrid fuzzy algorithm, even in cases with faulty detection, performs better than the common algorithm of FCM.



Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

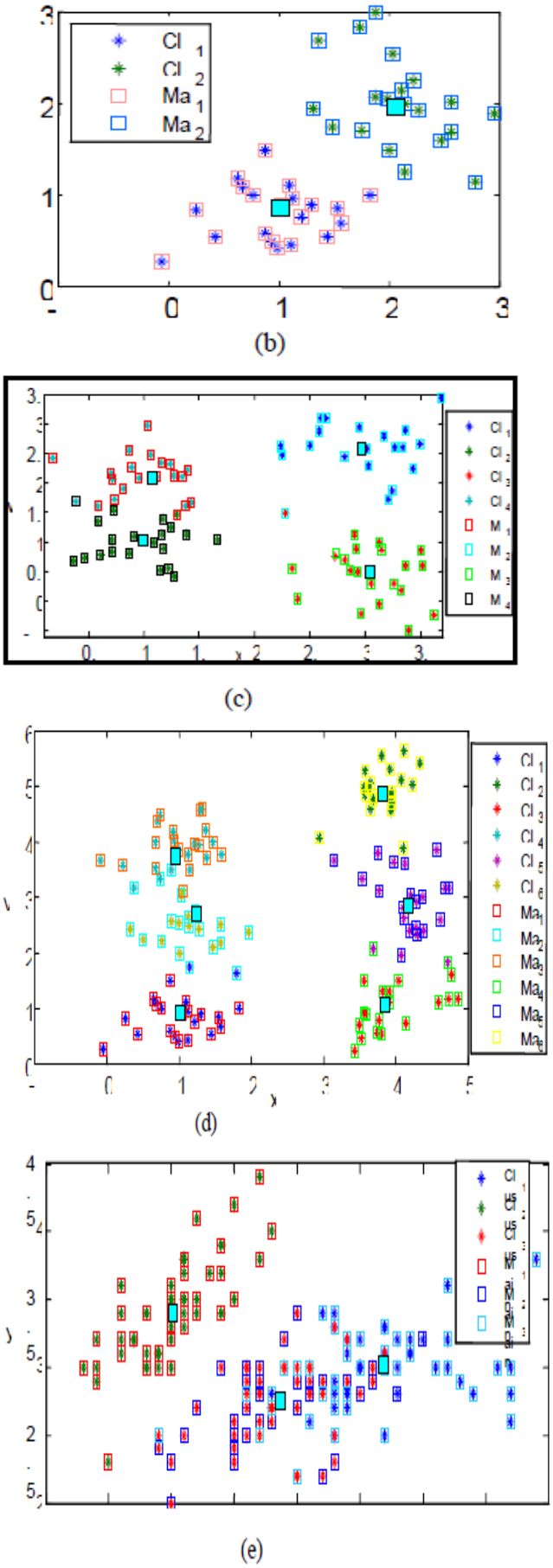


Fig. 2p: Fuzzy hybrid clustering (a) Hybrid KNN-FCM for S_1 (b) Hybrid KNN-FCM for S_2 (c) Hybrid KNN-FCM for S_3 (d) Hybrid KNN-FCM for S_4 (e) Hybrid KNN-FCM for IRIS, IRIS dataset is shown in only 2 dimensions out of 4 dimensions that have the most overlap

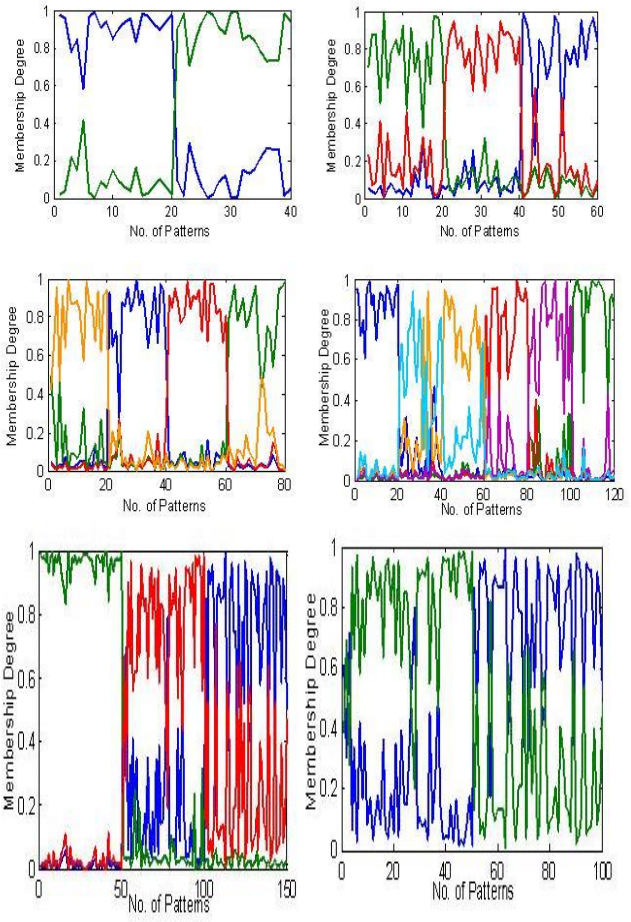


Fig. 3: Fuzzy membership functions of each dataset for clusters (a) Membership Function of Patterns – S2 (b) Membership Function of Patterns – S1 (c) Membership Function of Patterns – S3 (d) Membership Function of Patterns – S4 (e) Membership Function of Patterns – IRIS (f) Membership Function of Patterns – IRIS

