International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.227-238
http://dx.doi.org/10.14257/ijdta.2016.9.6.23

An Improved Affinity Propagation Clustering Algorithm Based on Entropy Weight Method and Principal Component Analysis

Wang Limin[1,2], Zhang Li[1,2], Han Xuming [3,*], Ji Qiang[1,2], Mu Guangyu[1,2] and Liu Ying[1,2]

[1]School of Management science and information engineering, Jilin University of finance and economics, Jilin, 130117, China
[2]Jilin Province Key Laboratory of Internet Fintech, Jilin University of Finance and Economics, Changchun 130117, China
[3]School of Computer Science and Engineering Changchun University of Technology,
Jilin, 130117, China
**Corresponding author:hanxvming@163.com*

Abstract

Traditional affinity propagation algorithm has inefficient results when conducting clustering analysis of high dimensional data because "dimension effect" lead to difficult find the proper class structure .In view of this, the author proposes an improved algorithm on the basis of Entropy Weight Method and Principal Component Analysis (EWPCA-AP). EWPCA-AP algorithm empowers the sample data by Entropy Weight Method, eliminate data irrelevant attributes by Principal Component Analysis, and travel with neighbor clustering algorithm, realization of high-dimensional data clustering in low dimension space. The numerical result of simulation experiment shows that the new EWPCA-AP algorithm can effectively eliminate the redundancy and irrelevant attributes of data and improve the performance of clustering. In addition, the proposed algorithm is applied in the area of the economy in our country and the clustering result is consistent with the real one. This algorithm provides a new intelligent evaluation method for Chinese economy.

***Keywords**: Affinity propagation; Principal component analysis; Entropy weight method*

1. Introduction

An important branch of unsupervised learning methods, cluster analysis, attempts to find groups of related patterns in data sets [18].Typically speaking, clustering describes the partitioning of a set of *N* data points into *K* groups using similarity measures, where a closer distance between data points within one cluster than that of data points between clusters is required. Presently, clustering has been applicable to numerous fields, such as pattern recognition [1], data mining [2], image segmentation [3], Biological information [4] *etc.*

Affinity Propagation (AP) clustering algorithm that is a new clustering algorithm based on message passing was proposed in 2007 by American scholars Frey and Dueck in Science journal [5].Unlike previous methods, Affinity Propagation clustering algorithm simultaneously considers all data points as potential exemplars, and it recursively transmits real-valued messages along edges of the network until a good set of centers and corresponding clusters is generated [6-7]. Affinity propagation has many advantages such as quick convergence, good precision *etc.* Because the Euclidean distance is used for clustering on high dimensional data, it cannot correctly reflect the real similar relations

ISSN: 2005-4270 IJDTA

among the data, and thus, the clustering effect is consequently affected. In order to overcome this shortcoming, an improved clustering algorithm, Affinity Propagation with Entropy Weight Method and Principal Component Analysis (EWPCA-AP), is proposed in this paper. EWPCA-AP algorithm empowers the sample data by Entropy Weight Method, eliminate data irrelevant attributes by Principal Component Analysis, and travel with neighbor clustering algorithm, realization of high-dimensional data clustering in low dimension space. Simulated experimental results show that the proposed AP algorithm is better than the traditional AP algorithm in clustering performance. Additionally in this paper, the proposed AP algorithm's practical application was demonstrated by applying in the area of the economy in our country and the clustering result is consistent with the real one. This algorithm provides a new intelligent evaluation method for Chinese economy.

2. Affinity Propagation Clustering Algorithm

Affinity propagation clustering algorithm is a clustering algorithm that is based upon affinity information propagation whose goal is to find the optimal representative set [8-10]. The algorithm simultaneously considers all data points as potential exemplars and avoids clustering results to select the initial representative point. Meanwhile, the algorithm takes as input a collection of real-valued similarities between data points, where the similarity $s(i, k)$ indicates how well the data point with index k is suited to be the exemplar for data point i. Through the information exchange of attraction degree, and therefore, a more ideal representative point set is obtained. The conventional affinity propagation clustering algorithm treats the Euclidean distance as a similar measuring method, for points x_i and x_k :

