Australian Journal of Basic and Applied Sciences, 5(8): 674-681, 2011
ISSN 1991-8178

A Modified Hybrid Fuzzy Clustering Algorithm for Data Partitions

[1]J. Hossen, [2]A. Rahman, [3]S. Sayeed, [2]K. Samsuddin, [2]F. Rokhani

[1]Institute of Advanced Technology (ITMA), Universiti Putra Malaysia (UPM), Malaysia.
[2]Faculty of Engineering, Universiti Putra Malaysia (UPM), Malaysia.
[3]Faculty of Information Science and Technology (FIST), Multimedia University (MMU), Malaysia.

Abstract: The clustering algorithm hybridization scheme has become of research interest in data partitioning applications in recent years. The present paper proposes a Hybrid Fuzzy clustering algorithm (combination of Fuzzy C-means with extension and Subtractive clustering algorithm) for data classifications applications. The fuzzy c-means (FCM) and subtractive clustering (SC) algorithm has been widely discussed and applied in pattern recognitions, machine learning and data classifications. However the FCM could not guarantee unique clustering result because initial cluster number is chosen randomly as the result of the classification is unstable. On the other hand, the SC is a fast, one-pass algorithm for estimating the numbers and center of clusters for a set of data. This paper presents the two different clustering algorithms and their comparison. First clustering algorithm is fuzzy c-means clustering, and second is subtractive clustering. Results show that the SC is better than FCM in respect of speed but not as good in accuracy, so a modified hybrid clustering algorithm is designed with all these parameters. The experiments show that the hybrid clustering algorithm can improve the speed, and reduce the iterative amount. At the same time, the hybrid algorithm can make the results of data partitions are more stable and higher accuracy.

Key words: FCM, Hybrid Fuzzy clustering, Partition coefficient, Subtractive clustering.

INTRODUCTION

Fuzzy logic starts with the concept of a *fuzzy set* (Zadeh, 1965). A fuzzy set is a set without a crisp, clearly defined boundary. It can contain elements with only a partial degree of membership. To understand what a fuzzy set is, first consider the definition of a *classical set*. A classical set is a container that wholly includes or wholly excludes any given element. In fuzzy logic is just a matter of generalizing the familiar yes/no (Boolean) logic. The truth values are multi-valued such as absolutely true, partly true, absolutely false, very true and so on are numerically equivalent in between 0 and1. Fuzzy sets are sets whose elements have degrees of membership. Fuzzy sets were introduced by Lotfi A. Zadeh as an extension of the classical notion of set. In classical set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition an element either belongs or does not belong to the set. Classical bivalent sets are in fuzzy set theory usually called *crisp* sets.

Fuzzy clustering (Chen, 2006) is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system's behavior. Clustering can also be thought of as a form of data compression where a large number of samples are converted into a small number of representative prototypes or clusters. In non-fuzzy or hard clustering data is divided into crisp clusters where each data point belongs to exactly one cluster. In fuzzy clustering, the data points can belong to more than one cluster, and associated with each of the points are membership grades which indicate the degree to which the data points belong to the different clusters. Fuzzy clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups (Jang, 1997). Data classifications is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. This paper presents the two different clustering algorithms and their comparison. First clustering algorithm is fuzzy c-means clustering, and second is subtractive clustering. Results show that the SC is better than FCM in respect of speed but not as good in accuracy, so a modified hybrid clustering algorithm is designed with all these parameters. The experiments show that the hybrid clustering algorithm can improve the speed, and reduce the iterative amount. At the same time, the hybrid algorithm can make the results of data partitions are more stable and higher accuracy.

Corresponding Author: J. Hossen, Institute of Advanced Technology (ITMA), Universiti Putra Malaysia (UPM), Malaysia.
E-mail: (jakir.hossen@mmu.edu.my, 006-062523382)

This paper is organized as follows: Section II provides a brief review on Subtractive clustering algorithm and Fuzzy C-means clustering with some extension in section III. Section IV presents the design of the proposed Hybrid Fuzzy clustering algorithm. Experimental results and Validation method with results based on standard data sets for the Hybrid Fuzzy clustering algorithm are compared with other existing data partitions algorithms (SC, FCM) in Section V and VI respectively. The final conclusions are drawn in Section VII.

