Study of Clustering Algorithm based on Fuzzy C-Means and Immunological Partheno Genetic

Hongfen Jiang, Yijun Liu, Feiyue Ye,Haixu Xi, Mingfang Zhu
College of Computer Engineering, Jiangsu Teachers University of Technology, Changzhou 213001, China
E-mail: jdhsff@163.com, {lyj, yfy, xihaixu, mfzhu2009}@ jstu.edu.cn

Junfeng Gu
College of Petroleum Engineering, Changzhou University, Changzhou 213001, China
E-mail: gujfemail@126.com

***Abstract*—Clustering algorithm is very important for data mining. Fuzzy c-means clustering algorithm is one of the earliest goal-function clustering algorithms, which has achieved much attention. This paper analyzes the lack of fuzzy C-means (FCM) algorithm and genetic clustering algorithm. Propose a hybrid clustering algorithm based on immune single genetic and fuzzy C-means. This algorithm uses the fuzzy clustering of Immune Partheno-Genetic to guide the number and the choice of the clustering centers. And then utilize FCM to make the clustering (IPGA-FCM). This algorithm not only overcomes the local optimal problem of FCM, the choice of the initial value is inappropriate, but also overcomes the contradictions between the search speed and clustering accuracy of the general genetic clustering algorithm. Then it applies the novel clustering algorithm to Chinese document clustering. The clustering algorithm is superior to other ordinary clustering algorithm and the result can embody the wide diversity and large amount of Chinese document. Experiments show the algorithm is effective.**

***Index Terms*—Clustering analysis, genetic algorithm, FCM, Immune mechanism, Text clustering**

I. INTRODUCTION

In practical work, classification is very important. Using the fuzzy mathematics method, fuzzy clustering analysis separates the similar things, and categorizes them. Fuzzy C-means (FCM) algorithm, first proposed by Dunn and promoted by Bezdek, is commonly used in the fuzzy clustering algorithm. FCM has been quoted in a variety of areas, such as pattern recognition, image processing, fuzzy modeling, and medical diagnosis. However, FCM algorithm is a local search algorithm. If the initial value is chosen properly, it will converge to the local minimum point [1]. Genetic algorithm is one of evolutionary algorithm, which has a strong global parallel search capability, especially in the occasion of large quantities, higher classification requirements, parallel computing required, the method shows the superiority that FCM can not match. However, although theoretically the method can overcome the local optimum and can find the global optimal solution. But it has the obvious drawback of “premature convergence” and the weak local searching ability [2].

To overcome the deficiencies of FCM and genetic algorithm, genetic algorithm (GA) was proposed combined with the FCM algorithm to solve the problem of local optimum [3], but the performance of the algorithm was still poor and its requirements to the field knowledge was strict, and showed convergence. To this end, someone utilized the genetic algorithm which combined immune genetic algorithm (IGA) and C-means algorithm [4]. Immune genetic algorithm was proposed recently, which was based on organism immune mechanism. It correspond the problem to be solved to the antigen, and the corresponding solution of the problem to the antibody. After combining antibody and antigen, a series of reactions occur against antigens. So the diversity of individuals in genetic algorithm can be ensured to improve the global search ability of genetic algorithm. It better solves the late precocious appearance of the general genetic clustering algorithm and can effectively converge to global optimal solution [5-7]. However, due to the introduction of immune mechanism, the evolution of each generation will increase the amount of calculation, which will extend the calculation time. Meanwhile, it is found that a suitable choice of threshold value ε was also difficult. As large, the genetic number is often not enough, which would prematurely terminate the algorithm; and small, often hereditary generations waste too much computation time, even sometimes algorithm does not terminate in the operation.

