An Insensitivity Fuzzy C-means Clustering Algorithm Based on Penalty Factor

Jiashun Chen [1, 2]
[1]Huaihai Institute of Technology /College of Computer Engineering Lianyungang China
[2]Nanjing University of Aeronautics and Astronautics /College of Information Science and Technology, Nanjing China
Email:1976139@163.com

Dechang Pi [2], Zhipeng Liu [2,3]
[2]Nanjing University of Aeronautics and Astronautics /College of Information Science and Technology, Nanjing China
[3]Nanjing University of Posts and Communications /College of Computer Science and Technology, Nanjing, China
Email: dc.pi@nuaa.edu.cn, Email:liuzhipengcs@139.com

***Abstract*— This paper analyzes sensitivity of Fuzzy C-means to noisy which generates unreasonable clustering results. We also find that Fuzzy C-means possess monotonicity, which may generate meaningless clustering results. Aiming at these weak points, we present an improved Fuzzy C-means named IFCM (Improved Fuzzy C-means). Firstly, we research the reason of sensitivity and find that constraint leads to sensitivity of algorithm, we propose abolish constraint; secondly, we replace membership with typicality for acquiring more reasonable clustering results; finally, we add penalty factor to objective function to avoid monotonicity and coincident clustering results. On the basis of these, we improve objective function and provide step of algorithm. Experiments on various datasets show that new algorithm recognizes noisy data effectively and makes cluster effect improve furthermore.**

***Index Terms*—IFCM, membership, constraint, noisy data**

I. Introduction

Cluster analysis as a branch of data mining has been applied in many fields. The target of cluster is partition sample data set into finite class, and makes intra-class maximal similarity and inter-class maximal difference. More and more clustering algorithms are coming out. Current clustering algorithms are thought to be of two kinds: hard cluster and soft cluster. Hard cluster is an absolute partition. Many algorithms are belong to this type, such as k-means [1], and improved k-means [2-4], and STING [5-6] and so on. K-Means is hard partition, which partition sample into some cluster strictly, that is to say sample only belongs to one cluster. This method hardly copes with some fuzzy problems. After fuzzy set theory was introduced, Bezhdek [7] proposed fuzzy cluster method which partitioned data set softly and abolished the idea of either 0 or 1, and used membership to denote relation between sample and class. In FCM, there are some constraints such as $\sum_{i=1}^{c} u_{ij} = 1, (j = 1, 2 \cdots, n)$, $0 < \sum_{i=1}^{c} u_{ij} < n$ which make clustering results become hard to interpret and understand. Clustering results are sensitive to noisy data and do not recognize outlier. Fig.1 is the Dataset X_{12} [8] distribution on coordinate axis, which shows that sample data x_6 and x_{12} are noisy. The distances from two points to two centroids are different, however, the membership of two points are the same. In fact, membership of x_{12} should far less than x_6, therefore, clustering results are difficult to interpret and understand. FCM is sensitive to noisy and could not recognize outlier.

TABLE 1.
CLUSTER RESULTS PRODUCED BY FCM ON DATASET X_{12}

dataset	U_1	U_2
x_1	0.0636	0.9364
x_2	0.0327	0.9673
x_3	0.0103	0.9897
x_4	0.1006	0.8994
x_5	0.0844	0.9156
x_6	0.5000	0.5000
x_7	0.9156	0.0844
x_8	0.9673	0.0327
x_9	0.9897	0.0103
x_{10}	0.8994	0.1006
x_{11}	0.9364	0.0636
x_{12}	0.5000	0.5000

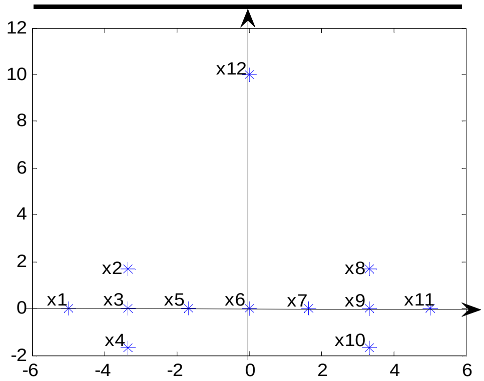

Figure1. Dataset X_{12} distribution on coordinate axis.

