Journal of Computational Information Systems 8: 2 (2012) 727–737
Available at http://www.Jofcis.com

Improved Fuzzy C-means Clustering Algorithm Based on Cluster Density

Xiaojun LOU*, Junying LI, Haitao LIU

Shanghai Institute of Micro-system and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China

Abstract

The fuzzy C-means algorithm (FCM) is one of the most popular techniques used for clustering. However, the conventional FCM uses the Euclidean distance as the similarity criterion of data points, which leads to limitation of equal partition trend for data sets. And the clustering performance is highly affected by the data structure including cluster shape and cluster density. To solve this problem, a distance regulatory factor is proposed to correct the similarity measurement. The factor is based on cluster density, which represents the global distribution information of points in a cluster. It is then applied to the conventional FCM for distance correction. Two sets of experiments using artificial data and UCI data are operated, and the results show that the proposed algorithm is superior to the other three clustering algorithms.

Keywords: Clustering; Fuzzy C-means; Regulatory Factor; Cluster Density; Distance Correction

1 Introduction

Cluster analysis is a branch in statistical multivariate analysis and unsupervised machine learning, which has extensive applications in various domains, including financial fraud, image processing, medical diagnosis, and text categorization [1-4]. Clustering is a process of grouping a set of objects into clusters so that the objects in the same cluster have high similarity but are dissimilar with objects in other clusters. The similarity criterion for distinguishing the difference between objects is generally measured by distance. Obviously, we put two data points in the same group if they are close to each other. And data points are from different groups if the distance between them is large. So, the performance of a clustering algorithm is highly affected by the similarity criterion.

The fuzzy C-means (FCM) algorithm is one of the most popular techniques used for clustering [5]. It is a kind of partitional clustering algorithm, which aims at partitioning a given data set into disjoint subsets so that specific clustering criteria are optimized. The conventional FCM method uses Euclidean distance as the similarity measurement of the data points. It has a

*Corresponding author.
Email address: louxjanan@gmail.com (Xiaojun LOU).

January 2012

limitation of equal partition trend for data sets and is only suitable for clustering data groups with hyperspherical shape [6]. In order to improve the performance of conventional FCM, a lot of researches have been done. Wu and Yang [7] replaced the Euclidean norm with a normalized distance function $1 - \exp(-\beta \|\boldsymbol{x_i} - \boldsymbol{v_c}\|^2)$, where β is a positive constant. Yang and Lin [8, 9] separately added a penalty term $w \ln \xi_i$ and a compensated term $\tau \tanh \xi_i$ to distance function. Graves and Pedrycz [10] used $1 - K(\boldsymbol{x_i}, \boldsymbol{v_c})$ as the distance measure, where $K(\boldsymbol{x_i}, \boldsymbol{v_c})$ is a kernel function. Kernel function can be defined in many forms, which could bring different performance. In [11], Mahalanobis distance is used by Xiang. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant. A feature-weighted distance [12, 13] was proposed to improve the performance of FCM, as the fact that the features differ from each other for contribution to clustering. Tsai [14] presented a new distance metric that incorporates the distance variation in a cluster to regularize the distance measure. And a hybrid distance metric [15] represented as the linear combination of several know distance metrics is proposed. That approach could benefit from advantages of different distance. Although there is a lot of study on distance measure, the approaches mentioned above not fully take into account the data distribution, and the shape character of data set is not well used.

In this paper, we study on the distribution of the data set, and introduce a definition of cluster density as the representation of the inherent character of the data set. A regulatory factor based on cluster density is proposed to correct the distance measure in the conventional FCM. It differs from other approaches in that the regulator uses both the shape of the data set and the middle result of iteration operation. And the distance measure function is dynamically corrected by the regulatory factor until the objective criterion is achieved. Two sets of experiments using artificial data and UCI data are operated. Comparing with some existing methods, the proposed algorithm shows the better performance.

The remaining content of this paper is organized as follows: Section 2 introduces the conventional FCM and some improved algorithms in related literature. Section 3 demonstrates the proposed algorithm in detail. The performance of the algorithm is evaluated in Section 4. And in Section 5 concludes this paper.

