2014

ISSN : 0974 - 7435 Volume 10 Issue 21

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(21), 2014 [13322-13328]

Credit risk classifications of e-commerce based on KPCA-MPSO-ANN

Jianping Wu [1*], Bangzhu Zhu[2]

[1]School of Business, Lingnan Normal University, Zhanjiang, 524048

[2]School of Economics and Management, Wuyi University, Jiangmen, 529020

ABSTRACT

The paper attempts to classify E-commerce credit risk in a clearer way by adopting KPCA, MPSO and ANN. In the KPC classification, the data was pre-processed in the first place, and then the eigenvalues and eigenvectors were extracted to reduce the dimensions of the E-commerce credit risk. Furthermore, the study searched the inertia weight and threshold of BP neutral network through the improved MPSO, and determined the inertia weight and threshold value BP neural network. The data of 13 enterprises was trained first, and that of another 5 enterprises was tested and predicted. And finally, the result was classified. The study proved that the KPCA-MPSO-ANN based analysis was quite effective, providing a sound basis, reference and empirical case for classification and evaluation of E-commerce enterprises. Besides, it is of some help to promote the development of E-commerce industry.

KEYWORDS

KPCA,improved MPSO; ANN; Credit risk classification.

INTRODUCTION

On September 22nd, 2014, the Wall Street Journal reported the news that the e-commerce giant Ma's Alibaba Group Holding has successfully listed in American market, which enabled Alibaba officially to be the largest IPO ever in history with a finance amount of $25 billion, and become the second largest Internet companies globally just after Google. This is the inevitable outcome of the development of electronic commerce. However, the e-commerce credit risk still exists, which makes the majority of scholars become greatly interested in it. There are few studies on the classifications of e-commerce credit risk. And in terms of level, the credit risk can be classified into level 1 to 4. In the trading of electronic business, the enterprises of better performance are preferable to choose, while the e-commerce credit risk level is dynamic, and how to choose enterprises among the various many is worthy of studying, and the distinguishing of the risk levels is the key to solve the problem. Since there are few studies on e-commerce credit risk classifications, the paper borrows other classifications to illustrate this problem. Yu Le'an and wang shouyang[1], studied and classified the stocks in Shanghai and Shenzhen by adopting Kernel Principal Component Clustering (KPCC); Min Dan conducted a research on stock and its classifications in light of Adaptive Resonance Theory (ART); Yang Mengzhao and Zhao Chunyang[3] studied stocks through employing Rough sets (RS) and Principal Component Analysis (PCA). Besides, Yu Le'an and Wang Shouyang[4] classified the credit risk based on Kernel Principal Component Analysis (KPCA) and Least Squares Fuzzy Support Vector Machine (LSFSVM), and Li Yunfei & Hui Xiaofeng[5] categorized the investment value of stocks using Support Vector Machine (SVM). Moreover, some of the scholars studied the classification of credit risks by adopting Artificial Neural Network (ANN) [6], Evolutionary Algorithms (EA)[7], K-Nearest Neighbor Algorithm (KNNA)[8]and Support Vector Machine (SVM) [9], which all have been proved to be greatly effective.

The previous studies show that there are various factors influence the risk of e-commerce enterprises. To study and classify the e-commerce credit risks of some good enterprises by taking the advantages of the financial information disclosure of the listed companies might be a sound solution while there are still two problems. First of all, due to too many financial indicators of listed enterprises, some scholars have to apply the Principal Component Analysis (PCA) to reduce dimensions. Whereas, the PCA a method based on linear conversion, which means that it is ineffective in solving the non-linear problems. In indicators of e-commerce credit risk, there are some non-linear problems. It has also been proved in previous studies that it is inappropriate to reduce the dimensions of e-commerce credit risk with only linear methods. In the second place, previous neural network showed some shortcomings such as fitting, poor generalization performance, slow convergence speed. And if cooperate it with Particle Swarm Optimization (PSO) algorithm, and determine the inertia weight and threshold of neural network with PSO, the PSO is easy to fall into local minimum values, because the PSO makes the results of the neural network inaccurate.

In view of the above two problems, this paper, by using KPCA, improved PSO algorithm MPSO and ANN to establish the frame model to classify e-commerce credit risk. As for the first problem, the KPCA is applied to solve the problem of extracting non-linear feature vector and for the second, MPSO algorithm is adopted to solve the local minimum values by dynamically search and regulate the inertial weight and threshold of neural network. Therefore, the paper, based on the principles of KPCA, MPSO and ANN, conducted a research on the classification of e-commerce credit risk by studying the data of risk indicators from 18 listed e-commerce enterprises.