Subtrative Clustering (SC):

Subtractive clustering originally came from mountain method (S. Chiu, 1994). It is a fast, one-pass algorithm for estimating the number and the centers of clusters for a set of data. Let *x* be the data set formed by concatenating the input data set *X* and the output data set *Y* of the system. Also, assume that each dimension of the data is normalized, that means data set *x* is bound by hypercube. Subtractive clustering treats each point as a potential cluster center and uses the following equation (1) as a measurement:

$$D_i = \sum_{j=i}^{n} \exp\left[\frac{\left|x_i - x_j\right|^2}{\left(\frac{r_a}{2}\right)^2}\right] \tag{1}$$

Where r_a defines the neighborhood radius for each cluster, $|\cdot|$ is Euclidean distance and *n* is the number of sampling points of the data set *x*. Using equation (1), subtractive algorithm will compute the potential for each point. The point with the highest potential, denoted by D_{c1} is selected as the first cluster center x_{c1}. Next, the potential of each data point x_i is updated uses the following equation (2) as follows:

$$D_i = D_i - D_{c1} \exp\left(\frac{\left|x_i - x_{c1}\right|^2}{\left(\frac{r_b}{2}\right)^2}\right) \tag{2}$$

Where r_b represents the radius of the neighborhood with significant potential reduction. Normally, r_b should be chosen to be higher than r_a to avoid closely spaced clusters. The next center is selected is the point with the highest potential. This process will continue until a stopping criterion is reached.

Let ε_h represent a threshold above which the point will definitely be selected as a center. Let ε_l represent a threshold below which the point will definitely be rejected. Four parameters, r_a, r_b, ε_l and ε_h influences the number of rules and accuracy of clustering. Normally, $r_b = 1.5\ r_a$, $0.15 \le r_a \le 0.3$, $\varepsilon_l = 0.15$ and $\varepsilon_h = 0.5$ are used. Better models can be achieved using different combination of these four parameters for different data sets. In this paper, we use subtractive algorithm to cluster the data set initially, and then, pass the number of clusters and centers of clusters to fuzzy C-means algorithm with extension, so that fuzzy C-means algorithm will refine the clustering as hybrid fuzzy clustering algorithm.

Fuzzy C-means Clustering with Some Extension:

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. This technique was originally introduced by Jim Bezdek in (1981) 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. The FCM algorithm is one of the most widely used fuzzy clustering algorithms (Newton, 1992; Lee, 2001). The FCM algorithm attempts to partition a finite collection of elements $X=\{x_1,x_2,\ldots\ldots,x_n\}$ into a collection of 'c' fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centers C, such that $C = c_i$, i =1, 2,…c and a partition matrix *U* such that $U = u_{ij}$, i =1, 2, ...c, j =1,...n where u_{ij} is a numerical value in [0, 1] that tells the degree to which the element x_j belongs to the i-th cluster. If the number of clusters is not properly assigned, the FCM algorithm might not produce meaningful fuzzy partitions (Chen, 2006). The following steps explain the implementation of FCM algorithm:

Step 1: Initialize the membership matrix U with random values between 0 and 1 such that the constraints in Equation (3) are satisfied.

$$\sum_{i=1}^{C} u_{ij} = 1.0, \tag{3}$$

Step 2: Calculate *c* fuzzy cluster centers c_i using equation (4).

$$c_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{i=j}^{N} u_{ij}^{m}} \tag{4}$$

Step 3: Compute the cost function according to equation (5).

$$J(U,c) = \sum_{i=1}^{c}\sum_{j=1}^{n} (u_{ij})^m (d_{ij})^2 \quad \text{, where } d_{ij} = (x_j - c_i) \tag{5}$$

Stop if either it is below a certain tolerance value or its improvement over previous iteration is below a certain threshold.

Step 4: Compute a new U using following equitation (6).

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}} \tag{6}$$

$\Delta = \| U^{l+1} - U^{l} \| = max_{i,j}|u_{ij}^{i+1} - u_{ij}^{l}|$. If $\Delta > \epsilon$, then set $i = i + 1$ and go to step 2. If $\Delta <= \epsilon$, then stop.