DISCUSSION

This work proposes a hybrid fuzzy clustering algorithm. Unlike the conventional fuzzy c-mean (FCM) method which is a semi-unsupervised clustering method, the proposed method in this paper follows an absolute unsupervised procedure to cluster of input data. While the FCM needs to be defined with some initial parameters, the proposed method doesn't need to any pre-defined initial parameters. The results showed the efficiency of the proposed method. However more confidence can be resulted with more simulation on real data sets.

CONCLUSION

Making use of FCM needs to specify some parameters, such as the number of clusters, algorithm-stopping measure on which the number of iterations are based on and fuzziness coefficient, i.e. m.

These parameters have great influences on results. Specifying initial values for parameters to obtain the first clustering, an iterative process up to reaching to stopping point is done ,in this initial values the user should arbitrarily choose the initial points as the clusters' center, and as you know different initial points may cause the target function to converge to the different local minimum points. Hence in

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

the hope of reaching to the global minimum point, the clustering would be repeated with different initial parameters.

In this paper a hybrid fuzzy method is proposed which is based on 2-layer clustering strategy. In the first layer, the unsupervised clustering by using KNN rule is done, and then the second layer containing one FCM iteration is performed. This algorithm has been tested and analyzed on six datasets. The results show that this algorithm can be used without any assumption about cluster's centers, fuzziness coefficient (m), algorithm stopping measure and the number of clusters. Furthermore it has been shown that for the clusters with considerable overlap ,the proposed method is more efficient than FCM.

REFERENCES

[1] Kuo-Lung Wu, Miin-Shen Yang - Alternative c-means clustering algorithms - Pattern Recognition35 (2002) 2267 – 2278.
[2] Dae-Won Kima, Kwang H. Lee, Doheon Lee, - On cluster validity index for estimation of the optimalnumber of fuzzy clusters - Pattern Recognition 37 (2004) 2009 – 2025.
[3] Oleg S. Pianykh - Analytically tractable case of fuzzy c-means clustering - Pattern Recognition 39(2006) 35 – 46
[4] Luis Rueda, Yuanquan Zhang - Geometric visualization of clusters obtained from fuzzy clusteringalgorithms - Pattern Recognition 39 (2006) 1415 – 1429
[5] Carl G. Looney - Interactive clustering and merging with a new fuzzy expected value –PatternRecognition 35 (2005) 2413 – 2423
[6] WitoldPedrycz, George Vukovich - Fuzzy clustering with supervision - Pattern Recognition 37(2004) 1339 – 1349
[7] Haojun Sun , Shengrui Wang, Qingshan Jiang - FCM-Based Model Selection Algorithms forDetermining the Number of Clusters - Pattern Recognition 37 (2004) 2027 – 2037
[8] Nabil Belacel, Pierre Hansen, NenadMladenovic - FuzzyJ-Means: a new heuristic for fuzzyclustering - Pattern Recognition 35 (2002) 2193 – 2200
[9] WeilingCai, Songcan Chen, Daoqiang Zhang - Fast and robust fuzzy c-means clustering algorithmsincorporating local information for image segmentation - Pattern Recognition 40 (2007) 825 – 838
[10] Chien-Hsing Chou, Chin-Chin Lin,Ying-Ho Liu, Fu Chang – A prototype classification method andits use in a hybrid solution for multiclass pattern recognition - Pattern Recognition 39 (2006) 624 – 634
[11] N. Zahid, M. Limouri, A. Essaid - A new cluster-validity for fuzzy clustering - Pattern Recognition32 (1999) 1089}1097
[12] Mario G.C.A. Cimino, Beatrice Lazzerini, Francesco Marcelloni – A novel approach to fuzzyclustering based on a dissimilarity relation extracted from data using a TS system - Pattern Recognition39 (2006) 2077 – 2091
[13] Michel MeH nard, Christophe Demko, Pierre Loonis - The fuzzy c-means: solving the ambiguityrejection in clustering - Pattern Recognition 33 (2000) 1219}1237
[14] Malay K. Pakhira, SanghamitraBandyopadhyay, UjjwalMaulik - Validity index for crisp and fuzzy clusters - Pattern Recognition 37 (2004) 487 – 501
[15] Wuhan, Hubei, China ",A Modified FCM Algorithm for MRI Brain Image Segmentation," 2008 International Seminar on Future Bio Medical Information Engineering, December18, 2008

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.