COMBINATION METHOD OF KPCA-MPSO-ANN

Rationale of KPCA

The Kernel Principal Component Analysis (KPCA) is a multivariate statistics method, which applies Kernel Method to the principal component analysis. In this process, multiple variables are mapped to a high-dimensional space through nonlinear function and the principal component analysis is carried on in the high dimensional space. On the basis of keeping the relevant variables unchanged, strive to make the original data information loss minimum, and conduct a comprehensive evaluation and analysis on the system to get the several variables that can reflect the characteristics of the original problem [10]. It has been widely used in nonlinear feature extraction, classification and pattern recognition, known as a kind of non-linear KPCA feature extraction method. However, it has never been applied in the classification of e-commerce credit risk. KPCA algorithm is helpful to avoid the complex problem of numerically solving the eigenvectors in feature space, which is about caculating the eigenvectors and eigenvalue of the kennel matrix.

Make a non-linear transformation of the original data space, map φ : $x \to X$, $R^N \to F$ to the feature space F, suppose F satisfies $\sum_{t=1+m}^{N} \varphi(D_t) = 0$, the non-linear KPCA then can be regarded as the principle component in F. In (1， 2， …， N), N is the length of the extending sequence, $\varphi(\bullet)$ refers to the non-linear mapping, and m is the embedded dimension. $\bar{c}$ is used to calculate the covariance matrix.

$$\bar{c} = \frac{1}{N-m} \sum_{t=1+m}^{N} \varphi(D_t)\varphi(D_t)^T \qquad (1)$$

In the equation (1), all the eigenvalues $\lambda(\lambda \geq 0)$ and elgenvetors V satisfy the following equation.

$$\lambda V = \overline{C}V \tag{2}$$

The linear expression of the sample vectors that V can be mapped to the feature space is:

$$V = \sum_{t=1+m}^{N} \alpha_t \varphi(D_t) \tag{3}$$

In (3), α_t is the equation coefficient and V is the feature vector in F.

Multiply (2) by $\varphi(D_t)$, it becomes

$$\lambda(\varphi(D_t), V)\lambda = (\varphi(D_t), \overline{C}V) \tag{4}$$

Define a matrix of (N-m) ×(N-m)as K

$$k_{ij} = K(D_i, D_j) = (\varphi(D_i), \varphi(D_j)) \tag{5}$$

I,j=1+m,…,N **(6)**

In (5), the calculation of equation coefficient a_i can be transformed into calculating eigenvectors and eigenvalues of matrix K

$$(N-m)\partial\lambda = \alpha K \tag{7}$$

In(7),the column vectors formed by ∂ is $\partial_t (t = 1,2,..., N-m)$, conduct normalization processing of V, at the time, the mapping of the time-delay series $\varphi(D_t)$ in V is as follows:

$$[V\varphi(D)] = \sum_{i=1}^{m} x_i^k [\varphi(D_t \cdot \varphi(D)] = \sum_{i=1}^{m} x_i^k k(x_i, x) \tag{8}$$

And the synthesizing evaluation function of KPCA is:

$$F(x) = \sum_{k=1}^{r} \sum_{i=1}^{m} \omega_k \alpha_i^k k(x_i, x) \tag{9}$$

In (9), ω_k is in the correspondence with the Kth contribution rate of the principle component, $\lambda_1, \lambda_2, ..., \lambda_n$ is the elgen values, and $\alpha_1, \alpha_2, ..., \alpha_n$ the feature vectors, the result of $\lambda_i / \sum_{t=1}^{m} \lambda_t$ refers to the contributions of λ_1 in the total variance.

$$(\sum_{t=1}^{k} \lambda_t / \sum_{t=1}^{m} \lambda_t) \geq 85\% \tag{10}$$

In (10), K is for the number of the selected principal components, 85% refers to the percentage of the selecting threshold, when the contribution rate of k principal components is greater than or equal to 85%, the first k principal components is the number of the selected components.

In the above deduction, if $\sum_{t=1+m}^{N}\varphi(D_t) \neq 0$, then K can also be shown as k^*

$$K^* = K - LK - KL + LKL \tag{11}$$

In (11), L represents the coefficient, the unity matrix, the eigenvalues and eigenvectors of CPCA can be evaluated as stated above.

