Science Journal of Education
2017; 5(3): 105-110
http://www.sciencepublishinggroup.com/j/sjedu
doi: 10.11648/j.sjedu.20170503.15
ISSN: 2329-0900 (Print); ISSN: 2329-0897 (Online)

Research of Enterprise Credit Rating Based on K-Means GMDH Model

Xiangyun Zhou*, Yixiang Tian

College of Economics and Management, University of Electronic Science and Technology of China, Chengdu, China

Email address:
zhouxy616@163.com (Xiangyun Zhou), tianyx@uestc.edu.cn (Yixiang Tian)
*Corresponding author

To cite this article:
Xiangyun Zhou, Yixiang Tian. Research of Enterprise Credit Rating Based on K-Means GMDH Model. *Science Journal of Education.* Vol. 5, No. 3, 2017, pp. 105-110. doi: 10.11648/j.sjedu.20170503.15

Received: March 15, 2017; **Accepted**: April 10, 2017; **Published**: April 12, 2017

Abstract: Since the outbreak of credit risk, researching on corporate credit rating has been brought into investors, the government and scholars focus. This paper constructs an optimal K-means clustering Group Method of Data Handling model can effectively improve the accuracy of rating results, reduce the computational complexity, and this paper proves the model under the least squares estimation can get the optimal results. This article uses Chinese corporate credit rating and financial indexes to study, comparing its results with the concequences of Hidden Markov GMDH model and other traditional neural network models. The empirical outcomes show that the K-means clustering GMDH model is better than Hidden Markov GMDH model and the remaining four neural network models, indicating that the method can effectively improve the accuracy of corporate credit rating assessment and reduce the cost of rating.

Keywords: K-Means GMDH Model, Enterprise Credit Rating, Credit Risk Management

1. Introduction

Credit risk has spread from private enterprises to state-owned enterprises, which leads to credit rating becoming a key issue of credit risk management. Chinese credit rating agencies continuously expand from the primary stage. Government, enterprises and investors constantly increase requirements on the quality of credit rating businesses. However, credit rating system of Chinese enterprises is still not perfect at present, and can't temporarily meet the demand of the country's financial market development. Therefore, improving the accuracy of credit rating assessment and strengthening the inspection of credit rating quality have great significance to Chinese, even worldwide credit rating service.

Credit rating is a method for investors to make judgement on the ability and willingness of issuers to fulfill their financial obligation. Rating agencies often hire professionals to complete the complex process, consuming a large amount of time and manpower for the rating results [1]. Meanwhile, in the credit market, there exist conflicts of interest among investors, enterprises to issue bonds and rating agencies, namely credit rating agencies collect the enterprsises' financial information for default risk assessment, and publish their credit rating in public; Enterprises need to pay for the service offered by credit rating agencies, but investors need not to pay. Enterprises will pursue higher credit rating for getting more financing through buying inflated rating. As for credit rating agencies, a for-profit organization, may ensure their own interests to offer low quality rating, which will damage the benefit of investors [2]. In conclusion, Due to information asymmetry between issuers and investors, also there are conflicts of interest, so the quality of credit rating is of great importance to healthy and ordered development of capital market.

Previous studies mainly from the following three aspects research enterprise credit rating:

(1) Quality and effect of the credit rating industries, Ping He and Meng Jin (2010) build real interest cost regression model, and analysis that the credit rating have influence on cost of issuing bonds in the primary market. The empirical results show that the enterprise credit rating has explanatory power to the issue of cost [3]. Laizong Kou and Yuzhang Pan (2015) consider that under issuer payment model and rating purchase mechanism, credibility of credit rating agencies

remains to be tested. In the ordinary least squares (OLS) regression, the competition degree of the rating agencies in China is taken as the instrumental variable, and the results show that the economic effect of credit rating on the cost of issuing bonds is significantly reduced [4].

(2) Enterprise credit rating system, B. Shi and G. Chi (2013) construct credit risk evaluation index system for small business loan which is consist of six criteria layers including the basic situation, guarantee and joint guarantee, solvency, profitability, operating capacity, macro-environment, and other thirteen indicators including the asset liability ratio, Engel coefficient and so on [5]. K. Kim and H. Ahn (2012) research on credit scoring of loan enterprises through some indicators such as the owner's equity, sales volume, total liabilities, average sales (sales / employees), the number of years of establishment, operating profit margin, the ratio of total assets to cash flow, etc [6].

