Proceedings of the 7th WSEAS International Conference on Evolutionary Computing, Cavtat, Croatia, June 12-14, 2006 (pp7-11)

Alternative Adaptive Fuzzy C-Means Clustering

SOMCHAI CHAMPATHONG*
Department of Computer Science
Faculty of Science
Khon Kaen University
THAILAND

SARTRA WONGTHANAVASU
Department of Computer Science
Faculty of Science
Khon Kaen University
THAILAND

KHAMRON SUNAT
Department of Computer Engineering
Faculty of Engineering
Mahanakorn University of technology
THAILAND

Abstract: -Fuzzy C-Means (FCM) clustering algorithm is used in a variety of application domains. Fundamentally, it cannot be used for the subsequent data (adaptive data). A complete dataset has to be static prior to implementing the algorithm. This paper presents an alternative adaptive FCM which is able to cope with this limitation. The adaptive FCM using Euclidean and Mahalanobis distances were compared to alternative adaptive FCM for performance evaluation purposes. Two different datasets were taken into consideration for the compared test. In this respect, adaptive FCM using Euclidean and Mahalanobis distances results in more misclassified data. By implementing synthesis dataset with outlier, adaptive FCM using Euclidean and Mahalanobis distances give 9% and 14% of misclassification, respectively. While implemented in alternative adaptive FCM the proposed method exhibits the promising performance by giving 2% of misclassification. This result shows similar manner for carrying out in iris dataset.

Keywords: Clustering, Euclidean distance, Mahalanobis distance, Alternative distance ,Adaptation, Outlier

1 Introduction

Fuzzy C-Means (FCM) clustering is analogous to a traditional cluster analysis. Cluster analysis or clustering is a method that groups patterns of data that in some sense belong together and have similar characteristics. FCM clustering assigned each data through the class centroid (prototype) in each class. Call value of each data in each class "membership value". The final membership values with FCM range between 0 and 1 for each data point, while the sum of values for a particular data across all classes equals to 1, equation (4). When subsequent data come basic FCM can not be use until data stable. Adaptive FCM is proposed by Stefano [10] to solve this problem. But some data has outlier adaptive FCM is not robust because outliers are effect to all cluster. This paper describes how to solve the adaptive outlier data cluster. The alternative adaptive Fuzzy C-Means is proposed to solve clustering for adaptive outlier data.

* Graduate student, Khon Kaen University, Thailand. Corresponding author: Tel: (66) 26448320

2 Basic FCM algorithm

FCM minimises an objective function using equation shown in (1):

$$J_{FCM} = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^m d^2(x_j, z_i) \quad (1)$$

where n is a number of data, c is a number of classes, z_i denotes the vector representing the centroid (prototype) of class i, x_j denotes the vector representing individual data j and $d^2(x_j - z_i)$ denotes the square distance between x_j and z_i according to a chosen definition of distance. m is the fuzzy exponent and ranges from (1, ∞). It determines the degree of fuzziness

of the final solution, that is the degree of overlap between groups. With $m = 1$, the solution is hard partition. As m approaches infinity the solution approaches its highest degree of fuzziness.

1. Choose the number of classes c, with $1<c<n$.

2. Choose a value for the fuzziness exponent m, with $m > 1$.

3. Choose definition of distance in the variable-space.

4. Choose a value for the stopping criterion ε

5. Initialize M=M$^{(0)}$, e.g. with random memberships

6. At iteration it= 1 2 3

(re) calculate $z = z^{(it)}$ using equation (2) and M$^{(it-1)}$

$$z_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m x_j}{\sum_{j=1}^{n} (\mu_{ij})^m} \qquad (2)$$

7. Re-calculate M=M$^{(it)}$ using equation (3) and $z^{(it)}$

$$\mu_{ij} = \frac{\left[1/d^2(x_j - z_i)\right]^{1/(m-1)}}{\sum_{i=1}^{c}\left[1/d^2(x_j - z_i)\right]^{1/(m-1)}} \qquad (3)$$

where

$$\sum_{i=1}^{c} \mu_{ij} = 1 \qquad (4)$$

8. If ||M$^{(it)}$-M$^{(it-1)}$|| < ε ,then stop; otherwise return to step 6.

3 Distance measures

3.1 Euclidean distance

Euclidean distance is defined by an equation (5) given following:.

$$EDji = \sqrt{(x_j - z_i)(x_j - z_i)^T} \qquad (5)$$

Where ED_{ji} is Euclidean distance vector representing distance between data x_j and prototypes z_i, T representing transpose matrix.