$$s(i,k) = -d_{ik}^2 = -\left\|x_i - x_k\right\|^2 \tag{1}$$

Before affinity propagation clustering algorithm, taking as input a real number $s(k, k)$ for each data point k, these values are named "preferences". These data points with larger values of $s(k, k)$ are more likely to be chosen as exemplars. The number of clusters is influenced by the values of the input preferences, the value of the input preference is even greater, the possibility of representative points is greater, and the number of clustering output is more numerous. Otherwise, the value of the input preference will be smaller, and the number of clustering output will be less. If a prior, all data points are equally suitable as exemplars, so namely all the $s(k, k)$ is the same value p. In traditional affinity propagation clustering algorithm, the shared value is defined the median of the input similarities or their minimum.

In order to select the appropriate representative point, there are two kinds of messages exchanged between data points, there are "responsibility" and "availability", which each represents a different competitive goal. The responsibility $r(i, k)$, means x_i point to candidate exemplar x_k that reflects the accumulated evidence for how well-suited x_k is to serve as the exemplar for x_i, taking into account other potential exemplars for x_i. The availability $a(i, k)$, means candidate exemplar x_k to x_i that reflects the accumulated evidence reflects the accumulated evidence for how appropriate it would be for x_i to choose x_k as its exemplar, taking into account the support from other points that x_k should be an exemplar. The larger $r(i, k)$ and $a(i, k)$ is, the larger the possibility that x_k is final class representative point. Affinity propagation is the iterative process that "responsibility" and "availability" update alternately. At the beginning, the availability are initialized to zero: $a(i,k)=0$. Then, the responsibilities and "responsibility" and "availability" update as follows:

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t. k' \neq k} \left\{a(i,k') + s(i,k')\right\} \tag{2}$$

$$a(i,k) \leftarrow \begin{cases} \min\left\{0, r(k,k) + \sum\limits_{i' s.t. i' \notin \{i,k\}} \max\left\{0, r(i',k)\right\}\right\} & i \neq k \\ \sum\limits_{i' s.t. i' \neq k} \max\left\{0, r(i',k)\right\} & i = k \end{cases} \tag{3}$$

When updating the messages, it is important that introducing the important parameter of damping factor λ to avoid numerical oscillations that arise in some circumstances. In each iteration, the updating results obtained *r*(*i*, *k*) and *a*(*i*, *k*) by weighing with messages from the previous iteration to form the final values used in matrix updates, namely:

$$r^{(t+1)}(i,k) \leftarrow (1-\lambda) r^{(t+1)}(i,k) + \lambda r^{(t)}(i,k) \tag{4}$$

$$a^{(t+1)}(i,k) \leftarrow (1-\lambda) a^{(t+1)}(i,k) + \lambda a^{(t)}(i,k) \tag{5}$$

Where $0 \leq \lambda \leq 1$, its default value is 0.5. *t* is the current number of iterations. The function of the damping factor *λ* is to improve convergence. When affinity propagation fails to converge for shake, we increase *λ* to eliminate shake. Iteration is updated alternately through the information of "attraction" and "ownership". Finally, determining data point *k* as the class representative point:

$$\arg\max_{k}\left(a(i,k) + r(i,k)\right) \tag{6}$$

3. An Affinity Propagation Algorithm Based on Principal Component Analysis and Entropy Weight Method

3.1 Entropy Weight Method

The entropy weight method is a kind of empowerment approach. In the process of concrete use, the entropy weight method uses information entropy to calculate the entropy weight of each index according to the variation degree of each index, thus we can draw the objective index weight through the entropy weight to modify the weight of each index [11,16].

The basic principle of entropy weight method:

According to the basic principle of information theory, information is a measure of the ordering degree of the system, while entropy is a measure of the disorder degree of the system. If the system is in a variety of different states, when the probability for each state is $p_i\left(i = 1, 2, ..., m\right)$, the entropy of the system is defined as:

$$e = -\sum_{i=1}^{m} p_i . \ln p_i \tag{7}$$

Obviously, when $p_i = {}^{1}\!/_{m} (i = 1, 2, ..., m)$, that is to say, when the probability of each state appears the same, the value of entropy is maximal $e_{\max} = \text{lnm}$.