Gao jian proposed the partheno genetic algorithm(PGA) based on immune mechanism to solve the optimization problem of data clustering [8]. Compared to the simple genetic algorithm, the partheno genetic had the following advantage: initial population was not required, there were not “premature convergence” problem, and simple operation. It was found that immune partheno genetic clustering algorithm could find the optimal solution close to global through less genetic number. Therefore we firstly utilize the immune partheno genetic clustering algorithm as initialization to find the optimal solution

---

Manuscript received May 1, 2012; revised June 24, 2012; accepted July 1, 2012.

doi:10.4304/jsw.8.1.134-141

close to global. Then use FCM algorithm to make solution. The evolution of each generation needs a longer computation time, but can decrease the evolution generations in the initialization phase. This method can not only resolve the local optimization problems, but also utilize the advantages of the FCM to improve the integral speed of convergence.

II. FUZZY C-MEANS ALGORITHM (FCM)

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m(\mu, v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d^2{}_{ik}, \quad 1 \le m < \infty \tag{1}$$

where $d_{ik} = \|x_k - v_i\|$ is any norm expressing the similarity between any measured data and the center, n is number of samples, c is a given number of categories, and $1 < c < n$, m is any real number greater than 1, v_i is the d-dimension center of the cluster, $\mu_{ik} \in [0,1]$ is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data,

and $\sum_{i=1}^{c}\mu_{ik} = 1, k = 1, 2, ..., n$.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership uik and the cluster centers vi by equation (2) and equation (3):

$$\begin{cases} \mu_{ik} = \dfrac{1}{\sum_{j=1}^{c}[\dfrac{d_{ik}}{d_{jk}}]^{\frac{2}{m-1}}} \quad if \quad d_{ik} \ne 0 \\ \mu_{ik} = 1 \quad if \quad d_{ik} = 0; i = 1, 2, ..., c; k = 1, 2, ..., n \\ \mu_{ik} = 0 \quad if \quad d_{ik} = 0, k \ne j \end{cases} \tag{2}$$

$$v_i = \frac{\sum_{k=1}^{n}(\mu_{ik})^m x_k}{\sum_{k=1}^{n}(\mu_{ik})^m}, i = 1, 2, ..., c \tag{3}$$

When $\max_{ik}\{|u_{ik}{}^{(j+1)} - u_{ik}{}^{(j)}|\} < \varepsilon$, this iteration will stop. ε is a termination criterion between 0 and 1, whereas j is the iteration steps.

But this procedure converges to a local minimum or a saddle point of J_m, if the choice of the initial clustering center is improper. On the other hand, as the number of class is large, FCM algorithm must repeat several times to get a satisfactory solution, and sometimes even no satisfactory solution.

III. HYBRID CLUSTERING ALGORITHM DESIGN

A. Encoding and Fitness Function

The performance of genetic algorithm is greatly affected by the choice of encoding mechanism. We use the encoding method of clustering center to encode the cluster prototype matrix P. Make the parameters of the cluster prototype connecting. According to their own range, a gene string, that is $b = \{\beta_1, \beta_2, ..., \beta_p, ..., \beta_i, ..., \beta_l, ..., \beta_{cp}\}$, is encoded by the quantify value. Each cluster prototype V_i has a corresponding set of parameters.

If 65536 degree encoding is used for each parameter, each gene string is composed by the string of $16 \times p \times c$ bit. At this point the search space is $2^{16 \times p \times c}$ points. More quantity series and higher precision, the computation is greater. When data volume is big, there’s only relation with p and c, and no with the data volume. Taking account into the convenience of encoding and decoding, we use binary encoding for each parameter. Where x is decimal form of arbitrary one-dimensional of the cluster center v_i, and $x \in (-b, b)$. x is transformed into sixteen bit binary form y according to the following formula. Where $y \in (0, 65536)$.

$$y = \frac{(b + x) \times (2^{16} - 1)}{2b} \tag{4}$$

According to the following formula when decoding:

$$x = -b + \frac{2b \times y}{(2^{16} - 1)} \tag{5}$$

The fitness function should be designed to meet the following conditions:

(1) Single-valued, continuous, non-negative and maximum.
(2) Reasonable and consistent.
(3) Less computation.
(4) Well versatility.