Above analysis shows that especial sample data are required for FCM algorithm. FCM could not cope with abnormal data in noisy environment, which show FCM is

Manuscript received December 28,2012; revised February 9,2013; accepted February 28, 2013.
Corresponding author: Chen Jiashun

doi:10.4304/jsw.8.9.2379-2384

sensitive to outlier. In order to overcome deficiency, we present a improved FCM named IFCM. Constraint of FCM has been abolished and we replace membership with typicality. At same time, penalty factor is added to objective function to avoid coincident clustering results and monotonicity. The rest of this paper is organized as follows: in section 2, relate works about FCM are introduced, and section3 present improved algorithm. After discussion of penalty factor, we modify objective function on the basis of FCM. According to minimize objective function, we acquired iterative equation of typicality and cluster center. On the basis of these work, we provide steps of IFCM. Section4, some experiments are conducted in different data sets, and we compare and analyze experimental results. Some conclusions are made in section 5.

II. RELATE WORK

After Bezdek [7] proposed FCM in 1981, as the research moves along, FCM has been applied in many fields. For example, FCM is applied in area of medicine to realize medical image cluster, which contribute to research of classification of diseases [9-12]. Combining FCM with self-organizing network [13] could be used in web log mining. Park et.al [14] applied FCM to deploy and optimize sensor node which improved energy efficiency and reduced cost. Power system applied FCM to forecast long-term load [15].

With increase use of FCM, improved FCM appeared a lot. The improved methods can be divided into two kinds and one is that FCM is improved by changing weight factor or fuzzy factor [16-18]. Literature [16] added weight to improve FCM. Literature [17] discussed that different fuzzy factors influence on membership. Liu [18] proposed an improved method named FCM-SM. They replace Euclidean with Mahalanobis distance to realize different type data set clustering. Unfortunately, in that paper we do not find discussion about sensitivity with noisy data, therefore, this drawback still exits. Former several improved methods change factors of FCM. The other method [19-24] is to combine FCM with other algorithms to realize their property of advantage complementary. He et.al [19] proposed an improved AFSA with adaptive Visual and adaptive step which combined artificial fish swarm algorithm with FCM to enhance outperformance of FCM. They used the ability for searching the global optimum to overcome local optimum and got better experimental results. Hesam [20] had similar target with literature [19] but they combined FCM with PSO to cope with local optimum. Prabhjot[21]et.al identified outliers and separated them from the data-set into one cluster before applying any clustering algorithm. This method scanned neighborhood and compared with threshold to determine outliers, which were similar to DBSCAN [22]. How to determine threshold is needed to further discussion. Executive time and efficiency of algorithm are hardly to ensure. Literature [23-24] combined FCM with fuzzy system to improve algorithm. Tamalika [24] improved FCM based on intuitionistic fuzzy set theory. They construct intuitionistic fuzzy set by using Yagertype[25] intuitionistic fuzzy generator, and replaced membership with intuitionistic membership and acquired intuitionistic membership iterative function. This algorithm was used to cluster medicine image and had good performance. All these algorithms cannot recognize noisy data effectively [26-28].