2 Related Work

In this section, we review the conventional fuzzy c-means clustering algorithm, and give short introduction of two existing improved approaches.

2.1 FCM

The most widely used fuzzy clustering method is the fuzzy c-means clustering, FCM for short [16]. The objective function of the FCM is to obtain a fuzzy c-partition for the given data set by minimizing the objective function J_{FCM}, seeing the following Eq. (1). The constraints related to the fuzzy membership degree are given, as expressed in Eq. (2).

$$J_{FCM}(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} \tag{1}$$

$$s.t. \quad u_{ij} \in [0,1], \sum_{j=1}^{n} u_{ij} > 0, \sum_{i=1}^{c} u_{ij} = 1 \tag{2}$$

The optimal minimum value of J_{FCM} is normally solved by the Alternative Optimization (AO) method. It relates to two updated equation, fuzzy membership degrees and cluster prototypes. The two update equations are given in Eqs. (3) and (4).

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{m-1}}} \tag{3}$$

$$\boldsymbol{v_i} = \frac{\sum_{j=1}^{n} u_{ij}^m \boldsymbol{x_j}}{\sum_{j=1}^{n} u_{ij}^m} \tag{4}$$

The distance between data points is defined as bellows:

$$d_{ij}^2 = \|\boldsymbol{x_j} - \boldsymbol{v_i}\|_A^2 = (\boldsymbol{x_j} - \boldsymbol{v_i})\boldsymbol{A}(\boldsymbol{x_j} - \boldsymbol{v_i})^T \tag{5}$$

Where $\boldsymbol{A}$ is a symmetric positive definite matrix. In the conventional FCM, the distance measurement is usually Euclidean distance, which means $\boldsymbol{A} = \boldsymbol{I}$ in Eq. (5).

The FCM uses the probabilistic constraint that the membership degree of a data point across all clusters sums to 1. The constraint is used to generate the membership degree updated equations for an iterative algorithm. The membership degrees resulting from the FCM and its derivations, however, do not always correspond to the intuitive concept of degree of belonging. Moreover, FCM is only suitable for hyperspherical clusters and has a limitation of equal partition trend for data sets.

2.2 FCM-M

In [17], Cai proposed an improved new algorithm called Fuzzy C-Means based on Mahalanobis distance function (FCM-M), and added a regulating factor of covariance matrix to each class in objective function. The Mahalanobis distance between a data point and a cluster centroid is defined as bellows:

$$\hat{d_{ij}^2} = (\boldsymbol{x_j} - \boldsymbol{v_i})^T \Sigma_{ij}^{-1} (\boldsymbol{x_j} - \boldsymbol{v_i}) \tag{6}$$

Where $\sum_{ij}^{-1}$ is the covariance matrix of $\boldsymbol{x_j}$ and $\boldsymbol{v_i}$.

The Mahalanobis distance is better adapted than the usual Euclidian distance to setting involving non spherically symmetric distributions. It is more particularly useful when multinormal distributions are involved, although its definition does not require the distributions to be normal.

Involving the regulatory factor based on Mahalanobis, the objective function is given by

$$J_{FCM-M}(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\Sigma}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m [(\boldsymbol{x_j} - \boldsymbol{v_i})^T \boldsymbol{\Sigma}_{ij}^{-1} (\boldsymbol{x_j} - \boldsymbol{v_i}) - \ln |\boldsymbol{\Sigma}_{ij}^{-1}|] \tag{7}$$

2.3 FCM-σ

In [14], Tsai and Lin proposed a new metric that takes the distance variation in each clusters as the regularization of the Euclidean distance. The new distance metric is defined as

$$\hat{d}_{ij} = \frac{\|\boldsymbol{x_j} - \boldsymbol{v_i}\|}{\sigma_i} \tag{8}$$

Where σ_i is the weighted mean distance in cluster i, and is given by

$$\sigma_i = \left\{ \frac{\sum_{j=1}^{n} u_{ij}^m \|\boldsymbol{x_j} - \boldsymbol{v_i}\|^2}{\sum_{j=1}^{n} u_{ij}^m} \right\}^{1/2} \tag{9}$$