The Fuzzy covariance matrix FC_i is calculated by the following equation (7):

$$FC_i = \frac{\sum_{j=1}^{n} u_{ij}^{m}(c_i - x_j)(c_i - x_j)^T}{\sum_{j=1}^{n} u_{ij}^{m}} \quad \text{, where } 1 \le i \le c \tag{7}$$

The numerator of Equation (7) is the fuzzy scatter matrix for i^{th} cluster. The standard deviation of each membership function can be represented by using the root of diagonal element of *FCi* as follows:

$$\sigma_{ik} = \sqrt{Diag(FC_i)}, \quad \text{where } 1 \le k \le d \tag{8}$$

where *d* are the dimensions of input vectors. σ_{ik} and the i^{th} cluster center c_i are used to initialize the parameters of Gaussian membership functions. Also, after finishing clustering, fuzzy rules will be obtained. Each cluster will represent a rule.

The FCM algorithm assigns a small noise as a same membership value 1/*c* to each cluster. To overcome this handicap due to the false representation, a modified α–cut method is utilized to remove the noise. Then, to fit those processed membership values for each fuzzy set, a modified asymmetric Gaussian membership function as defined by (Hossen, 2011) is chosen for the adaptive membership function scheme that provides more flexibility as equation (9) as follows.

$$\mu_{ij} = \mu_{ij1} \times \mu_{ij2} \tag{9}$$

$$\mu_{ij1} = e^{-\left(\frac{x_{kj}-v_{ij1}}{\sigma_{ij1}}\right)^2} \times Index_{v_{ij1}} + (1 - Index_{v_{ij1}})$$

$$\mu_{ij2} = e^{-\left(\frac{x_{kj}-v_{ij2}}{\sigma_{ij2}}\right)^2} \times Index_{v_{ij2}} + (1 - Index_{v_{ij2}})$$

$$\textit{if } x_{kj} \le v_{ij1}, Index_{vij1} = 1, \ \textit{otherwise } 0$$

$$\textit{if } x_{kj} \ge v_{ij2}, Index_{vij2} = 1, \ \textit{otherwise } 0$$

Proposed Hybrid Fuzzy Clustering Algorithm:

Normally all the clustering techniques are to find cluster centers which is a way to tell where the heart of each cluster is located. Fuzzy C-means clustering relies on knowing the number of cluster *a priori*. In that case, the algorithm tries to classify the data into the given number of clusters. The performance of the fuzzy C-means clustering is very dependent on the choice of the initial cluster center and tends to converge to a nearby local optimum. Many research teams are trying to develop global optimizers for C-means clustering (Y.-K. Hu and Y. P. Hu, 2006; Mohanad A., 2008; Bainian Li, 2010; M. Chen and S. Wang, 1999).

In this paper, we use subtractive clustering to find the initial number of clusters and initial centers as subtractive clustering starts by finding the first large cluster and then go to find the second, and so on. But the problem with subtractive clustering is that the accuracy is dependent on the good choice of four parameter, such as, r_a, r_b, ε_l and ε_h. Therefore, this paper uses hybrid fuzzy clustering technique for data classifications applications.

The Hybrid Fuzzy clustering algorithm is presented as follows:

Step 1: For i = 1, .., n, calculate the potential D_i using the equation (1) as follows.

$$D_i = \sum_{j=i}^{n} \exp\left[\frac{|x_i - x_j|^2}{\left(\frac{r_a}{2}\right)^2}\right]$$

Step 2: Set n_c = 1, consider the highest potential of data point as D_{c1} and the location of that point as x_{c1} as first cluster center.

Step 3: Update each point potential using the equation (2) as follows.

$$D_i = D_i - D_{c1} \exp\left(\frac{|x_i - x_{c1}|^2}{\left(\frac{r_b}{2}\right)^2}\right)$$

Step 4: If $max\ D_i \ge \varepsilon_h D_{c1}$ is true, accept x_{ci} is the next cluster center, continue until getting the final (all) cluster center from whole set of data.

Step 5: If *max* $D_i < \varepsilon_h D_{c1}$ is true, go to Step 4, otherwise, check if the point provides a good trade-off between having a sufficient potential and being sufficiently far away from existing cluster centers. If this is the case, this point is selected as the next cluster center.

Step 6: Calculate u_{ij} using the equation (6) as follows.

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

Step 7: Update fuzzy cluster center using the equation (4) as follows.

$$c_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{i=j}^{N} u_{ij}^m}$$

Step 8: Compute the cost function according to the Equation (5) as follows.

$$J(U,c) = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij})^m (d_{ij})^2$$

Stop if it is below a certain tolerance value or its improvement is below a certain thresh. Otherwise, go to Step 4.