The number of evaluated project is *m*, the number of evaluation indicators is *n*, which formed the original evaluation matrix $R = \left(r_{ij}\right)_{m\times n}$

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & r_{m3} & r_{m4} \end{pmatrix}_{m\times n} \tag{8}$$

Among, r_{ij} is the value of the *i* project under the *j* index.

The entropy weight of the *j* index:

$$w_j = {\left(1-e_j\right)} \Big/ {\sum_{j=1}^{n}\left(1-e_j\right)} \tag{9}$$

3.2. Principal Component Analysis

The principal components analysis, also called the main component analysis, is a method of data dimensionality reduction. The basic idea is to produce the optimal simplification of a multi variable system in which the *p* variable values x_1, x_2,…, x_p described in the data sheet of the original system re-adjust and combine, and extract *m* (*m* ≤ *p*) synthesize variable f_1, f_2, …, f_m, which can maximize the overlap of the information of data object described in the original system. Dimensionality reduction and the simplification goal can then be obtained [12-14]. They can be represented:

$$\begin{cases} f_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ f_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\ \quad\vdots \\ f_m = a_{1m}x_1 + a_{2m}x_2 + \cdots + a_{pm}x_p \end{cases} \tag{10}$$

Where f_m is the principal component, and *m* ≤ *p*. Generally, the cumulative contribution of *m* principal components is 85%, so in keeping the most information, the goal of dimensionality reduction can be achieve.

3.3. Affinity Propagation Clustering Algorithm Based on Entropy Weight Method and Principal Component (EWPCA-AP Clustering Algorithm)

The raw data is weighted by using entropy weight method. Then, the principal component is introduced into the affinity propagation clustering algorithm in this paper. The processed sample is reduced dimension by the principal component with a focus on keeping most information.

The EWPCA-AP clustering algorithm steps are as follows:

1 Making the characteristics of the data normalized, such that $x_{ik} \in [0,1]$ (*k* = 1,2, ..., *m*), characterized by the designated normalized formula:

$$x_{ik}{}' = \frac{x_{ik} - \min\left\{x_{jk}, j=1,2,\ldots,n\right\}}{\max\left\{x_{jk}, j=1,2,\ldots,n\right\} - \min\left\{x_{jk}, j=1,2,\ldots,n\right\}} \tag{11}$$

2 Calculate weight of each attribute that are raw data.

3 The raw data is weighted by using entropy weight method.

4 Using the idea of principal component analysis method, select the top *m* principal components so that the cumulative contribution rate reaches above 85%, in order to achieve dimensionality reduction for the processed sample.

5 Initialize $a(i,k)=0$, calculate similarity matrix *s* according to Eq.(1). And initialize the value *p* according to the following equation, namely:

$$p = \alpha \bullet median(s(:)) \tag{12}$$

6 Update iteration *a*(*i*, *k*) and *r*(*i*, *k*).

7 The representative points are obtained according to Eq. (7).

8 The algorithm will stop if the results of the algorithm do not change or reaches a predetermined maximum number of iterations, otherwise returns to 5.

4. Simulation Experiment and Analysis

In this section, the improved affinity propagation clustering algorithm is applied to the clustering experiments by using the Wine, Spect and Ecoli datasets in the UCI database. First, the entropy weight method is applied to weight the raw date. Second, the principal component analysis method is applied to the dimensionality reduction. Third, the Euclidean distance measuring method is used to calculate the similarity matrix and adjust the bias parameter *p* according to the data set information and determine the optimal number of clusters. Finally, evaluate the effective of clustering by using Silhouette. The data characters selected and the numbers of principal components are shown in Table 1.

Table 1. The Data Characteristics and the Number of Principal Components of the Three Data Sets

Date Set	Instances	Attributes	Classes	Number of Principal Components
Wine	178	13	3	8
Spect	187	22	2	9
Ecoli	336	8	3	5

4.1. Silhouette Effective Index

Silhouette index has been widely used for its good evaluation of clustering structure. It can reflect the capacity characteristic of class inseparability and class divisibility, which can not only be used to evaluate the optimal number of clustering but also can be used to evaluate the quality of clustering.