Different from the Mahalanobis distance, the new distance measure normalizes the distance based on the spread of data points from the centroid in a cluster. It is not normalized with respect to the covariance between features. The FCM-σ searches for C clusters by minimizing the objective:

$$J_{FCM-\sigma}(\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \frac{\|\boldsymbol{x_j} - \boldsymbol{v_i}\|^2}{\sigma_i} \tag{10}$$

3 Proposed Approach

The current existing fuzzy c-means based clustering algorithms either consider only hyperspherical clusters in data space or describe only weighted features. However the density of data points in a cluster could be distinctly different from other clusters in a data set. The conventional distance measure only evaluates the difference between two individual data points. It ignores the global view of the data distribution. Considering the shape of the data set, a distance regulatory factor based on cluster density is proposed to correct the similarity measurement.

3.1 Cluster density

Given a data set $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$, for every data point $\boldsymbol{x_i}$, the dot density is usually defined as:

$$z_i = \sum_{j=1, j\neq i}^{n} \frac{1}{d_{ij}}, \ d_{ij} \leq e, 1 \leq i \leq n \tag{11}$$

Where e is the effective radius for density evaluating. The value of e should be set appropriately according to the application scenarios, because it affects the result directly. The dot density is relatively higher when e is set bigger. To avoid this uncertainty, we can simplify the Eq. (11) when the distribution of data set is uniform and these are no outliers. The new definition is:

$$z_i = 1/\min(\{d_{ij}\}), \ 1 \leq i \leq n \tag{12}$$

We use the inverse of the minimum distance value among the neighbors as the dot density approximately.

However, dot density considers only the local distribution of the data set. Cluster density is proposed to solve this problem. Cluster density is the weighted linear combination of dot densities as expressed in Eq. (13).

$$\hat{z}_i = \frac{\sum_{j=1}^{n} \alpha_{ij} w_{ij} z_j}{\sum_{j=1}^{n} \alpha_{ij} w_{ij}},\ 1 \le i \le c \tag{13}$$

Where α_{ij} is the category label of data point $\boldsymbol{x_j}$, and w_{ij} is the weight of $\boldsymbol{x_j}$.

$\alpha_{ij} = 1$ when $\boldsymbol{x_j}$ most likely belongs to the cluster i, otherwise $\alpha_{ij} = 0$. And w_{ij} is a positive constant which can be adjusted by users. Cluster density $\hat{z}_i$ considers the global shape of the data set in a cluster and uses the dynamical membership degrees in the iteration process. Furthermore, using cluster density instead of dot density can reduce the computation consumption during clustering.

3.2 FCM based on cluster density (FCM-CD)

Using the cluster density, the distance measure is corrected as Eq. (14).

$$\hat{d}_{ij}^2 = \frac{\|\boldsymbol{x_j} - \boldsymbol{v_i}\|^2}{\hat{z}_i},\ 1 \le i \le c, 1 \le j \le n \tag{14}$$

Thus, the optimization expression can be written as follows base on Eqs. ((1),(13) and (14)), as seen in Eq. (15):

$$J_{FCM-CD}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{X}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m \|\boldsymbol{x_j} - \boldsymbol{v_i}\|^2 \frac{\sum_{k=1}^{n} \alpha_{ik} w_{ik}}{\sum_{k=1}^{n} \alpha_{ik} w_{ik} z_k} \tag{15}$$

Applying Lagrange Multiplying Method to Eq. (15), we can obtain the two update equations given in Eqs. (16) and (17).

$$\boldsymbol{v_i} = \frac{\sum_{j=1}^{n} u_{ij}^m \boldsymbol{x_j}}{\sum_{j=1}^{n} u_{ij}^m},\ 1 \le i \le c \tag{16}$$

$$\boldsymbol{u_{ij}} = \frac{\hat{d}_{ij}^{-2/(m-1)}}{\sum_{k=1}^{c} \hat{d}_{kj}^{-2/(m-1)}} \tag{17}$$

Comparing the conventional FCM method, the proposed algorithm (FCM-CD) brings in the cluster density based regulatory factor as the distance correction in similarity measurement. The process of stepwise regression involves of these steps as follows:

1) Choose the number of clusters c, fuzziness index m, iteration error ε, maximum iterations T, and initialize the membership degree matrix $U^{(0)}$.