FCM used membership to denote relation between sample and class. FCM had good performance for some indistinct samples. FCM introduced membership; therefore objective function was modified on the basis of hard partition. Objective function was defined as follows:

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

Where $D_{ij}^2=\left\|x_j-v_i\right\|^2$ was Euclidean distance which denoted distance from jth sample to the ith clustering center. The letter m was fuzzy factor which was used to change membership and benefited to interpret clustering results. FCM was subject to three constraints as follows：

- $u_{ij}\in[0,1]\ \ (i=1,2,\cdots,c;\ j=1,2,\cdots,n)$
- $\sum_{i=1}^{c}u_{ij}=1\ \ (j=1,2,\cdots,n)$
- $0<\sum_{i=1}^{c}u_{ij}<n\ \ \ (i=1,2,\cdots,c)$

Membership and clustering center iterative function were acquired by minimizing objective function.

$$u_{ij}=1\Bigg/\sum_{k=1}^{c}\left(\frac{D_{ij}^2}{D_{kj}^2}\right)^{\frac{1}{m-1}} \tag{2}$$,

$(i=1,2,\cdots,c;\ j=1,2,\cdots,n)$

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}}\quad(i=1,2,\cdots,c) \tag{3}$$

The main steps of FCM are given as follows

Step1:initializing parameters m and ε , and the maximal iterative times ITER_TIME, and setting the number of class c, generating randomly initial membership matrix U^0;

Step2: computing clustering center by using formula (3);

Step3: amending membership according to formula (2);

Step4: if the number of iterative times is great than ITER_TIME or $\left|J^l-J^{l-1}\right|<\varepsilon$, then stop, or go to step2, and the number of iterative times add one.

Experimental results show that FCM could not recognize noisy data, and clustering results are not consistent to the fact, and hard to understand and interpret. In order to overcome these drawbacks, we present an improved FCM based on penalty factor.

III. IMPROVED FUZZY C-MEANS(IFCM)

A. Definition Objective Function

From above analysis, we know the constraints are the main reason that FCM is sensitive to noisy and has bad performance. If we abolish constraints, according to the requirement of convergence, objective function $J(U)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}D_{ij}^{2}$ has minimal value when $u_{ij}=0$. Obviously, it is meaningless. We define a new objective by replacing membership with typicality and adding penalty factor to avoid monotonicity and coincident clustering results.

Definition1: Supposing sample $X=\{X_1,X_2,\cdots,X_n\}\subset R^p$, R^p denote p _dimensional real vector space, if a fuzzy c partition is the optimum partition, and then exiting $T=[t_{ij}]_{c\times n}$ and $V=(v_1,v_2,\cdots,v_n)$ and $c(1\le c\le n-1)$ satisfy the below (formula 4)function for any T,V.

$$J(T,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}(t_{ij}^{m}D_{ij}+\lambda t_{ij}^{m}\log t_{ij}^{m}) \qquad (4)$$

Where m is fuzzy factor, and λ is a constant, and $1\le i\le c, 1\le j\le n$. we replace u_{ij} in FCM with t_{ij} which is the typical value that denotes the jth sample belonging to the ith class. $T=\left[t_{ij}\right]$ is typicality matrix. We abolish constraint $\sum_{i=1}^{c}u_{ij}=1$,which makes clustering results be insensitive to noisy and be easier to interpret and understand. With the number of class increase, the first part $\sum_{i=1}^{c}\sum_{j=1}^{n}(t_{ij}^{m}D_{ij})$ in formula 4 is monotonous. The second part $\lambda t_{ij}^{m}\log t_{ij}^{m}$ is introduced to avoid this monotonicity. According to the requirement of minimize objective function, we derive two sides of formula 4, and make $\frac{\partial J(T,V)}{\partial t}=0$,and then get some formulas as follows:

$$\frac{\partial J(T,V)}{\partial t_{ij}}=mt_{ij}^{m-1}D_{ij}^{2}+\lambda m(mt_{ij}^{m-1}\log t_{ij}^{m}+t_{ij}^{m}\cdot\frac{1}{t_{ij}})=0 \qquad (5)$$

$$\Rightarrow mt_{ij}^{m-1}D_{ij}^{2}+\lambda mt_{ij}^{m-1}(m\log t_{ij}+1)=0$$

$$\Rightarrow \log t_{ij}=\frac{-D_{ij}^{2}}{m\lambda}-\frac{1}{m}$$

$$\Rightarrow t_{ij}=\exp(\frac{-D_{ij}^{2}}{m\lambda})\cdot\exp(-\frac{1}{m}) \qquad (6),$$

$$V_i=\frac{\sum_{j=1}^{n}t_{ij}^{m}x_j}{\sum_{j=1}^{n}t_{ij}^{m}} \qquad (7)$$

Theorem1: penalty factor that is introduced to objective function can effectively penalize monotonicity of objective function.