Suppose one data set composed of *n* data points is divided into *k* clustering C_i(i =1,2,…, *k*), *a*(*t*) is the average dissimilarity or distance that is obtained by sample point *t*, and all the other samples in cluster C_i, *d*(*t*, C_i) are the average dissimilarity or distance by *t* and all the other samples in another cluster C_j, then $b(t)=\min\{d(t,c_i)\}$, *i* =1,2,…, *k* and *i*≠ *j*. Thus, the formula for silhouette index sample *t* is calculated as:

$$Sil(t) = \frac{[b(t)-a(t)]}{\max\{a(t),b(t)\}} \tag{13}$$

The average $Sil \in [-1,1]$ value for all samples in a cluster represents the capacity (intra-class average distance) and reparability (minimum class distance). The average *Sil* value for all samples in a data set can reflect the quantity of the cluster, in which the bigger the Silhouette index the better the cluster's quantity.

4.2. Comparison and Analysis

Set parameters φ by manual continuously, parameters φ as shown in Table 2. When the AP algorithm and the EWPCA-AP algorithm achieve the best cluster in Table 3, experimental result contrast as shown in Figure 1,convergence times contrast as shown in Figure 2.

Table 2. The Best Weight Coefficient

Date Set	Weight Coefficient		
	AP	PCA-AP	EWPCA-AP
Wine	16	11	17
Spect	4	9	16
Ecoli	19	30	20

Table 3. AP Algorithm and EWPCA-AP Algorithm Cluster Results Comparison

Date Set	Classes	Best Number of Clustering		
		AP	PCA-AP	EWPCA-AP
Wine	3	3	3	3
Spect	2	5	2	2
Ecoli	3	3	3	3

It can be seen from Table 3 that the Spect data set cannot achieve the fact cluster from the AP algorithm. It can be seen from Figure 1 that the three data sets precision from an affinity propagation algorithm based on principal component analysis is smaller than the precision from EWPCA-AP clustering algorithm, and the abscissa coincidence is in the figure. It can also be seen that the precision from an affinity propagation algorithm based on principal component analysis and entropy weight method is more than the other two kinds of clustering algorithm. Therefore, the precision from an affinity propagation algorithm based on principal component analysis and entropy weight method is high and, therefore, can significantly improve the quality of the cluster.