2) Get the initial centroids using Eq. (17).

3) Calculate the dot density of every data point using Eq. (12).

4) And when the iteration index is $t(t = 1, 2, \cdots, T)$, updating the membership degree matrix $U^{(t)}$ and cluster centroids $V^{(t)}$ using Eqs. (17) and (18).

5) Calculate the value of the objective function $J^{(t)}$ using Eq. (15).

6) If $|U^{(t)} - U^{(t-1)}| < \varepsilon$ or $t = T$, then stop the iteration and get the membership degree matrix U and the cluster centroids V, otherwise set and return to step (4).

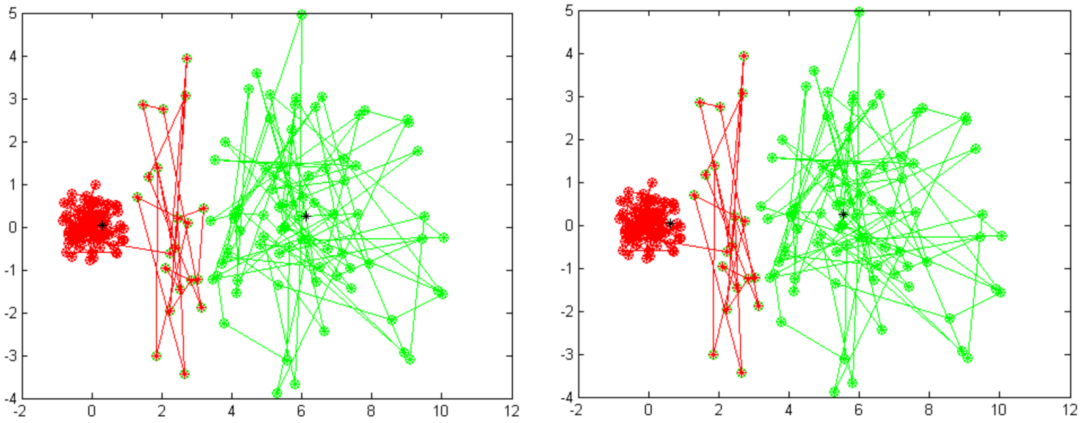
4 Numerical Experiments

This section evaluates the performance of the proposed FCM-CD algorithm through numerical experiments. The experiments include two groups: group one using artificial data sets to analysis the advantage of the FCM-CD and group two using the well-known public UCI data sets to check the ability of the FCM-CD in clustering.

Four clustering methods including the conventional FCM, FCM-M, FCM-σ, FCM-CD, are evaluated and compared with each other in both groups experiments. And we choose the same parameters by setting $m = 2, \varepsilon = 10^{-5}$, and $T = 100$.