(3) Enterprise credit rating model, statistical and artificial intelligence technology are mainstream credit rating method. P. Hajek and K. Michalak (2013) use feature selection to train the multilayer perceptron and radial basis function for predicting the credit rating, the results emphasize that the algorithms are used to classify the credit rating of enterprises, which can improve the predicting accuracy [7]. Ge-Er Teng and Chang-Zheng He (2013) combine hidden Markov model (Anastasios Petropoulos, 2016; Robert J. Elliott, 2014) with Group Method of Data Handling (GMDH) model (Yi-Xiang Tian, 1999), using mixed model to estimate consumer credit rating [8].

Based on above literatures, the rating quality of rating agencies can affect the cost of enterprise financing, so it is very important to improve the quality of enterprise credit rating. In recent years, most of the scholars have constructed enterprise credit rating model by using statistical methods and artificial intelligence technology, however, the big data and multiple indicators increase the computational complexity, which lead to low accuracy of credit rating by traditional methods. The financial indicators of enterprises have a relative relationship, so the clustering algorithm is used to explore the implicit relationship between the indicators, and the sample data are classified according to the characteristics of the indicators. The clustering results based on a certain order being added to the GMDH model in each layer, with the last filter varibles, are the next layers' input varibles to estimate. This method, which reduce the computational complexity and improve the accuracy of calculation, can avoid too many input variables and collinearity problem. This paper combining with the existing ideas of establishing credit rating model, uses rating indexes from previous literatures and contructs K means-GMDH model to estimate Chinese enterprise credit rating for efficiently reducing computational complexity. The results are compared with those of hidden Markov-GMDH model, traditional GMDH model, multilayer perceptron, radial basis function and artificial neural network model. The empirical results show that K-means GMDH model can significantly improve the accuracy of enterprise credit rating.

2. K Means GMDH Modeling Method

2.1. GMDH Model

(1) The sample data set W can be divided into training data set A and test data set B, W=A+B;

(2) The combination of two input varibles are generated for C_m^2 neurons and are constructed into transfer function, which is commonly used Kolmogorov-Gabor polynomial function as reference function, such as two binomial function:

$$y = f(x_i, x_j); i, j = 1, 2, ..., m; i \neq j \tag{1}$$

$$y = f(x_i, x_j) = \alpha_0 + \alpha_1 x_i + \alpha_2 x_j + \alpha_3 x_i x_j + \alpha_4 x_i^2 + \alpha_5 x_j^2 \tag{2}$$

(3) In first layer, the algorithm establishes intermediate models used to estimate the transfer function coefficients on training data set by the least squares method and calculates the sum of squares of errors between the real value and the calculated value of the model, selecting the model of least square fitting error on test data set. Filter variables satisfy $m_1 \leq C_m^2$, combine input variables in pairs generating for C_m^2 neurons as the second input varibles which is selected by the external criteria of GMDH [17];

(4) Repeat the above process, the model can produce intermediate models of the third layer, the fourth layer up to N th layer. When the external criteria of N th layer becomes optimal, the iterative process is stopped and the optimal complex model is obtained. Mean square error (*MSE*) and root mean squared error (*RMSE*) can measure the deviation between the predicted value and the true value. When mean squared error and root mean squared error are smaller, the predicting accuracy is higher.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_{obs,i} - X_{\mathrm{mod}el,i})^2 \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{\mathrm{mod}el,i})^2}{n}} \tag{4}$$

2.2. Theoretical Derivation

Suppose $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ...\boldsymbol{x}_n\}$ are n spatial data, clustering the original sample *X*, C_j is the clustering center of the J class, The distance between x_i and c_j is:

$$d(x_i, c_j) = \sqrt{(x_i^1 - c_j^1)^2 + ... + (x_i^k - c_j^k)^2 + ... + (x_i^n - c_j^n)^2} \tag{5}$$

In this paper, the purpose is to solve the optimal clustering center c_i, c_j, minimizing the distance between GMDH intermediate variables q and clustering center[12]. Objective function is:

$$q_m^{l+1} = \arg\min_{c_i,c_j}[d^2(c_i,q_i^l)+d^2(c_j,q_j^l)]; i \neq j, l = 1,2...,n \quad (6)$$

s.t.