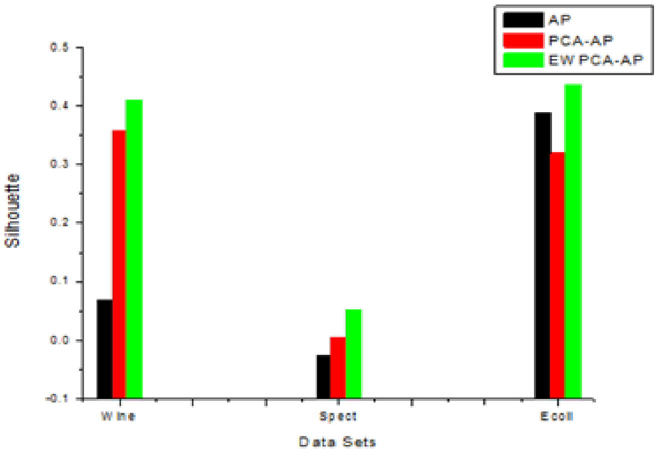


Figure 1. Experimental Result Contrast Chart

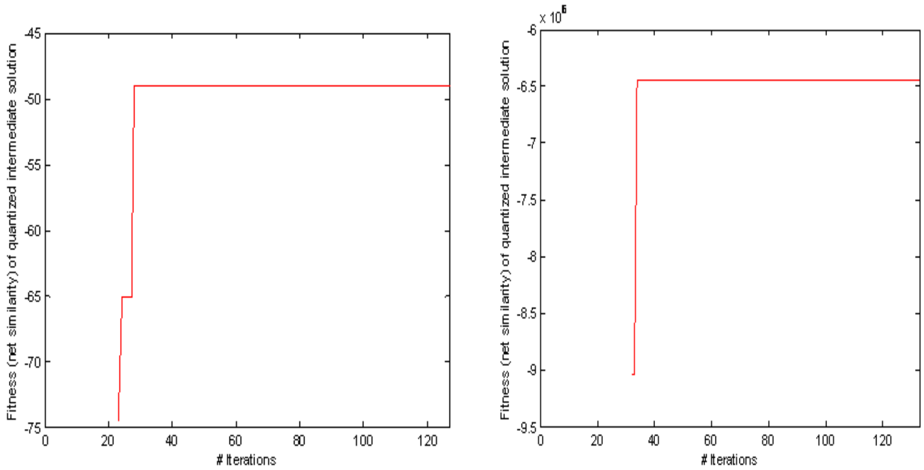


Figure 2. Convergence Times Contrast

It can be seen from Figure 2 that proved EWPCA - AP clustering algorithm convergence times less than the classic AP clustering algorithm. Therefore, the proposed improved algorithm convergence speed is fast.

5. Application of EWPCA-AP Clustering Algorithm in China's Regional Economy

5.1. Data Selection

Application of the EWPCA-AP clustering algorithm can be used to better reflect the economic situation of our China’s region and serve as an effective tool to help our government make the correct policy decision. In this paper, we select 12 the economic indicators which can reflect the economic development in different areas: regional gross domestic product (GDP), GDP per capital (GDPPC), the first industrial added value (FDA), the secondary industry (SDA), the tertiary industry value increase (TDA), the consumer price index (CPI) value, the social fixed assets investment (SFAI), social total retail sales of consumer goods (RSCG), (TI) total imports and exports (TE), urban residents per capital disposable income (UHPCDI), industrial value (IA) [17], the data as shown in Table 4.

Table 4. Economic Indicators and Relevant Data of Various Provinces and Cities

Region	GDP	GDPPC	FDA	SDA	TDA	CPI	SFAI	RSCG	TI	TE	UHPCDI	IA
Beijing	19500.6	93213	161.8	4352.3	14986.4	103.3	6847.06	83751000	36585710.8	6324622.3	403210	3536.89
Tianjin	14370.2	99607	188.5	7276.7	6905	103.1	9130.25	44704000	7950339.9	4902477.8	32293.6	6678.6
Hebei	28301.4	38716	3500.4	14762.1	10038.9	103	23194.23	104007000	2392030.1	3096267.9	22580	13194.76
Ningxia	2565.1	39420	223	1265	1077.1	103.4	2651.14	6105000	66544.9	255246.4	21833	944.5
Xinjiang	8360.2	37847	1468.3	3766	3126	103.9	7724.46	20391500	529210.6	2226980.4	19874	3024.27

5.2. Clustering Analysis of the Economic Situation of Our China's Region

Using the entropy weight method to empower the raw data of each attribute. Then, in reserving most of the information and given the improved data of economic indicators for our 31 provinces and cities for the principal component extraction, extracting the first 7 principal components makes the aggregate contribution rate more than 85%. By continuously adjusting the *p* value gives the optimal clustering results, as illustrated in Figure 3.

From Figure 3, we can see that the improved EWPCA-AP clustering algorithm of China's 31 provinces and cities clustering results are basically in accord with China's regional distribution characteristics: eastern coastal developed area, the western region underdeveloped. In order to demonstrate the proposed clustering algorithm is feasible and practical for regional economic evaluation in China. In this paper, the total GDP of the two categories of regional economy is illustrated by comparison, as shown in Figure 4.