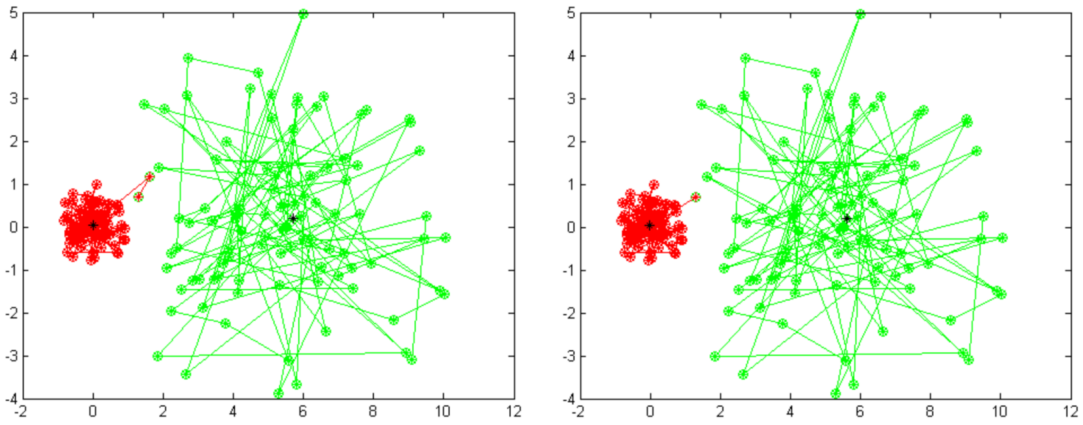
4.1 Group one: artificial data

In order to show the differences of clustering methods visually, artificial data sets are created in a two-dimensional plane for experiments. First, two circular clusters are simulated, as seen in Fig. 1. They are both normally distributed in a circle. The centroids are (0, 0) and (5.5, 0); the radiuses are 1 and 5. The two circles have a small overlapping area. Each cluster has 100 data points.



(a) FCM (b) FCM-M

The clustering results in Fig. 1 reveal that FCM and FCM-M have poor performance when the densities of the clusters are different and FCM-σ and FCM-CD get the higher clustering accuracy



(c) FCM-σ (d) FCM-CD

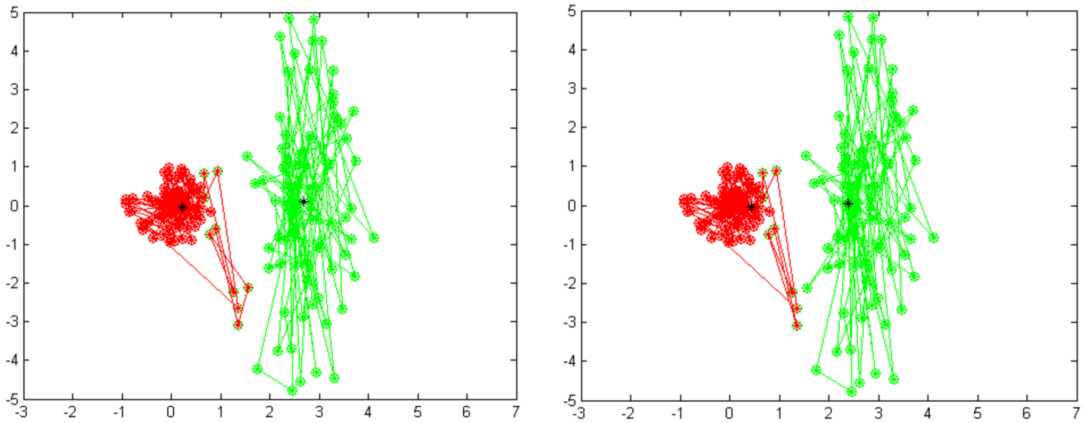
Fig. 1: Two circular clusters with different density (Points in the same color are in the same cluster; the connecting lines represent the clustering results; the black dots are centroids after clustering)

due to the regulatory factor. The clustering errors and absolute deviations of centroids of the four methods are shown in Table 1.

Table 1: Clustering result with two circular clusters

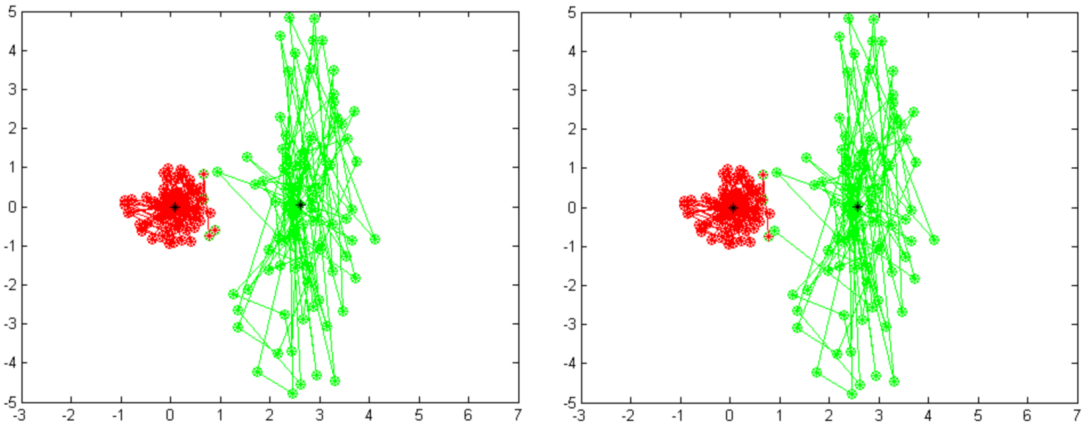
Result	Clustering method			
	FCM	FCM-M	FCM-σ	FCM-CD
Clustering error	10.0%	9.5%	1.0%	0.5%
Centroids deviation	0.97	0.88	0.33	0.26