Proof: Because with the number of class increase the first part $\sum_{i=1}^{c}\sum_{j=1}^{n}(t_{ij}^{m}D_{ij})$ of objective function decreases monotonously. $\because t_{ij}^{m}\log t_{ij}^{m}\le 0$, $\therefore t_{ij}^{m}\log t_{ij}^{m}\le 0$, $\therefore$ the second part $\lambda t_{ij}^{m}\log t_{ij}^{m}\ge 0$ can restrain to decrease while $\lambda<0$;Conversely, $\lambda>0$ can restrain to increase. In conclusion, $\lambda t_{ij}^{m}\log t_{ij}^{m}\ge 0$ penalize monotonicity of objective function.

B. Improved FCM

The main steps of improved FCM as follows:

Algorithm1:

Input: initializing parameters m,ε,λ and typicality matrix T^0;

Output: typicality matrix T and cluster center V;

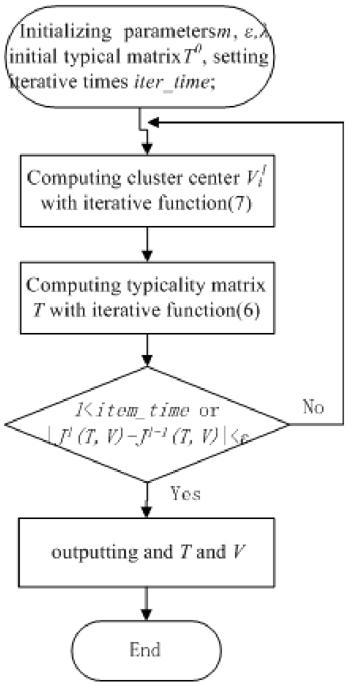

Figure2. Flow chart of algorithm

IV. Results and Analysis

Experiment 1:

Dataset:Artificial experimental dataset

Algorithm:FCM,IFCM

Dataset in experiment is an artificial data which is generated randomly by using Gaussian distribution. Dataset can be divided two classes, the first class is that mean is 120 and variance is100; the second class is that mean is 90 and variance is 100. In each class, there are 15 data points, and each point includes two attributes.

Clustering results generated by FCM and IFCM are show in Figure3. From Figure3 there are two clear classes, therefore we know that FCM and IFCM have good performance on dataset. In order to test sensitivity of two algorithms to noisy, we add some noisy data to original dataset. Figure4 shows the clustering results by FCM after adding noisy. Clustering centers of two classes have obviously deviated from original position in Figure4. Clustering results generate a lot of resubstitution errors by FCM. FCM is sensitive to noisy. By contrast, clustering results in Figure5 are better than in Figure4. We also recognize two classes but clustering center do not deviate. Although there are some incorrect data points in Figure5 but the number is far less than in Figure4. IFCM evidently lowered insensitivity, and has better performance.