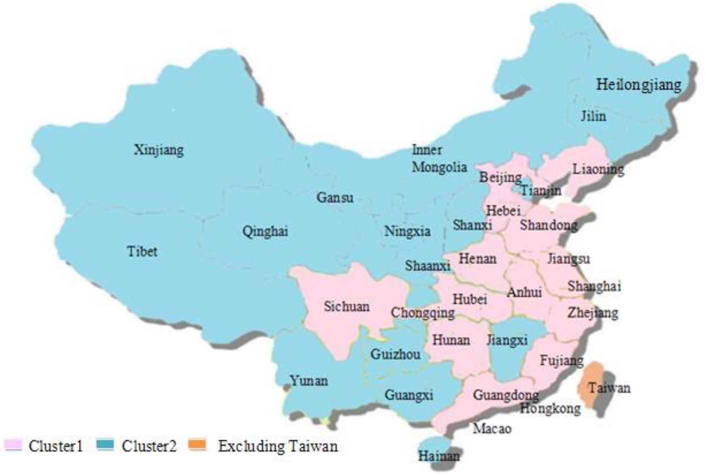


Figure 3. Clustering Results of Regional Economic Situation

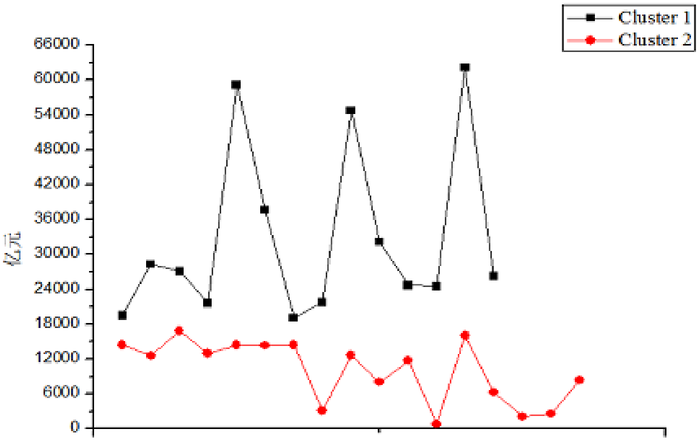


Figure 4. Regional GDP Comparison Chart

GDP refers to the final results of the activities of all permanent units in the region during the period of a certain period of time [15]. It reflects the economic performance of a region.It is the best indicators of regional economy.It can be seen from Figure 3 that the first kind of regional GDP is higher than the second. It shows that the economic strength of the first category is relatively strong, the industrial structure is more reasonable, the economic strength of the second regions is weak, and the industrial structure needs to be improved.At the same time, at the request of the national policy of our country, we should give priority to arrangement of the resource exploitation and infrastructure projects in the second area, increase of regions of the second type of poverty alleviation efforts and strengthen the first region and second region economic cooperation and technical cooperation, guide more foreign investment to the second region.

6. Conclusions

Because the affinity propagation clustering algorithm clustering effects deals with more attributes and the information concerning overlapping samples is not good, this paper proposed an affinity propagation algorithm based on principal component analysis and entropy weight method. The algorithm used the method of entropy weight to weaken the raw data attribute, and then combined then principal component analysis on the improved data dimensionality reduction, which then effectively overcomes the effect of redundant information and improves the quality of clustering. The experimental results show that the proposed algorithm can effectively improve the clustering performance and clustering speed. In addition, the improved affinity propagation clustering algorithm is applied to the analysis data of various provinces and cities of our country economic indicators, satisfactory results were obtained, which the government in formulating economic policies provide a reference for the, for the various provinces and cities of our country economy provides a new intelligent evaluation method.

Acknowledgements

The research was supported by the National Science Foundation of China under grant No. 61202306, 61472049, 61402193 and 61572225, the National Society Science Foundation of China under grant No. 15BGL090, the Ministry of Education, the Foundation under grant No. 13YJAZH130, the Application Basis Foundation of Jilin Provincial Science & Technology Department under Grant No. 201215119, 20130522177JH and 2013s0101072JC, the Society Science Research Foundation of Jilin

Province under Grant No. 2014B166, the support project on the excellent talent of Jilin Province in colleges and universities under grant No. 2014159.

References

[1] J. Tian, B. Y. Zhang and W. Y. Yang, "Gird Pattern Recognition Based on Clustering of Self-organizing Maps", Geomatics and Information Science of Wuhan University, vol. 11, no. 17, **(2013)**.

[2] M. Verma, M. Srivastava, N. Chack, A. K. Diswar and N. Gupta, "A comparative study of various Clustering Algorithms in data mining", International Journal of Engineering Research and Applications (IJERA) ISSN. 2248-9622, **(2012)**.

[3] M. Gong, Y. Liang, J. Shi, W. Ma and J. Ma, "Fuzzy c-means clustering with local information and kernel metric for image segmentation", Image Processing, IEEE Transactions on., vol. 22, no. 2, **(2013)**, pp. 573-584.

[4] H. L. Yang, L. Zhu and B. Hang, "Feature Selection and Sample Classification for SELDI-TOF Mass Spectrometry Data Based on Affinity Propagation Clustering", Chinese Journal of Biomedical Engineering, vol. 1, **(2013)**, pp. 14-20.