Therefore the optimal clustering results, based on the objective function, correspond to the minimum values of the objective function. That is the more effective of the clustering, the smaller of the objective function, and the bigger of the fitness. The fitness function $f(.)$ can be defined by the objective function $J_m(\mu, \upsilon)$.

$$f = \frac{1}{J_m(\mu, v) + \xi} \tag{6}$$

ξ is a given constant. According to the design requirement, the $\xi = 1$ and the fitness function is the following formulation.

$$f = \frac{1}{J_m(\mu, v) + 1} \tag{7}$$

B. Immune Operator Choice

The choice of vaccination adopts adaptive suppression based on individual concentration and promoting selection strategy [9]. This method has the advantage: the larger value of individual fitness, the choice probability is more. It is guaranteed that the individuals with big value of fitness can be kept in the evolutionary group. Accelerate the convergence of the algorithm. The higher of individual concentration, the probability of selection is small. The diversity of individuals is ensured in the evolutionary group.

C. Partheno Genetic Algorithm Operation

PGA defines the genetic arithmetic operators, such as gene selection, gene transposition, gene translocation and gene variation [10]. Where, the selection of arithmetic operators is same with GA. The remaining operators are defined as follows:

a) Definition I: Gene transposition operators is some probability p_c, based on witch exchange the gene of some position on a chromosome. The location of the gene be exchanged is randomly generated.

b) Definition II: Gene translocation operators is some probability p_s, based on witch turn some gene of the substring on a chromosome backward one by one. And the gene of the last position of the substring shifts to the front position. The substring and its length are randomly generated.

c) Definition III: Gene variation operators is some probability p_i, based on witch some gene of the substring on a chromosome is randomly re-assigned. The substring and its length are randomly generated.

D. Algorithm Flow

/*Initialize immune partheno genetic and find the approximate solution close to global */

- Initialize each parameter: fuzzy index m, precision requirement ε, group size M, and crossover probability P_c, α;
- When $t=0$, initial group is generated;
- Calculate individual fitness $f(M_i(t))$ and concentration d;
- If $t \ge \max\_gen$, decode the best individual, and turn 6th step. Else continue.
- According to $p = \alpha p_f + (1-\alpha) p_d$, duplicate 85% to 95% from big to small. And individual is randomly generated to add to the group to make the scale remain as M. Obtain transition matrix $M'(t)$;
- Carry on the partheno genetic operation, and produce the next generation group $M(t+1)$. Turn to 2nd step;

/* solving by using of FCM */

- Calculate the cluster prototype $\{v_i, 1 \le i \le c\}$, $p=0$;
- Recalculate the membership matrix U;
- Calculate object function J^P;
- If $\left|J^P - J^{P-1}\right| < \varepsilon$, then export U, J, $\{v_i, 1 \le i \le c\}$. Algorithm terminates. Else turn to 6th step.

IV. EXPERIMENTAL RESULTS

Experiment depends on two set of data separately by FCM, GA, IGA, PGA-FCM (initialized by PGA firstly, and then soled by FCM), and IPGA-FCM (initialized by IPGA firstly, and then soled by FCM), which is proposed in this paper. Select fuzzy index $m=2$, precision requirement $\varepsilon = 10^{-6}$, group size $M=70$, crossover probability $P_c = 0.8$, mutation probability $P_v = 0.01$, and $\alpha = 0.65$.

A. Data I

Select standard IRIS data as the test sample data set. This data set is already family by all, and it is widely used as reference data for pattern recognition. IRIS data set consists of 150 sample points in four-dimensional space. The total sample set includes three kinds of IRIS: Setosa (A1), Versicolor (A2), and Virginica (A3). And each has fifty samples. Distinction between class A1 and A2, A3 is obvious, and there is some overlap between A2 and A3. It is not easy to separate A2 from A3.