Rationale of MPSO

Particle Swarm Optimization (PSO) was an intelligent algorithm put forward by the American social psychologist James Russell Kenned and electrical engineers Eberhart in 1995, inspired by the social behavior in nature of crowds of fish and birds, which is based on a simple social pattern. For PSO, each "particle" can be compared to a bird in the searching space, which can be used to solve and optimize problems. All particles search, follow and memorize the current optimal particles in the solution space which has a fitness value determined by optimization function and a speed that determines their flight direction and distance. PSO algorithm finally finds the optimal solution through iteration, but initially it needs to be initialized into a group of random particles. By chasing two extreme values, particles, in each iteration, would update their locations. One is global extreme value pbest, which is the optimal solution at current found by the entire group, the other is individual extreme value pbest, the optimal solution found by the particle itself. When the present fitness value equals to the previous one, the particle swarm will fall into the local minimum value. In order to the problem, PSO is improved in the paper by applying dynamic adjusting inertia weight.

When the present fitness value equals to the previous one, the particle swarm will fall into the local minimum value. So it can be predicted that in all foreseeable iterations afterward, if the two inertia weights are equal, it will no longer produce such an equivalent inertia weight. Moreover, if the number of iterations afterwards is the fitness value of inertia weight produced by the current iteration, the inertia weight is bound to change, and this change is irreversible, which will make the PSO fall into local extremum. The following is the method of improved PSO. The previous improved algorithms of the inertia weight were done by the following equation:

$$z = z_{\max} - \frac{z_{\max} - z_{\min}}{iter_{\max}} \times iter \tag{12}$$

In (12), iter is the number of the present iterations, $iter_{\max}$ is the maxium number of iterations, $z_{\min}$ is the minimum value of inertia weight, while $z_{\max}$ is the maximum inertia weight. It is improved on the basis of (12), and there are serious disadvantages in the above improvement of inertia weight, because the improvement relies on the current number of iterations and the maximum number of iterations, which will make it fall into a local minimum value. To avoid these deficiencies, this paper proposes a dynamic adjustment algorithm of inertia weight, a method that increases first and decreases later, controlled by the algorithm itself to improve PSO.

The basic idea is: with the dynamic inertia adjustment strategy, firstly, choose the inertia weight of the particles of smaller initial values to make strong initial developing ability of particles. Compare the previous fitness value of iteration f_{t-1} with the current fitness value, if the two fitness values satisfy $f_t=f_{t-1}$, the particles will fall into local minima. To strengthen the overall searching ability of particles, the linear function: z = a z + z (a is the random number in the range of 0-1) could be applied, which will help to increase the inertia weight of particles, then the particle will be enabled to jump out of the trap of the local extreme values. But when the inertia weight keeps increasing to 1.4, the exploring ability of particles is the strongest, while the connection among particles will be the weakest, that is, the particle developing ability comes to the weakest. On the contrary, to strengthen the developing ability of particles, it can also use a linear function: z = z - a to z (a is the random number in the range of 0-1),then the connections among partials will be enhanced by reducing the inertia weight, repeat the method, then make the particles themselves control and adjust inertia weight dynamically, and the connection among partials grows stronger and stronger, so the exploring ability and development capacity will be in a dynamic and strategic balance. The adjustment formula of inertia weight is shown as follows:

$$z = \begin{cases} z + a \times z & f_t = f_{t-1}\text{且} \quad z < 1.4 \\ z - a \times z & f_t = f_{-1}\text{且} \quad z \geq 1.4 \\ z & f_t \neq f_{t-1} \end{cases} \tag{13}$$

In (13), “a”refers to the random number in the range of 0-1, z represents the linear function, f_t is the fitness value of tth generation, and f_{t-1} refers to the fitness value of t-1th generation. The above formula expresses the dynamic balance of the particle swarm in overall and local searching[11].

The rationale of ANN

Artificial Neural Networks (ANN) is widely used in economic management and other fields, known as an excellent nonlinear approximating and assessment tool. It is unlike traditional statistical method with strong learning ability. Without the description in advance of the internal relations between data, function or distributions used can meet the requirements of any accuracy and function. ANN is not just a model of nonlinear system, strong it has generalization ability, thus can also be used for knowledge reasoning. MLP network and BP neural network are taken as examples to illustrate in this paper.