Then a circular cluster and an elliptic cluster are simulated, as seen in Fig. 2. The centroids are (0, 0) and (5.5, 0); the radius of the circle is 1. The major radius and minor radius of the ellipse are 5 and 2. Each cluster has 100 data points.



(a) FCM (b) FCM-M

Fig. 2 shows that the conventional FCM method has the biggest clustering error when the data set is not spherical distributed. And the proposed FCM-CD method performs best. The errors



(c) FCM-σ (d) FCM-CD

Fig. 2: A circular cluster and a elliptic cluster (Points in the same color are in the same cluster; the connecting lines represent the clustering results; the black dots are centroids after clustering)

of the four clustering methods (FCM, FCM-M, FCM-σ and FCM-CD) are 4.5%, 4.0%, 2.0% and 1.5%, as shown in Table 2. The absolute deviations of centroids are also given in Table 2.

Table 2: Clustering result with a circular cluster and a elliptic cluster

Result	Clustering method			
	FCM	FCM-M	FCM-σ	FCM-CD
Clustering error	4.5%	4.0%	2.0%	1.5%
Centroids deviation	0.46	0.57	0.21	0.15

A good clustering algorithm should be tolerant to different densities and various cluster shapes. The comparing experiments above show that FCM-CD is superior to the FCM and the other two existing FCM based methods.

4.2 Group two: UCI data

In this group of the experiments, the well-known public real data sets are applied. We choose the Iris, Wine and SPECT Heart data sets, which are widely used for classifications and clustering. Table 3 gives the description of the data sets.

Table 3: Data sets description

Data set	Number of clusters	Number of instances	Number of attributes
Iris	3	150	4
Wine	3	178	13
SPECT Heart	2	267	22

Four clustering methods are operated with these data sets and the results are shown in Table 4-6. We give out the number of misclustering in each cluster and the total error rate. We can see that the proposed FCM-CD outperforms the other algorithms in the real data sets, although FCM-M has nearly the same performance. Taking advantages of the mahalanobis distance, FCM-M has some more intrinsic strengths than Euclidean based methods. And the error rates of the Wine data and the SPECT Heart data are relatively higher due to the nature of data sets. These two data sets have high-dimensional attributes, which will easily cause misclusterings without applying suitable feature selection method. And none of the four methods mentioned in this paper have consideration of the feature weight. That's why the error rates reach nearly fifty percent. However, the experiments still can prove that the FCM-CD is superior.

Table 4: Clustering result of Iris data

Clustering method	Misclustering number in cluster 1	Misclustering number in cluster 2	Misclustering number in cluster 3	Error rate
FCM	0	13	3	10.67%
FCM-M	1	6	4	7.33%
FCM-σ	0	11	5	10.67%
FCM-CD	0	9	5	9.33%

Table 5: Clustering result of Wine data

Clustering method	Misclustering number in cluster 1	Misclustering number in cluster 2	Misclustering number in cluster 3	Error rate
FCM	40	25	23	49.44%
FCM-M	37	44	4	47.75%
FCM-σ	19	18	50	48.88%
FCM-CD	33	22	25	44.94%

Table 6: Clustering result of SPECT heart data

Clustering method	Misclustering number in cluster 1	Misclustering number in cluster 2	Error rate
FCM	93	28	45.32%
FCM-M	54	42	35.96%
FCM-σ	90	39	48.31%
FCM-CD	48	41	33.33%

5 Conclusions

As the fact that the conventional fuzzy c-means clustering algorithm is only suitable for hyper-spherical clusters, a improved clustering approach based on cluster density (FCM-CD) is proposed.