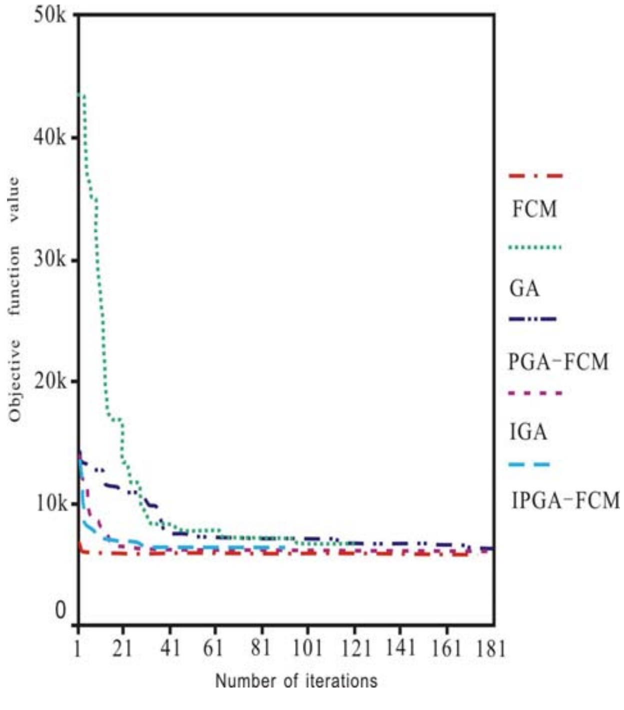

Figure 1. Variation curves of the clustering objective function value on data I

Make fuzzy clustering experiment on the IRIS data, and the Figure1 shows the variation curves of the objective function value of five fuzzy clustering algorithms with the changes of the iteration number. The curves in Fig. 1 are all the average results of the thirty tests. Although on the end of the algorithm, the objective function value is very close, the convergence rate of the

IPGA-FCM algorithm is faster than the other four methods.

Make experiments 30 times on the IRIS data, respectively using four kinds of algorithms. The average objective function value and the classification accuracy of the four algorithms are showed in Table 1.

TABLE 1.
COMPARISON DATA OF CLUSTERING RESULTS

Algorithm	Objective Function value	Number of misclassification	Classification
FCM	6050.57	16	89.33%
GA	6345.51	12	92%
IGA	6050.57	12	92%
PGA-FCM	6050.57	12	92%
IPGA-FCM	6050.57	10	93.33%

B. Data II

The data used by literature 11th is nineteen indicators of eleven sample data of the actinolite of some iron ore area [11]. And the data should be divided into 5 classes. Make experiments and find that the result of this literature is not the global optimum. The optimum classifications are the following: {1,3,4,6,7,8,11,15}, {2,16,17}, {9,10,12,13}, {5,18}, {14,19}. Make clustering on this data by five algorithms separately ten times. Using FCM clustering algorithm 10 times, obtain the global optimal clustering result 3 times and the local optimum 7 times. All get the best clustering results by using partheno genetic algorithm and the algorithm in this paper. But the calculation speed of the latter is faster than the former.

Make fuzzy clustering experiment on data II, and the Figure.2 shows the variation curves of the objective function value of the three fuzzy clustering algorithms, that is FCM, PGA-FCM, and IPGA-FCM, with the changes of the iteration number. The curves in Figure.2 are all the average results of the ten tests. The average objective function minimum of IPGA-FCM clustering algorithm is much smaller than that of FCM and PGA-FCM.