$$d(c_j,q_j^l) = \sqrt{(c_j-q_j^1)^2+(c_j-q_j^2)^2+...+(c_j-q_j^l)^2} \quad (7)$$

$$d(c_i,q_i^l) = \sqrt{(c_i-q_i^1)^2+(c_i-q_i^2)^2+...+(c_i-q_i^l)^2} \quad (8)$$

Seek respectively partial derivative of c_i, c_j:

$$\frac{\partial q_m^{l+1}}{\partial c_i} = 2[(c_i-q_i^1)+(c_i-q_i^2)+...+(c_i-q_i^l)] = 0 \quad (9)$$

$$\frac{\partial q_m^{l+1}}{\partial c_j} = 2[(c_j-q_j^1)+(c_j-q_j^2)+...+(c_j-q_j^l)] = 0 \quad (10)$$

$$c_i = \frac{q_i^1+q_i^2+...+q_i^l}{l} \quad (11)$$

$$c_j = \frac{q_j^1+q_j^2+...+q_j^l}{l} \quad (12)$$

q_i^1, q_j^1 represent the i th and j th intermediate variables in the first layer of GMDH model, q_m^{l+1} is final output variable of GMDH model, optimal clustering center ci, cj respectively represent i th and j th intermediate variables in the first layer to the l th layer. The results show that the optimal clustering center is in line with the definition of the K means clustering center. Therefore, under the least squares estimation, K means clustering GMDH model can get the optimal solution.

3. Main Steps of K Means GMDH Model

It is difficult to determine the number of neurons in each layer of the GMDH model. When the number of reserved neurons is incorrect setting, it will cause useful information prematurely eliminated and affect the quality of final model. For avoiding the problem, Zengguo Li et al (2012) present the method of adding initial variables to improve the transmission of GMDH model. In this method, the variables of initial layer are all added to each layer, which will lead to collinearity and low operation speed and accuracy due to a large number of input variables per layer.

The principle of K means clustering GMDH model: Cluster analysis of original sample data uses K means clustering, then the clustering results are respectively added to each layer of the GMDH model in a certain order, with the last filter variables, as the next input variables to estimate. The optimal GMDH model is stopped until the last cluster result is trained.

This model can effectively reduce the input variables of each layer, prevent the loss of useful information, decrease the calculation of complication and improve the accuracy of estimation. The Ge—Er Teng et al (2013) establish hidden Markov-GMDH model to improve the accuracy of consumer credit rating, but the initial parameters of hidden Markov are generated by random numbers, which may affect the speed and accuracy of the method. Although the traditional neural network model is widely used in the prediction of enterprise credit rating (P. Hajek, 2013; Michele Azzollini, 2011), the accuracy of prediction is still to be improved. According to the defects of the existing credit rating methods, this paper constructs K-means GMDH model and compares the results with those of hidden Markov-GMDH model, traditional neural network models for testing rating efficiency of this new method.

Specific steps are as follows:

Step 1: Set the initial clustering number K before clustering. Select K objects from the N data objects as the initial clustering centers, and Other objects are assigned to the most similar classes according to the similarity of the cluster centers [19].

Step 2: Calculate the clustering centers for each update class. Suppose J th class in the sample is $\{x_{j1}, x_{j2}, ...x_{jn}\}$, namely including nj, so clustering center of this class is $c_j = (c_j^1, c_j^2, ..., c_j^k, ..., c_j^n)$, and c_j^k is the k th attributes of c_j

$$c_j = (x_{j1}^k + x_{j2}^k + ... + x_{jn}^k) \Big/ n_j \quad (13)$$

Step 3: Repeat this process until convergence of the standard measure function, from the performance of the process, the value of clustering centers after updating should be consistent with which is before updating. Use mean square error *J* as the standard clustering measure function.

$$J = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-c_i)^2}{n-1}} \quad (14)$$

Step 4: The sample data are proportionally divided into training set and test set, each kind of results obtained by K means clustering algorithm is added as an input variable to each layer of the GMDH model in a certain order with the previous filtering variables.

Step 5: Use the least squares method to estimate weight of transfer function on training set, select reserved neurons in the first layer and b th class ($b<i$, $b \neq a$) from clustering results as input variables in the second layer, and so on. Using additive method of clustering results up to the final group, the GMDH model is trained to get the optimal model.