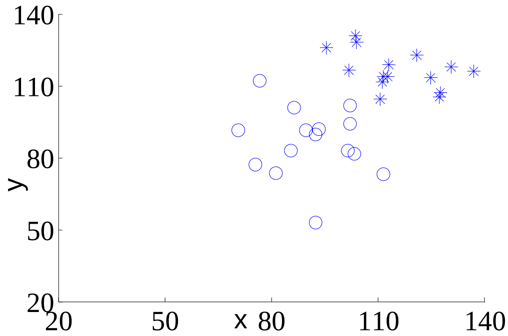

Figure3.Cluster results produced by FCM, IFCM on artificial dataset

TALBE2.
THE NUMBER AND RATE OF RESUBSTITUTION ERROR PRODUCED BY FCM, IFCM

algorithms	FCM	IFCM
resubstitution error	10	2
Rate error	66.67%	13.33%

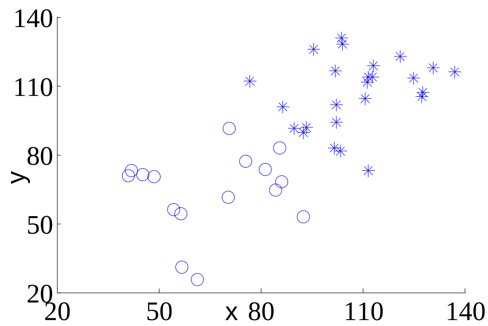

Figure4.Clustering result produced by FCM on artificial dataset with noisy data

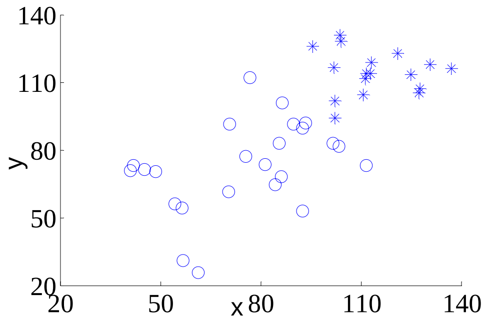

Figure5.Clustering result produced by IFCM on artificial dataset with noisy data

Table2 shows that the number and rate of resubstitution error produced by FCM are greater than IFCM. IFCM overcomes sensitivity to noisy, and get good results.

Experiment2:

Dataset: X_{12}

Algorithm:FCM, IFCM

Dataset X_{12} includes twelve data points and each data point has two attributes. Talbe1 shows coordinate value of each point. Except for data point x_6 and x_{12}, the other ten points can be divided two classes which distribute two sides of y coordinate. Fig.1 shows distribution on the axis. We easily identify that data point x_6 and x_{12} are noisy. Distances from x_6 to two cluster centers are equal but less than distance from x_{12} to two cluster centers. Setting ε =0.000001 and $iter\_time$ =100, Cluster_n=2.

Table3 show the membership produced by FCM and IFCM. As a whole, IFCM is better than FCM about dataset partition. From table3,we know that typicality of data points $\mathbf{x}_3$ 和 $\mathbf{x}_9$ are equal to one,because in real data set these two points are cluster center, which shows that IFCM is easier to find cluster center. Data points $\mathbf{x}_6$ and $\mathbf{x}_{12}$ are noisy. Membership of two data points are the same, and do not recognize them as noisy. However, in table3, typicality of these two data points are different, and $T(x_6) = \mathbf{0.7383}$, $T(x_{12}) = \mathbf{0.3730}$. Clustering results produced by IFCM are easier to understand and interpret. In conclusion, IFCM is insensitive to noisy and gets better clustering results.

TABLE3.
MEMBERSHIP AND TYPICALITY PRODUCED BY FCM AND IFCM ON X_{12}

Data set	FCM		IFCM	
	U_1	U_2	T_1	T_2
$\mathbf{x}_1$	0.0636	0.9364	0.4160	0.9271
$\mathbf{x}_2$	0.0327	0.9673	0.4598	0.9891
$\mathbf{x}_3$	0.0103	0.9897	0.4659	1.0000
$\mathbf{x}_4$	0.1006	0.8994	0.4561	0.9438
$\mathbf{x}_5$	0.0844	0.9156	0.5527	0.9961
$\mathbf{x}_6$	0.5000	0.5000	0.7383	0.7383
$\mathbf{x}_7$	0.9156	0.0844	0.9961	0.5527
$\mathbf{x}_8$	0.9673	0.0327	0.9891	0.4598
$\mathbf{x}_9$	0.9897	0.0103	1.0000	0.4659
$\mathbf{x}_{10}$	0.8994	0.1006	0.9438	0.4561
$\mathbf{x}_{11}$	0.9364	0.0636	0.9271	0.4160
$\mathbf{x}_{12}$	0.5000	0.5000	0.3730	0.3730