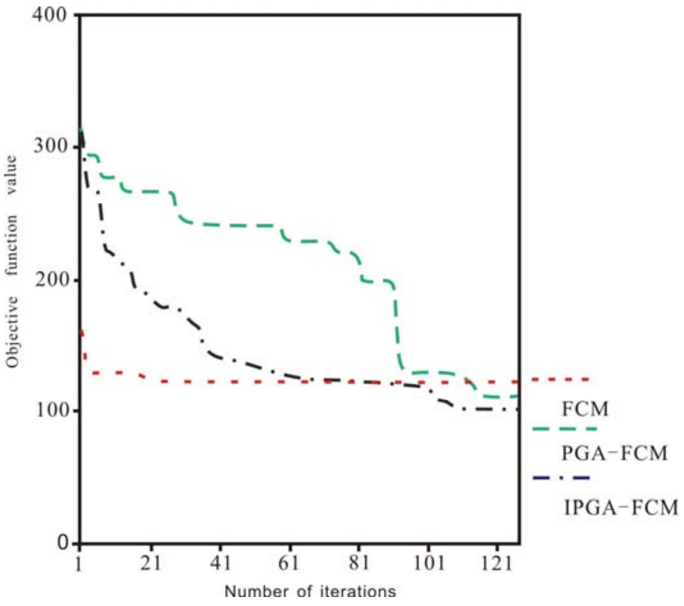

Figure 2. Variation curves of the clustering objective function value on data II

C. Data III (the former 3 dimension of data II)

Make clustering on this data 10 times by the four algorithms and utilize FCM clustering algorithm 10 times. Obtain the global optimal clustering result 3 times as showed in Figure.4, and get the local optimal result 7 times as showed in Figure.5. However get the best clustering results 10 times by using the algorithm proposed in this paper.

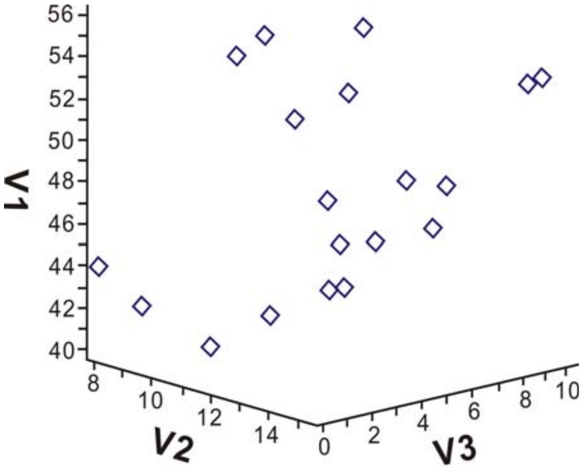

Figure 3. clustering samples

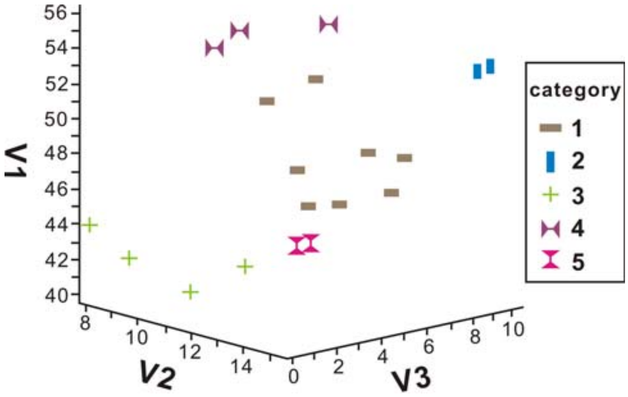

Figure 4 . Global optimal clustering result

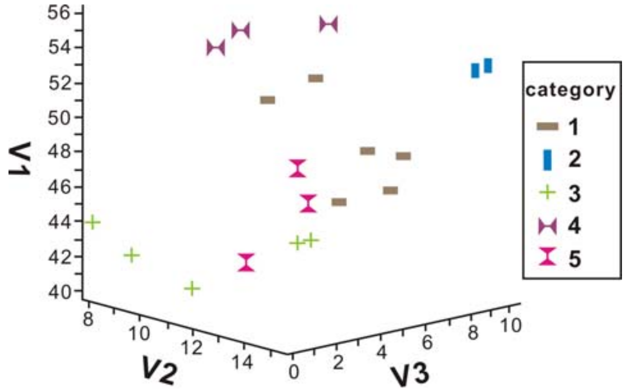

Figure 5. Local optimal clustering result

The global optimal solution can be obtained by using PGA-FCM or by IPGA-FCM (proposed in this paper). By comparing the certain requirements of evolution generation, showed in table 2, the algorithm proposed in this paper is more efficient.

TABLE2.
COMPARISON DATA OF EVOLUTION GENERATION

Algorithm	Evolution generation	FCM cycles	Running Time(S)	Objective function value
PGA-FCM	>=80	25	9.5	50.633
IPGA-FCM	>=10	25	9.7	50.633

V. APPLICATION IN TEXT CLUSTERING

After automatic segmentation, feature express, and dimensionality reduction preprocessing, the Chinese text can be transferred to m-dimension unit vector. In the processing of clustering, normalize the class centre vector m_j , and the class concept vector $c_j = m_j / \|m_j\|$ is introduced. The formula $\|m_j\|$ indicates the module of the centre vector.