3.2 Mahalanobis distance:

Mahalanobis distance is defined in equation (6):

$$MD_{ji} = \sqrt{(x_j - z_i)A^{-1}(x_j - z_i)^T} \qquad (6)$$

where MD_{ji} is Mahalanobis distance vector representing distance between data x_j and prototypes z_i, T represents a transpose matrix. A is variance-covariance matrix defined by equation (7)

$$A = \frac{\sum_{j=1}^{n} (x_j - z_i)^T (x_j - z_i)}{n-1} \qquad (7)$$

3.3 Alternative distance

Alternative distance [9] is given in equation (8).

$$AD_{ji} = \exp(-\beta D^2(x_j, z_i)) \qquad (8)$$

where AD_{ji} is alternative distance vector representing distance between data x_j and prototypes z_i, $D^2(x_j - z_i)$ is square Euclidean distance between data x_j and prototypes z_i, β defined by equation (9)

$$\beta = \left(\frac{\sum_{j=1}^{n} \| x_j - \bar{x} \|^2}{n} \right)^{-1} ; \ \bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \qquad (9)$$

4 Adaptive FCM

Basic FCM is used to cluster static data. It will be much convenient to use the knowledge already gained in partitioning a given set to classify more data. This knowledge is condensed into the prototypes $\{z_i; i = 1,...,c\}$ and can be used to classify subsequent data without reprocess whole data [13]. To design a classifier for a new entry x_{n+1} on the basis of the prior knowledge $\{z_i; i = 1,...,c\}$, equation (3) can be used again to determine the memberships of new data, after the distance between x_{n+1} and each of the prototypes z_i have been determined. Contrary to the previous iterative solution of equation (3), now

the prototypes z_i remain unchanged and equation (3) is used only once, to obtain the memberships of new point.

$$\mu_{i,n+1} = \frac{\left[1/d^2(x_{n+1} - z_i)\right]^{1/(m-1)}}{\sum_{i=1}^{c}\left[1/d^2(x_{n+1} - z_i)\right]^{1/(m-1)}} \qquad (10)$$

Notice that condition (4), which can now be written as

$$\sum_{i=1}^{c} \mu_{i,n+1} = 1 \qquad (11)$$

Making the clustering procedure adaptive can be done by reflecting the changing membership values into the prototype locations, which in the case of $n+1$ data points can be written as

$$z_i = \frac{\sum_{j=1}^{n}(\mu_{j,i})^m x_j + (\mu_{n+1,i})^m x_{n+1}}{\sum_{j=1}^{n}(\mu_{j,i})^m + (\mu_{n+1})^m} \qquad (12)$$

5 Alternative Adaptive FCM

Adaptive FCM effected for subsequent data that has outliers. This paper proposed how to solve this effect. The algorithm of alternative adaptive FCM is similar to adaptive FCM except that it takes alternative distance (8) only and update prototypes are replaced by equation (13)

$$\frac{S(n) + (\mu_{n+1,i})^m \exp(-\beta_{n+1} D^2(x_{n+1} - z_i))x_{n+1}}{M(n) + (\mu_{n+1,i})^m \exp(-\beta_{n+1} D^2(x_{n+1} - z_i))} \qquad (13)$$

.

Where

$$S(n) = \sum_{j=1}^{n}(\mu_{ij})^m \exp(-\beta_n D^2(x_j - z_i))x_j \qquad (14)$$

and

$$M(n) = \sum_{j=1}^{n}(\mu_{ij})^m \exp(-\beta_n D^2(x_j - z_i)) \qquad (15)$$

6 Testing Algorithm

Data were separated into two groups for the test: one is for clustering using FCM only, the other is subsequent data used in adaptive FCM and alternative adaptive FCM. This testing step are depicted in Fig. 1. There are 2 types of dataset carried out in the test as follows:

6.1 Synthesis data set: It is composed of known class data. This data is separated into 2 groups with 2 attributes. Range of these data is [-3 3] and this data has outlier 1 vector (100, 0). Mean value of group one is (-1.8194, 0.0637), variance (0.4674, 0.5394). Mean values of group two is (-0.1735, 1.9798), variance (0.4969, 0.5274). In testing step this paper entry data into two steps. One entry 90 vectors to basic FCM and entry 11 vectors (has outlier) to adaptive FCM or alternative adaptive FCM step.