V. CONCLUSION

This paper discuss sensitive problem about FCM algorithm, and propose an improved Fuzzy C-means named IFCM. We firstly analyze the main reason of sensitivity in FCM and find that constraints make clustering results be unreasonable. Therefore, improved algorithm abolished constraints, and replace membership with typicality. In order to overcome monotonicity and coincident clustering results, we add penalty factor to objective function. On the basis of these, we construct objective function. At last, we provide the main step of IFCM. In the last part, some experiments are on different data sets. Experiment one on data set generated by Gaussian distribution show that IFCM recognizes the number of class effectively and acquires better clustering result and has less resubstitution errors. Experiment two shows that IFCM distinguishes noisy from data set and is easier to find cluster center. New algorithm has good performance in noisy environment, is apt to find outlier. This work is just the first step, and there are many challenging issues discussed above. We are currently investigating into detailed issues as a further study.

ACKNOWLEDGEMENTS

This research is sponsored by Qing Lan Project and National High Technology Research and Development Program of China (863 Program, No. 2007AA01Z404).

REFERENCES

[1] J.A.Hartigan, M. A.Wong, A K-Means Clustering Algorithm. Journal of the Royal Statistical Society,Series C (Applied Statistics), vol.28(1) , 1979 ,pp.100-108.

[2] L. Yi, S.Y. Lu, F. Fotouhi, Y.P. Deng, S.J. Brown, FGKA: a Fast Genetic K-means Clustering Algorithm. Proceedings of the 2004 ACM symposium on Applied computing, pp.622-623,2004.

[3] L.Wei, Modified K-Means Clustering Algorithm. Image and Signal Processing, vol.4, 2008, pp.618 – 621.

[4] V. R.Patel, R.G.Mehta, Modified k-Means Clustering Algorithm.Computational Intelligence and Information Technology, vol.250,2011,pp. 307-312.

[5] W.Wang, J.Yang, and R. Muntz, STING+: an approach to active spatial data mining.15th International Conference on Data Engineering,pp.116 – 125,1999.

[6] J.S.Chen, D.C.Pi, A Similar Sub-trajectory-based algorithm for Moving Object Trajectory Clustering. Information. Vol.15(4), 2012,pp.1645-1662.

[7] J. C.Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.

[8] N.R. Pal, K.Pal, J.C. Bezdek, A new hybird c-means clustering model. Proceedings of the IEEE International Conference on Fuzzy Systems, Piscataway: IEEE Press, vol.1, pp.179-184,2004.

[9] K.Dervis, O.Celal, Fuzzy clustering with artificial bee colony algorithm.Scientific Research and Essays, vol.5(14), 2010,pp.1899-1902.

[10] Z.X.Ji, Y.Xia, Q.Chen, Q.Sun, , Xia,D.H., Fuzzy c-means clustering with weighted image patch for image segmentation. Applied Soft Computing, vol.12, 2012,pp.1659–1667.

[11] X.Li, X.Lu, J.Tian, P.Gao, H.G.Kong, G.W.Xu, Application of Fuzzy c-Means Clustering in Data Analysis of Metabolomics. analytical chemistry,vol.81(11),2009, pp.4468–4475.

[12] H.Ozyavru, N.Ozkurt, S.Men, Segmentation of multiple sclerosis plagues by robust fuzzy clustering with spatial information. International Symposium on Innovations in Intelligent Systems and Applications, pp.420 – 423, 2011.

[13] X.Zhang, L.Z.Duan, Y.Q.Zhan, G.F.Wang, Research of Algorithm Based on Improved Self-Organization Neural Network of Fuzzy Clustering.Advances in Intelligent and Soft Computing, vol.139, 2012,pp.191-197.

[14] Y.K.Park, M.G.Lee, K.K.Jung, J. J.Yoo, S. H Lee, H.S. Kim, Optimum Sensor Nodes Deployment Using Fuzzy C-Means Algorithm. International Symposium on Computer Science and Society. pp.389-392,2011.