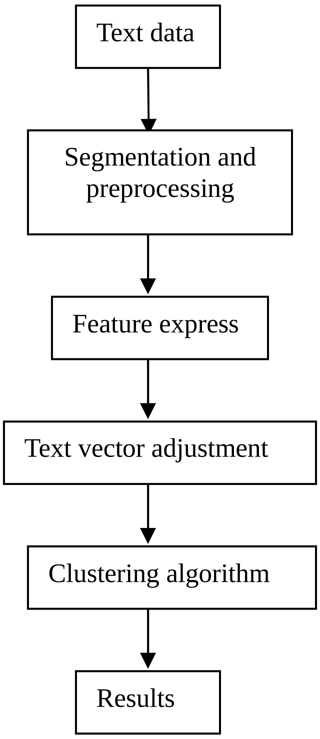

Figure 6. Structure figure of text clustering experiments

The test texts are downloaded from sina.com. Figure6. is the test structure figure.

And the modular of segmentation and preprocessing is derived from Institute of Computing of the Chinese Lexical Analysis System. There are five types of Chinese text, the securities (105), the domestic politics (160), the finance and economics (180), the tourist (58), and the science and technology (37). 40 texts of these Chinese texts not only belong to the securities class, but also to the finance and economics class. 10 texts belong to both the tourist and the finance and economics. 20 texts belong to both the domestic politics and the finance and economics. And 7 texts belong to both the science and technology and the finance and economics. Chose 1000 words when make dimensionality reduction processing by use of segmentation algorithm. Respectively utilize the IPGA-FCM and FCM algorithm to make clustering of the five types of texts. And $c = 5$, and threshold $\varepsilon = 0.3$. The clustering results are shown in table 3 and table 4.

TABLE 3.
CLUSTERING RESULTS OF FCM

Class	Number of text selection	Number of leakage selection	Number of wrong selection	Proportion of wrong or leakage selection
Securities	105	3	10	12.38%
Domestic politics	160	16	7	14.38%
Finance and economics	180	21	20	22.91%
Tourist	58	2	3	8.62%
Science and technology	37	3	3	16.21%
Average proportion of wrong or leakage selection				14.90%

The wrong or leakage election average accuracy of IPGA-FCM is smaller than that of FCM, by comparing the data in table 3 and table 4.

We can find that both the IPGA-FCM and FCM can distinguish the texts which belong to multiple classes. But the IPGA-FCM algorithm is more effective than FCM. Of course, the IPGA-FCM has some disadvantage. As the dimension of text vector increases, the clustering speed may be slow. Further study is needed at a later time.

TABLE 4.
CLUSTERING RESULTS OF IPGA-FCM

Class	Number of text selection	Number of leakage selection	Number of wrong selection	Proportion of wrong or leakage selection
Securities	105	3	9	11.42%
Domestic politics	160	16	4	12.5%
Finance and economics	180	15	19	18.99%
Tourist	58	1	2	5.17%
Science and technology	37	2	3	13.51%
Average proportion of wrong or leakage selection				12.32%

V. CONCLUSIONS

The algorithm proposed in this paper has solved the problem of local optimum in the FCM algorithm and avoid the fluctuations phenomenon in the genetic latter part of GA. The number of iteration times on the sensitive clustering data has been obviously less than the modified genetic algorithm (PGA-FCM). And the classification accuracy has been also higher than the modified genetic algorithm. We have got the obvious advantages by using the algorithm proposed in this paper.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China Grants (No. 61142007), Youth Research Fund (KYY10041) of Jiangsu Teachers University of Technology and Jiangsu's research project on modern educational technology (2011-R-18859).

REFERENCES

[1] Bezdek J C.Recent, Convergence results for the fuzzy c-means clustering algorithm, Classification, vol. 5, n. 2, pp. 237-247, 1998.