Step 9: calculate the standard deviation (width) of each membership function using the equation (8) as follows.

$$\sigma_{ik} = \sqrt{Diag(FC_i)}$$

Experimental Results:

The Subtractive Clustering algorithm is designed and implemented using MATLAB whose results are shown in the Fig. 1. The Figure shows, there are two clusters formed with one center each but there is some overlapping.

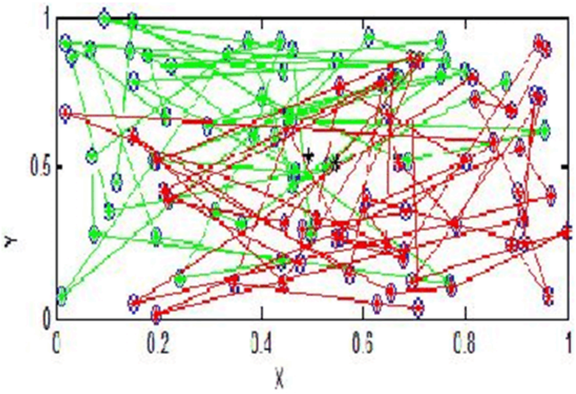

Fig. 1: Subtractive clustering based data partitions.

The FCM algorithm is also designed and implemented using MATLAB whose results are shown in the Fig. 2. The Figure shows, there are two clusters formed with one center each but there is also overlapping between the datasets of two clusters that makes the complexity to differentiate the data of the two clusters.

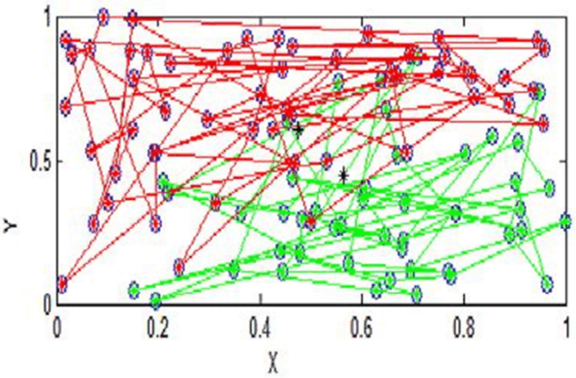

Fig. 2: FCM clustering based data partitions.

So a proposed hybrid clustering algorithm is designed which reduces the overlapping of the data completely. This algorithm separate the data of the two clusters completely and also gives the better performance in respect of speed and rate of convergence as shown in Fig. 3.

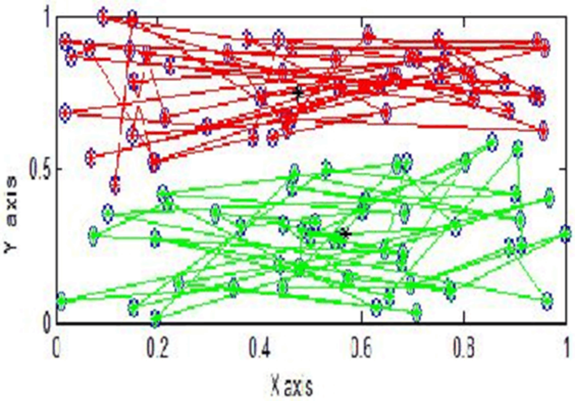

Fig. 3: Hybrid Fuzzy clustering based data partitions.

We carry out many trials using the standard data sets. During the process we alter the number of data point and the other parameters and we can find that the results we gain using two methods are same at the most time but the number of iteration of FCM is greater than the number of the other one. We also calculate the speed in term of iterations as shown in table II, and we find out after many trials that the iterations which is been taken is less than other algorithms in many cases and the accuracy is almost same but for some cases it is slightly greater than other algorithms.

Table 1: Accuracy and Iterations of three clustering algorithms.

DataTypes	Accuracy (%)			No of Iteration		
	SC	FCM	Hybrid	SC	FCM	Hybrid
Data1	78	81	82	16	33	14
Data2	88	90	92	17	36	15

Therefore, the solution obtained by the method of hybrid is better than the solution obtained by the methods of SC, FCM. Because estimating the optimum number of clustering is very difficult, and the method of combining subtractive clustering with FCM is better than the methods of SC and FCM.