Considering of the global dot density in a cluster, a distance correction regulatory factor is built and applied to FCM. Two groups of experiments using artificial data and UCI data are operated. Two other FCM based clustering methods are introduced to compare with the proposed FCM-CD algorithm. The experiment results reveal that FCM-CD has a good tolerance to different densities and various cluster shapes. And FCM-CD shows a higher performance in clustering accuracy.

Acknowledgement

This work is supported by the National Science and Technology Major Projects of China and the Major State Basic Research Development Program of China.

References

[1] R. Filipovych, S. M. Resnick, C. Davatzikos. Semi-supervised cluster analysis of imaging data. *NeuroImage*, 2011, 54 (3): 2185 – 2197.

[2] J. Jiang, W. Li, X. Cai. Cluster behavior of a simple model in financial markets. *Physica A: Statistical Mechanics and its Applications*, 2008, 387 (2 – 3): 528 – 536.

[3] Yan Li, Edward Hung, Korris Chung. A subspace decision cluster classifier for text classification. *Expert Systems with Applications*, 2011, 38 (10): 12475 – 12482.

[4] L. Zhang, L. Y. Zhang, S. Y. Chen, et al. Research on customer classification based on fuzzy clustering. *Journal of Computational Information Systems*, 2007, 3 (5): 1971 – 1976.

[5] Witold Pedrycz, Partab Rai. Collaborative clustering with the use of Fuzzy C-Means and its quantification. *Fuzzy Sets and Systems*, 2008, 159 (18): 2399 – 2427.

[6] X. F. Liu, H. L. Zeng, B. C. Lv. Clustering analysis of dot density function weighted fuzzy c-means algorithm. *Computer Engineering and Applications*, 2004, 40 (24): 64 – 65.

[7] K. L. Wu, M. S. Yang. Alternative C-means clustering algorithms. *Pattern Recognition*, 2002, 35 (10): 2267 – 2278.

[8] M. S. Yang. On a class of fuzzy classification maximum likelihood procedures. *Fuzzy Sets and System*, 1993, 57 (3): 365 – 375.

[9] J. S. Lin. Fuzzy clustering using a compensated fuzzy hopfield network. *Neural Processing Letters*, 1999, 10 (1): 35 – 48.

[10] Daniel Graves, Witold Pedrycz. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. *Fuzzy Sets and Systems*, 2010, 161 (4): 522 – 543.

[11] S. M. Xiang, F. P. Nie, C. S. Zhang. Learning a Mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, 2008, 41 (12): 3600 – 3612.

[12] W. L. Hung, M. S. Yang, D. H. Chen. Bootstrapping approach to feature-weight selection in fuzzy C-means algorithms with an application in color image segmentation. *Pattern Recognition Letters*, 2008, 29 (9): 1317 – 1325.

[13] C. H. Cheng, J. W. Wang, M. C. Wu. OWA-weighted based clustering method for classification problem. *Expert System with Applications*, 2009, 36 (3): 4988 – 4995.

[14] D.M. Tsai, C. C.Lin. Fuzzy C-means based clustering for linearly and nonlinearly separable data. *Pattern Recognition*, 2011, 44 (8): 1750 – 1760.

[15] J. Wang, S. T. Wang. Double indices FCM algorithm based on hybrid distance metric learning. *Journal of Software*, 2010, 21 (8): 1878 – 1888.

[16] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. MA, USA, Springer, 1981.

[17] J. Y. Cai, F. D. Xie, Y. Zhang. Fuzzy c-means algorithm based on adaptive Mahalanobis distance. *Computer Engineering and Applications*, 2010, 46 (34): 174 – 176.