[2] Xiaoping Wang and Liming Cao, Genetic Algorithm – Theory, Application and Software. Xi’an Jiaotong University Press, 2002.

[3] Jianzhuang Liu, Weixin Xie, and Jianjun Huang, Clustering Analysis by Genetic Algorithms, ACTA ELECTRONICAL SINICA, vol. 23, n. 11, pp. 81-82, 1995.

[4] Jian Gao, Cluster Analysis Based on C-Means and Immune Genetic Algorithm, Computer Engineering, vol. 29, n.12, pp. 65-66, 2003.

[5] Andrew Watkins, Jon Timmis, and Lois Boggess, Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learing Algorithm, Genetic Programming and Evolvable Machines, vol. 5, n. 3, pp. 78-780, 2004.

[6] Jon Timmis, Artificial immune systems-today and tomorrow, Natural Computing, vol. 6, n. 1, pp. 55-57, 2007.

[7] L.N. de Castro, J. I. Timmis, Artificial immune systems as a novel soft computing paradigm, Soft Computing, vol. 7, n. 8, pp. 34-35, 2003.

[8] Jian Gao, A Parthian-Genetic Algorithm Based on Immune Mechanism for Data Clustering Optimization, Computer Engineering and Applications, vol. 12, pp. 89-90, 2003.

[9] Hongmei Yao, Xiaorong Mei, and Xianyi Zhuang, A Learning Algorithms of Fuzzy Neural Network Based on Fuzzy Clustering and Genetic Algorithms, Journal of Heilongjiang Commercial College (Natural Sciences Edition), vol. 15, n. 1, pp. 21-25, 1999.

[10] Maojun Li, Richeng Luo, and Tiaosheng Tiao, Analyis on the Genetic Operators of Partheno-Genetic Algorithm, Systems Engineering and Electronics, vol. 23, n. 8, pp. 84-87, 2001.

[11] Wei Zhang and Fuzheng Pan, A fuzzy clustering based on genetic algorithm, Journal of Hubei University (Natural Science Edition), vol. 24, n. 2, pp. 101-104, 2002.

[12] Mingfang Zhu,Changjie Tang, Shucheng Dai, Yong Xiang, Shaojie Qiao. Predicting the Total Workload in Telecommunications by SVMs[C]//Proceedings of the Third International Conference on Innovative Computing, Information and Control (ICICIC2008): June 18-20, 2008,

Dalian, China.(EI：084011616985, Inspec:Accession Number: 10184853).

[13] Rigui Zhou, Fengmei Wang and Xiaozhou Xia, Design approach for stepped impedance filter based on neural network model, Journal of Information and Computational Science, Vol. 9, n.5,pp.1395-1401,2012.

[14] Argyris Kalogeratos, Aristidis Likas, Document clustering using synthetic cluster prototypes[J], Data & Knowledge Engineering, 2011, 70(3), 284-306.

[15] Wen Zhang, Taketoshi Yoshida, Xijin Tang, QingWang, Text clustering using frequent itemsets[J], Knowledge-Based Systems, 2010, 23(5), 379-388.

[16] Chuan Li, Changjie Tang, Tianqing Zhang, Zhonghua Yu, Tianqing Zhang, Qihong Liu,Mingfang Zhu and Yongguang Jiang. Mining Multi-dimensional Frequent Patterns in Traditional Chinese Medicine Prescriptions. Q.Yang and G.Webb(eds) :PRICAI2006, LNAI Vol 4009 pp251-260, 2006, Springer Verlag Berlin Heidelberg 2006. SCI LWSCIE2007000011 #.

[17] Yintian Liu, Changjie Tang, Chuan Li, Mingfang Zhu and Tao Zeng. Dynamic Data Distribution of High Level Architecture Based on Publication and Sbuscription Tree. Lecture Notes in Computer Science Vol. 4016. Berlin: Springer-Verlag, 2006. 532~543 SCI检索号 UT ISI: 000239658700045 #©

[18] Tao Zeng, Changjie Tang, Yintian Liu, Jiangtao Qiu, Mingfang Zhu.Mining h-Dimensional Enhanced Semantic Association Rule Based on Immune-Based Gene Expression Programming, L. Feng et al. (Eds.): WISE 2006 Workshops,LNCS 4256,pp. 49 – 60,2006. © Springer-Verlag Berlin Heidelberg 2006 ,SCI: IDS=BFF82#©.