[15] C.W.Zhang, Z.G.Yang, Improved fuzzy clustering algorithm in Long-Term load forecasting of power system..International Conference on Computer Science and Information Technolog,vol.9 pp.556-560,2010.

[16] L.Liu, J.Z.Zhou, X.L.An, Y.H.Li, Q.Liu, Improved Fuzzy Clustering Method Based on Entropy Coefficient and Its Application. Lecture Notes in Computer Science, vol.5264, 2008,pp.11-20.

[17] H. F.Li, F. L.Wang, S. J.Zheng, L.Gao, An Improved Fuzzy C-Means Clustering Algorithm and Application in Meteorological Data. Journal Advanced Materials Research, vol.181,2011, pp.545-550.

[18] H.C.Liu, B.C.Jeng, J.M.Yi, Y.K.Yu, Fuzzy C-Means Algorithm Based on Standard Mahalanobis Distances.Proceedings of the 2009 International Symposium on Information Processing. pp.422-427,2009.

[19] S.He, N.Belacel, H.Hamam, Bouslimani Y, Fuzzy Clustering with Improved Artificial Fish Swarm Algorithm. International Joint Conference on Computational Sciences and Optimization, vol.2, pp.317-321,2009.

[20] H. Izakian, A.Abraham, Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. Expert Systems with Applications, vol.38, 2011, pp.1835–1838.

[21] P.Kaur, A.Gosain, Improving the performance of fuzzy clustering algorithms through outlier identification. International Conference on Fuzzy Systems, 2009, pp.373-378.

[22] M.Ester, H.P.Kriegel, J.Sander,A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. International Conference on Knowledge Discovery and Data Mining, pp.226-231,1996.

[23]A. Celikyilmaz, and I. B. Turksen. Enhanced Fuzzy System Models With Improved Fuzzy Clustering Algorithm.IEEE TRANSACTIONS ON FUZZY SYSTEMS, vol.16(3), 2008, pp.779 – 794.

[24] T.Chaira, A novel intuitionistic fuzzy C means clustering algorithm and its application tomedical images. Applied Soft Computing, 2011,vol.11,pp.1711–1717.

[25] P.Burillo, H.Bustince, Entropy on intuitionistic fuzzy set and on interval-valued fuzzy set, Fuzzy Sets and Systems, vol.78, 1996, pp.305–316.

[26] C.Y,Lui, X.Q.Zhang, X.F.Li, Y.N.Liu, J.Yang, Gaussian Kernelized Fuzzy c-means with Spatial Information Algorithm for Image Segmentation. Journal of computers, vol. 7(6), 2012,pp. 1511-1518.

[27] S.D.Li, X.Q.Lv, C.L.Zhan, S.C.Shi, Regional Division of Police Patrols Based on Adaptive FCM Clustering and MMAS. Journal of computers, vol.6 (2), 2011,pp.313-320.

[28] X.H.Hu, T.Mu, W.H.Dai, H.Z.Hu, G.H.Dai, Analysis of Browsing Behaviors with Ant Colony Clustering Algorithm. Journal of computers, vol. 7(12), 2012, pp. 3096-3102.

Chen Jiashun was born in 1976. He received M.S. degree in computer science and technology from China University of geosciences, in 2006. Now, he is a Ph.D. candidate in Nanjing University of Aeronautics and Astronautics. His research interests are data mining and distributed

systems.

Pi Dechang was born in 1971. He was a Ph.D. of Nanjing University of Aeronautics and Astronautics (NUAA) of China and now he is a professor and Ph.D. supervisor in NUAA. His research interests are data mining and database systems etc.

Liu Zhipeng was born in 1980. He received the B.S. and M.S. degree in computer science and technology from Nanjing University of Posts and Telecommunications, in 2002, 2005, respectively. Now, he is a Ph.D. candidate in Nanjing University of Aeronautics and Astronautics. His research interests are data mining and distributed systems.