4. Empirical Test

4.1. Data Selection and Preprocessing

Nowadays, there have been six large enterprise credit rating agencies in China, such as Dagong Global Credit Rating, Credit International Assessment, China Lianhe Credit Rating, Peng Yuan Credit Rating, Shanghai New Century Credit Evaluation Investment Service and Zhong Cheng Trust Investment company. In this paper, we use the Wind database to collect credit ratings and finance indicators of the 1254 companies from the six agencies, main indicators are: The previous enterprise credit rating, current ratio, quick ratio, interest coverage ratio, the rate of return on total assets, the rate of return on net assets, the ratio of profit to net sales, inventory turnover, the growth rate of main business revenue and the ratio of total cash debt. (Anastasios Petropoulos, 2016; Qing-Miao Li, 2012). Credit rating organizations classify credit scores according to enterprise credit risk and financial conditions. There are fourteen scores from AAA, AA+, AA to CCC, CC, C. The most of researchers use linear transformation theory to orderly evaluate socres, namely AAA is equal to 8, AA+ is 7, AA is 6, AA- is 5,…, BBB and below BBB is 1. For eliminating heteroscedasticity of the data and reducing the estimation error, we evaluate before logarithm of rating value and standardize sample data [21].

4.2. Empirical Analysis

In this paper, we firstly use matlab software to divide all input variables into K means clustering for simply determining the classifications of sample data. Figure 1 shows that sample data can be classified four groups.

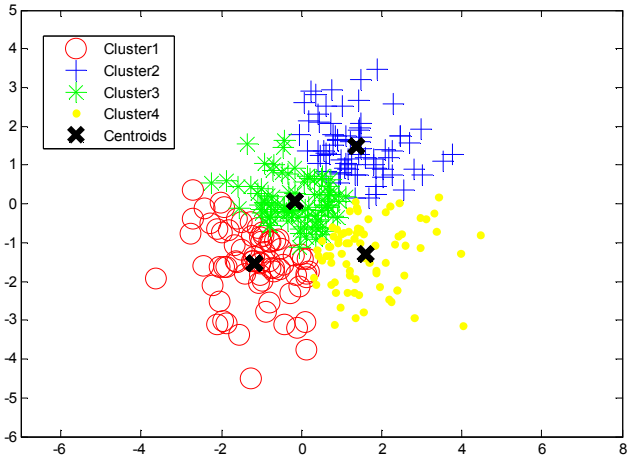

Figure 1. *Results of K means clustering.*

Table 1. *Test results of K means clustering.*

	Cluster	Error	Test	
	Mean Square	**Mean Square**	**F**	**sig**
Zscore: current ratio (X1)	282.526	0.550	816.131	0.000
Zscore: quick ratio (X2)	291.096	0.536	506.411	0.000
Zscore:interest coverage ratio (X3)	93.625	0.852	67.898	0.000
Zscore: the rate of return on total assets (X4)	315.316	0.497	411.929	0.000
Zscore: the rate of return on net assets (X5)	216.551	0.655	221.583	0.000
Zscore: the ratio of profit to net sales (X6)	78.984	0.875	69.855	0.000
Zscore: inventory turnover (X7)	36.020	0.944	24.302	0.000
Zscore: the growth rate of main business revenue (X8)	21.440	0.967	15.254	0.000
Zscore: the ratio of total cash debt (X9)	134.125	0.787	98.192	0.000

Secondly, we use SPSS software to analysis efficiency of K means clustering. The result shows that the previous enterprise credit rating is classified one group, other nine financial indicators are divided into three groups. After 10 iterations, the algorithm changes little in the clustering center. Table 1 shows significance level of K means clustering of sample data, which is below 1%. There are significant differences between the nine indicators in the three groups, so the K means clustering results are effective.

Because of the different length of each K means clustering result, adding clustering results to each layer of GMDH in different sequences can effect computational speed and accuracy. This paper uses matlab software to evaluate computational speed and accuracy of K-means GMDH models with different adding order, selects the optimal model with highest accuracy in a certain operating time. This method can avoid collinearity problem, reduce the computational complexity and improve the accuracy of the evaluation results.

Table 2 shows the results of computational speed and accuracy of six K-means GMDH models. The first layer of GMDH is added to the third clustering group, the second layer is added to the second clustering group and the third layer is added to the first clustering group, which consumes a little time and gets the highest accuracy. So the paper chooses this K means-model as optimal enterprise credit rating model.