Validation Method with Results:

To evaluate the efficiency of clustering algorithms, Partition coefficient (PC) have been employed in the data partition experiments. Bezdek (1981) has been defined a performance measure (reduce overlapping data in the clusters) based on minimizing the overall content of pair wise fuzzy intersection in *U,* the partition matrix. Trauwaert (1988) has also proposed cluster validity index for fuzzy clustering. The Partition Coefficient (PC) index equation is defined as following equation in (10).

$$PC = \frac{1}{N\sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{2}} \tag{10}$$

The *PC* index indicates the average relative amount of membership sharing done between pairs of fuzzy subsets in *U* by combining into a single number, the average contents of pairs of fuzzy algebraic products. The index values range in 1/*C,* where *C* is the number of clusters. The more the *PC* is closer to 1, the more the result is distinct; on the contrary the more the *PC* is closer to1/*C,* the more the result is vaguer. For this experiment, the PC index results have been shown in the table I as following.

Table 2: Partition Coefficient of three clustering algorithms.

Partition Coefficient (PC)		
Subtractive Clustering	Fuzzy C-means	Hybrid Clustering
0.44	0.60	0.97

Conclusion:

In this paper, we find out that Subtractive clustering which is a fast, one-pass algorithm technique, takes less time than FCM for data partitions but the accuracy was not good as FCM and also there was overlapping problem which was not solved. So a modified hybrid fuzzy clustering algorithm has been designed and simulated for data partitions applications. Result shows that Hybrid algorithm takes less number of iterations to create data sets as compared to FCM and SC. And also it takes less time & it solves the overlapping between the two clusters data. So it can be concluded from this proposed work that Hybrid fuzzy clustering algorithm is better for data classifications and provides fast convergence solutions to Fuzzy logic problems.

REFERENCES

Bezdek, J.C., 1981. Pattern Recognition and Fuzzy Objective Function Algorithms, *Plenum Press*, pp. 65-86, New York.

Bainian, Li., Kongsheng Zhang, and Xu. Jian, 2010. Similarity measures and weighed fuzzy c-mean clustering algorithm, International Journal of Electrical and *Computer Engineering under WASET,* 6(1): 1-4.

Chen, M. and S. Wang, 1999. Fuzzy clustering analysis for optimizing fuzzy membership functions, *Fuzzy Sets and Systems*, 103: 239-254.

Chen, W. J., M. L. Giger and U. Bick, 2006. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast enhanced MRI images, *Acad. Radiol*, 13(1): 63-72.

Chiu, S., 1994. Fuzzy model identification based on cluster estimation, *Journal of Intelligent and Fuzzy Systems*, 2(3): 267-278.

Hossen, J., A. Rahman, K. Samsudin, F. Rokhani, S. Sayeed, R. Hasan, 2011. A Novel Modified Adaptive Fuzzy Inference Engine And Its Application to Pattern Classification, World Academy of Science, Engineering and Technology, 75: 1201-1207.

Hu, Y.-K. and Y.P. Hu, 2006. Global optimization in clustering using hyperbolic cross points, *Pattern Regonition*, vol, 10.

Jang, J.-S. R., C.-T. Sun, E. Mizutani, 1997. Neuro- Fuzzy and Soft Computing - A Computational Approach to Learning and Machine Intelligence, *Prentice Hall.*

Lee, H.-M., C.-M. Chen, J.-M. Chen and Y.-L. Jou, 2001. An Efficient Fuzzy Classifier with Feature Selection Based on Fuzzy Entropy, *IEEE Transaction on Systems,Man, and Cybernetics – Part B: Cybernetics*, 31(3): 426-432.

Mohanad, A., M. Mohammad, R. Abdullah, 2008. Optimizing of Fuzzy C-Means clustering Algorithm Using GA, *World Academy of Science, Engineering and Technology,* Vol. 39.

Newton, S.C., S. Pemmaraju and S. Mitra, 1992. Adaptive Fuzzy Leader Clustering of Complex Data Sets in Pattern Recognition, *IEEE Transaction on Neural Networks*, 3(5): 794-800.

Trauwaert, E., 1988. On the meaning of Dunn's partition coefficient for fuzzy clusters, *Fuzzy Sets and Systems*, 25: 217-242.

Zadeh, L.A., 1965. Fuzzy Sets, *Information and Control*, 8(3): 338-353.