6.2 Iris data set: It is comprised of Iris Setosa, Iris Versicolor, Iris Virginica. In testing step this paper entry data into two steps. One entry 120 vectors to basic FCM and entry 30 vectors to adaptive FCM or alternative FCM step.

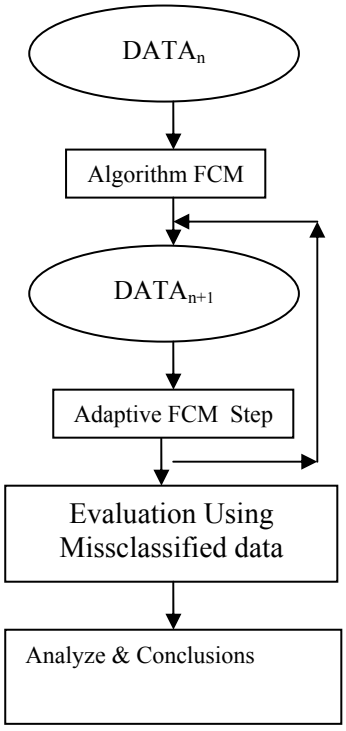

Fig.1: Testing algorithm

7 Results

7.1 Results when using synthesis data

Adaptive FCM with Euclidean distance results in 7 data points of the misclassification, while Mahalanobis distance gives 14 data points of the misclassification. Figures 2 and 3 show these results according to implementing in synthesis data.

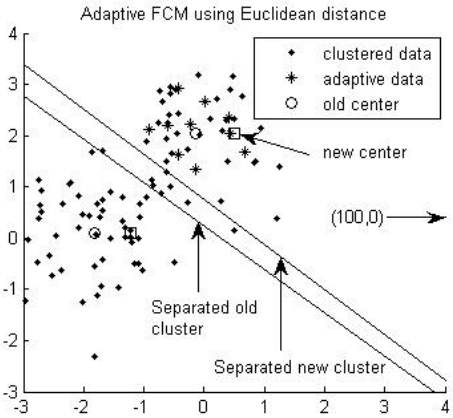

Fig.2 Results from adaptive FCM using Euclidean distance with outlier data.

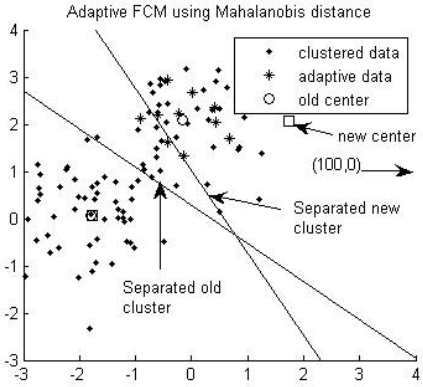

Fig.3 Results from adaptive FCM using Mahalanobis distance with outlier data.

Figure 2 shows two clusters derived from adaptive FCM as applied to with/without adaptive data with outliers. A separated old cluster line isolates two data clusters without adaptive data, while a separated new cluster line concerns adaptive data with outliers. Figure 3 is explained in similar manner, but applied Mahalanobis distance measure.

Figure 4 shows alternative adaptive FCM algorithm, it gives the same line for old and new separated cluster lines.

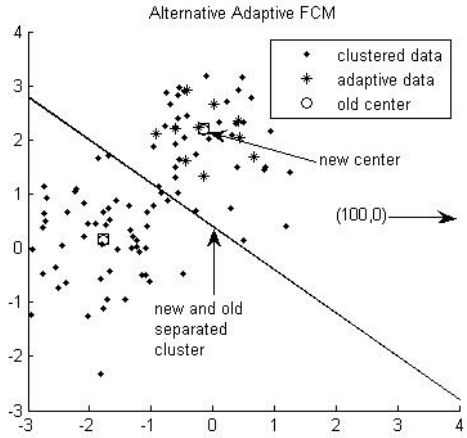

Fig.4 Results from alternative adaptive FCM with outlier.

7.2 Results when using Iris data set

When using adaptive FCM with Euclidean distance, there are 12 missclassified data points as shown in Fig. 6. Adaptive FCM using Mahalanobis distance gives 48 misclassified data points as depicted in Fig. 7.

When applying alternative adaptive FCM, it results 12 misclassified data points as shown in Fig. 8.