[19] PENG Jing, YANG Dons-Qin, TANG Shi-Wei, FU Yan, JIANG Han-Kui, A Novel Text Clustering Algorithm Based on Inner Product Space Model of Semantic[J], CHINESE JOURNAL OF COMPUTERS, 2007, 30(8), 1354-1362.

[20] Guan Renchu,Pei Zhili,Shi Xiaohu,Yank Chen,and Liana Yanchun, Weight Affinity Propagation and Its Application to Text Clustering[J], Journal of Cor -mputer Research and DeveloprnenL, 2010,47(10), 1733- 1740.

[21] XU Sen, LU Zhi-mao,GU Guo-chang, Spectral clustering algorithms for docu -ment cluster ensemble problem[J], Journal on Communications, 2010, Vol. 31 No.6,58-66.

[22] G. Forestier, P. Ganrski, C. Wemmert. Collaborative clustering with background knowledge [J]. Data & Knowledge Engineering, 2010, 69(02):211-228. http://dx.doi.org/10.1016/j.datak.2009.10.004

[23] Hao ZhanGang, A New Text Clustering Method Based on KGA[J], JOURNAL OF SOFTWARE, Vol. 7, n.5, pp.1094-1098, 2012.

[24] Shi Kansheng, Li Lemin, He Jie, Zhang Naitong, Liu Haitao, Song Wentao. Improved GA-based text clustering algorithm[C]. 2011 4th IEEE International Conference on Broadband Network and Multimedia Technology, IC-BNMT 2011, pp. 675-679, 2011.

Hongfen Jiang was born in September 1979, and in Changzhou City, Jiangsu Province, China. In July 2005, she was graduated from China University of Petroleum, and obtained a master's degree on Computer Software and Theory. She has worked at the School of Computer Engineering in Jiangsu Teachers University since 2005. Her main research interest is the data mining and artificial intelligence.

E-mail: jdhsff@163.com

Junfeng Gu was born in November 1978, and in Changzhou City, Jiangsu Province, China. In July 2004, he was graduated from China University of Petroleum, and obtained a master's degree on Geophysics and Information Technology. He has worked at the School of Petroleum Engineering in Changzhou University. His main research interest is geophysics and petroleum engineering.

E-mail: gujfemail@126.com

Yijun Liu was born in 1978, and in Changzhou City, Jiangsu Province, China. In July 2004, she was graduated from Nanjin University, and obtained a master's degree on Computer Application. She has worked at the School of Computer Engineering in Jiangsu Teachers University since 2006. Her interest includes data mining and computer intelligence.

E-mail: lyj@jstu.edu.cn

Feiyue Ye works in School of Computer and Information Science, the University of South Australia as a visiting scholar from 07 Mar 2009 to 08 Sept. 2009. He received the Ph.D. degree from College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, in June 2006 and his Masters degree (by research) in engineering from Xi'an University of Architecture and Technology in 1997 and he got his bachelor's degree in engineering from Xian University of Architecture and Technology in 1984. His current research interests include data mining and intelligent information system. He has co-authored more than 30 journal papers, conference papers, book chapters.

E-mail: yfy@ jstu.edu.cn

Haixu Xi was born in October 1981, and in Changzhou City, Jiangsu Province, China.In July 2006, he was graduated form Nanjing normal University and obtained a master's degree on Educational Technology. He was worked at the school of Computer School of Computer Engineering in Jiangsu Teachers University of Technology since 2006. His research area includes digital teaching resource development and multimedia technology.

E-mail: xihaixu@jstu.edu.cn

Mingfang Zhu was born in 1970, Ph.D. and associate professor, student member of China Computer Federation. His interests include data mining, computation intelligence.

E-mail: mfzhu2009@jstu.edu.cn