Table 2. *Time and accuracy of different K means clustering GMDH models.*

Clustering results added to the GMDH model in different order	Time(s)	Test MSE(%)	Train MSE(%)
1,2,3,4	1.2078	0.0568%	0.0244%
1,3,2,4	1.5628	0.064%	0.0267%
2,1,3,4	0.8486	0.0721%	0.0348%
2,3,1,4	0.9078	0.0654%	0.0498%
3,1,2,4	1.312	0.0533%	0.0227%
3,2,1,4	1.2702	0.0371%	0.0166%

Fig. 2 and Fig. 3 respectively show the accuracy of K means-GMDH model on train data and test data, the variation tendencies of estimated value and real value are almost identical. *MSE* and *RMSE* are statistical indicators for measuring the deviation between observed value and real value. The smaller the value of the two indexes, it represents that the model has a better accuracy. The value of traning data *MSE* is 0.0166%, *RMSE* is 1.29% and the value of test data *MSE* is 0.0371%, *RMSE* is 1.93%, indicating the model has high predicted accuracy. In conclusion, this method can effectively evaluate credit rating.

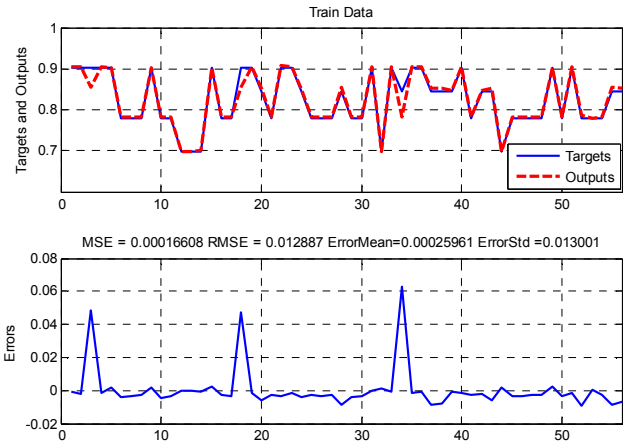

Figure 2. *Accuracy of K means GMDH model on train data.*

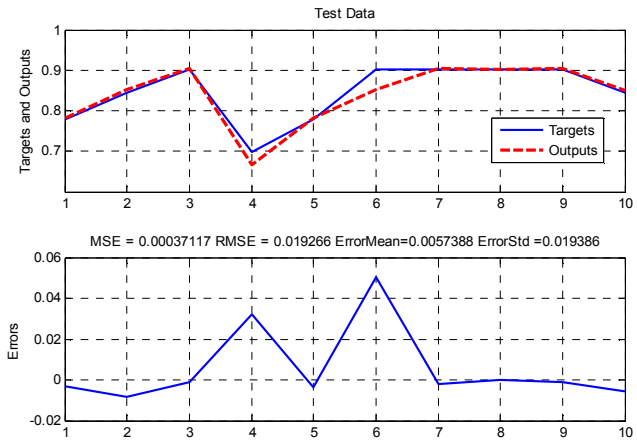

Figure 3. *Accuracy of K means GMDH model on test data.*

In order to directly test the effectiveness of K-means GMDH model, this paper compare the results of this model with those of hidden Markov GMDH, traditional GMDH, multilayer perceptron, radial basis function, artificial neural network model on the same data set.

Table 3 summarizes statistical results of the six neural network models, shows that MSE of hidden Markov GMDH, traditional GMDH, multilayer perceptron, radial basis function, artificial neural network on train data set and test data set are all higher than the same indicators of K-means GMDH, which indicates K-means GMDH model can adequately mining big data information and improve evaluation accuracy.

Table 3. *Compared results of six neural network models.*

	Train data MSE(%)	Test data MSE(%)
Radial basis function	20.3	22.9
Multilayer perceptron	7.7	9.1
Artificial neural network	1.04	1.42
GMDH	0.144	1.2
HMM-GMDH	0.1398	0.0922
K means-GMDH	0.0166	0.0371

5. Conclusion and Prospect

At first, this article theoretically proves that K-means GMDH model can minimize the distance between each clustering centers and the intermediate variables in the least squares estimation, namely this model can get optimal predicting results. Then, we use K means clustering algorithm to quickly mining big data information, each kind of results obtained by clustering is added as an input variable to each layer of the GMDH model in a certain order with the previous filtering variables. Next, the paper compares time and accuracy of K-means GMDH with those of hidden Markov GMDH, traditional GMDH, multilayer perceptron, radial basis function, artificial neural network, which shows K-means GMDH model, an effective predicting method for enterprise credit rating, can get the best prediction in less time.