[5] B. J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points", Science, vol. 315, **(2007)**, pp. 972-976.

[6] L. Zhang, Y. Chen, Y. Ji and Z. Jinsong, "Research on K-means algorithm based on density", Application Research of Computers, vol. 11, **(2011)**, pp. 4071-4073+4085.

[7] X. F. Lei, K. Q. Xie and F. Lin, "An Efficient Clustering Algorithm Based on Local Optimality of K-means", Journal of Software, vol. 7, **(2008)**, pp. 1683-1692.

[8] W. M. Lu, C. Y. Du and B. G. Wei, "Distributed Affinity Propagation Clustering Based on Map Reduce", Journal of Computer Research and Development, vol. 49, no. 8, **(2012)**, pp. 1762-1772.

[9] Y. D. Fu and J. L. Lan, "Kernel-based adaptation for affinity propagation clustering algorithm", Application Research of Computers, vol. 29, no. 5, **(2012)**, pp. 1644-1647.

[10] J. P. Zhang, F. C. Chen and S. M. Li, "Parallel Affinity Propagation Clustering Algorithm Based on Hybrid Measure", Computer Science, vol. 40, no. 7, **(2013)**, pp. 167-172.

[11] H. Qin and D. Luo, "New Uncertainty Measure of Rough Fuzzy Sets and Entropy Weight Method for Fuzzy-Target Decision-Making Tables", Journal of Applied Mathematics, **(2014)**.

[12] D. Paul and I. M. Johnstone, "Augmented sparse principal component analysis for high dimensional data", arXiv preprint arXiv., **(2012)**, pp. 202+1242.

[13] Z. Ma, "Sparse principal component analysis and iterative thresholding", The Annals of Statistics. Vol. 4, **(2013)**, pp. 772-801.

[14] R. Bro, "Smilde, A.K.Principal component analysis", Analytical Methods, vol. 6, **(2014)**, pp. 2812-2831.

[15] M. Drehmann and K. Tsatsaronis, "The credit-to-GDP gap and countercyclical capital buffers: questions and answers", BIS Quarterly Review, **(2014)**.

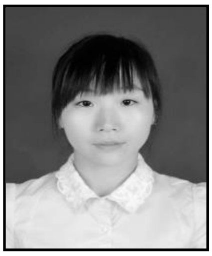
[16] Y. Zhang, "TOPSIS Method Based on Entropy Weight for Supplier Evaluation of Power Grid Enterprise", 2015 International Conference on Education Reform and Modern Management. Atlantis Press, **(2015)**.

[17] L. Xiaobo, "The space evolution research of economic competition in Golden Delta Counties of the Yellow River", 2014 Conference on Informatization in Education, Management and Business (IEMB-14). Atlantis Press, **(2014)**.

[18] M. R. Anderberg, "Cluster analysis for applications. Office of the Assistant for Study Support Kirtland Afb N Mex, **(1973)**.

Authors

Limin Wang, Female, born in 1975. Received her Ph.D. degree in Jilin University in 2007. Now she is a professor, membership of China Computer Federation. Recently several years, she has published more than 60 papers Her current research interests include machine learning, data mining and finance Engineering.



Li Zhang, Female, born in 1989. M. S degree candidate in Jilin University of Finance and Economics his main research interests include data mining, machine learning and intelligent computation.

Xuming Han, Male, born in 1971. Received his Ph.D. degree in Jilin University. Now he is a professor, membership of China Computer Federation, his main research interests include machine learning, data mining, evolutionary algorithms and intelligent computation.

Qiang Ji, Male, born in 1989. M. S degree candidate in Jilin University of Finance and Economics his main research interests include data mining, machine learning and intelligent computation.

Guangyu Mu, Female, received her Ph.D. degree in Jilin University in 2011.She is an professor of Jilin University of Finance and Economic. Her research interests focus on Computer Application Technology and Trust Mechanism in Social Commerce.

Ying Liu, Female, received her Ph.D. from Chinese Academy of Sciences in 2013.She is an associated professor of Jilin University of Finance and Economic. Her research interests focus on Machine Learning, Finance Engineering and Data Mining.