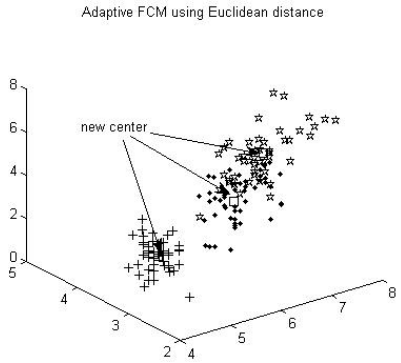

Fig.5 Results from adaptive FCM using Euclidean distance for iris data set.

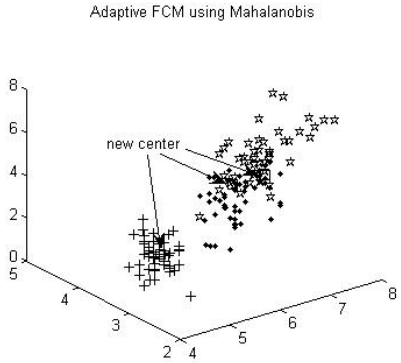

Fig.6 Results from adaptive FCM using Mahalanobis distance for iris data set.

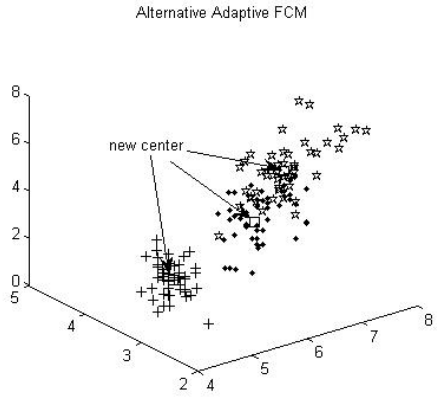

Fig.7 Results from alternative adaptive FCM for iris data set.

8 Conclusions

Alternative adaptive FCM is the promising algorithm to cluster adaptive outlier data due to its robustness. It takes summation β in clustered data and subsequent data into consideration while other do not take care these concepts. Adaptive FCM using Mahalanobis distance does not give a good result in adaptive outliers since it gains distance by using variance-covariance matrix (old and subsequent data) in separation resulting more misclassification. The summary result in details was given in Table 1.

Future work we should to take care Mahalanobis distance to modify by generated co-factor between previous data and adaptive data.

Table 1: *Summary of compared methods (figures shown percentage of errors)*

Missclassified data	Synthesis data	Irisdata set	Iris with outlier data
Adaptive FCM Using Euclidean distance	9%	8%	33%
Adaptive FCM Using Mahalanobis distance	14%	32%	16%
Alternative Adaptive FCM	2%	8%	9%

References:

[1] A.K. Jain, M.N. Murty, P.J. Flyn, Data Clustering: Review *ACM Computing Surveys*, Vol.31 no.3 Sep.1999.

[2] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, Newyork,1981.

[3] Kantardzic M., *Data Mining: Concept Model, Method, and Algorithms,* IEEE Press, USA,2003.

[4] N.Belacel, P. Hansen and N. Mladenovic, "Fuzzy J-Means: a new heuristic for fuzzy clustering" *Pattern Recognition* 35(2002) 2193-2200.

[5] N.Belacel, etc., "Fuzzy J-Means and VNS Methods for clustering Genes from Microarray Data" *Internationl Journal of Bioinformatics*,Oxford University Press. March 2004. NRC 46546.

[6] Peter J.Deer, Peter Eklund, "A study of parameter values for a Mahalanobis Distance fuzzy classifier" *Fuzzy Sets and Systems* 137 (2003) 191-213.

[7] R.De Maesschaluck, D.Jouan-Rimbaud, D.L.Massart, "The Mahalanobis distance" *Chemometrics and Intelligent Laboratory Systems* 50 (2000) 1-18.

[8] Somchai Champathong, et al. "Distance Measure for Fuzzy C-Means Clustering Algorithm" *The Joint Conference on Computer Science and Software Engineering* ,Thailand (2005) 129-136.

[9] Kuo-Lung Wu, Miin-Shen Yang, "Alternative C-means clustering algorithms" *Pattern Recognition* 35 (2002) 2267-2278.

[10] Stefano Marsili-Libelli, Andreas Müller, "Adaptive fuzzy pattern recognition in the anaerobic digestion process" *Pattern Recognition Letters* 17 (1996) 651-659.