With the development of data mining, this technology has been highly attracted attention by scholars in enterprise credit rating. The application of data mining technology is worth further exploration in credit risk management.

Acknowledgements

This paper is one of the phased achievements of National Social Science Fund 《Research on Aggregation, Diffusion and Evolution of Chinese Economic Shocks from Sovereign Credit Rating Downgrade and Policy Choices》 (14BJY174).

References

[1] Xiaoyang Zhou et al. The issuer-pays business model and competitive rating market: rating network structure [J]. The Journal of Real Estate Finance and Economics,2016,53:1-26.

[2] Bolton Patrick, Freixas Xavier, Shapiro Joel. The credit ratings game [J]. The Journal of Finance,2012, 67(1):85-111.

[3] Ping He, Meng Jin. The influence of credit rating in Chinese bond market [J]. Financial Research,2010,358(4):15-28.

[4] Zonglai Kou, Yuzhang Pan, Does China's credit rating really affect the cost of issuing bonds? [J]. Financial Research, 2015, 424(10):81-98.

[5] B. Shi, G. Chi. A Credit risk evaluation index screening model of petty loans for small private business and its application [J]. Advances in information Sciences and Service Science- s, 2013, 5(7):1116-1124.

[6] K. Kim, H. Ahn. A corporate credit rating using multi-class support vector machines with an ordinal pairwise partitioning approach [J]. Computer &Operation Research, 2012, 39(8):1800-1811.

[7] P. Hajek, K. Michalak. Feature selection in corporate credit rating prediction [J]. Knowledge-Based Systems,2013,51:72-84.

[8] Ge-Er Teng, Chang-Zheng He. Jin Xiao. Customer credit scoring based on HMM/GMDH hybrid model [J]. Expert Systems with Applications, 2013,36:731-747.

[9] Anastasios Petropoulos, Sotirios P. Chatzis, Stylianos Xanthopoulos. A novel corporate credit rating system based on Student's t hidden Markov models [J]. Expert Systems with Application, 2016,53:87-105.

[10] Robert J. Elliott, Tak Kuen Siud,e, Eric S. Fung. A Double HMM approach to Altman Z-scores and credit ratings [J]. Expert Systems with Applications, 2014, 41:1553-1560.

[11] Yixiang Tian. Comparative analysis and empirical analysis of the different level of GMDH algorithm in the medium and long term forecasting model [J]. Forecasting,1999,6:73-75.

[12] Geer Teng, Changzheng He, Jin Xiao,Yue He, Bing Zhu, Xiaoyi Jiang. Cluster ensemble framework based on the group method of data handling [J]. Applised Soft Computing, 2016,43:35-46.

[13] Hao Zhang, Guanglong Dai. Improvement of distributed clustering algorithm based on min-cluster [J]. Optik, 2016, 127:3878-3881.

[14] Jiali Lin. Analysis on the factors influencing the rating quality of credit rating agencies [D]. Hang Zhou: Zhe Jiang University, 2014:30-56.

[15] Yechen Qin,Reza Langari,Liang Gu. A new modeling algorithm based on ANFIS and GMDH [J]. Journal of Intelligent & Fuzzy Systems, 2015, 29:1321-1329.

[16] Michele Azzollini, Vincenzo Pacelli. An Artifical Neural Network Approach for Credit Risk Management [J]. Journal of Intelligent Learning Systems and Applications, 2011, 3(2):103-112.

[17] Qiumin Li, Yixiang Tian. The k-nearest neighbor-based GMDH prediction model and its application [J]. International Journal of Systems Science, 2014, 45(11):2301-2308.

[18] Zengguang Li, Jin Wang. Improvement of GMDH parameter model and its application in coal price research [J]. Systems Engineering, 2012. 30(6):105-110.

[19] Wei Hu. Improved hierarchical K means clustering algorithm [J]. Computer Engineering and Applications, 2013, 49(2):157-159.

[20] Miaojing Li, Jiyi Wu. Credit evaluation of small and medium sized enterprises in e-commerce environment [J]. Systems Engineering Theory and Practice, 2012, 32(3):555-560.

[21] Haoming. Zhang, Chunyan Miao, Zhiqi Shen. Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings [J]. Neurocomputing, 2014, 128(5):285-295.