(1)Model structure of MLP (P;q)

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i \vartheta(\sum_{j=1}^{p} \beta_{ij} y_{t-j} + \beta_{j0}) + \alpha_t \quad \textbf{(14)}$$

In (14), y_t is the output of the neural network, β_{ij} is the link weight from jth element to inputted vector to ith neuron, $\beta = (\beta_1, \beta_2, ..., \beta_q)'$, and the deviation α_i is independent identically distributed, $\vartheta(\bullet)$ is with nonlinear characteristics which is used to activate the function of the hidden layer neurons[12].

(2) BP artificial neural network

BP artificial neural network is usually composed of input layer, a number of hidden layer and output layer. A neuron is expressed by a node, and several nodes from a group. The lower layer upper layer nodes are connected by weight, using the Internet connection to enhance the communication among layers, and there is no connection among the nodes in each layer. The typical BP neural network is a network model of a hidden layer with three layers, as shown in figure 1.

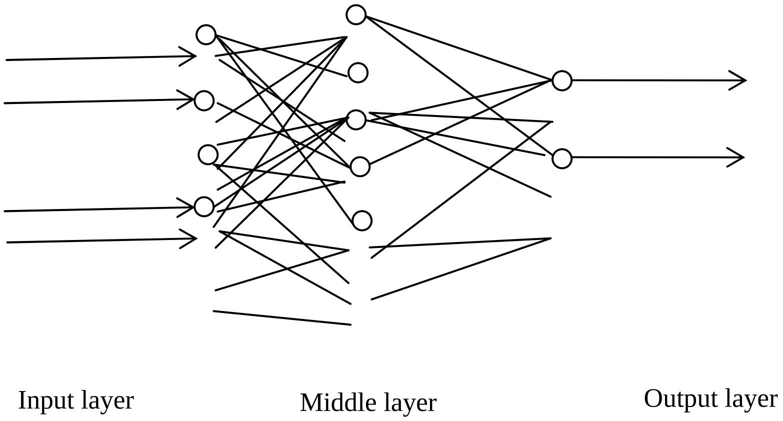

Figure 1 : The structure of BP neural network model

In the neural network model, y_t can be expressed by the following formula:

$$y_t = \beta_0 + \sum_{j=1}^{n} \beta_j \vartheta(\beta_{0j} + \sum_{i=1}^{m} \beta_{ij} y_{t-1}) + \alpha_t \quad \textbf{(15)}$$

In (15), β_{ij}, j=1,2,…,n, β_j, i=1,2,…,m; j=0,1,2,…,n, is the model parameter; n refers to the node number of the hidden layer; m is the node number of the input layer, and $\vartheta(.)$ is the transfer function, whose function is

$$\vartheta(x) = \frac{1}{1+e^{-1}}.$$

So the BP neural network model described by (16) actually reflects the nonlinear function mapping relation between observation value input at the early stage and output $y_{i,}$ shown as follows:

$$y_t = f(y_{t-1}, y_{t-2}, ..., y_{t-p}, \beta) + \alpha_t \tag{16}$$

In (16), β is the parameter vector of the whole model; f (.)is decided by network structure and weight parameters, and BP neural network is the representative model of ANN[13].

CASE STUDY

In order to test the classifying effect of KPCA-MPSO-ANN model, the paper borrows the data from Wang Xinhui's Mater thesis[15], "On the International E-commerce Credit Risk Warning Based on BP Neural Network".

Choosing study samples and indicators

The study samples in this paper are the 18 e-commerce companies from Wang Xinhui's[14] master thesis and the indicators selected are: ratio of sales X_1, net assets income rate X_2, return rate of total assets X_3, total assets turnover X_4, ratio of profits to cost X_5, stock turnover X_6, accounts receivable turnover X_7, current asset turnover X_8, times of interest earned X_9, digital certificates X_{10}, quick asset ratio X_{11}, current ratio X_{12}, rate of capital accumulation X_{13}, platform service credit X_{14}, rate of fixed assets renewal X_{15}, total assets growth rate X_{16}, delinquencies X_{17}, payment delay X_{18}, growth rate X_{19}. The above 19 indicators is the classifying indexes of the e-commerce credit risk. Make matrix with the original data of the 18 e-commerce enterprises credit risk indicators, the 18 enterprises are expressed by A to Z.

TABLE 1 is about the 18 e-commerce companies. Due to the limited space, among the 19 indicators, only some of them are listed in the table. Before the model calculation, the risk should be divided into four levels, namely, risk level 1, risk level 2, risk level 3 and risk level 4.

Risk level 1 shows the excellent state. The overall grading is in the range of (75,100), and the expected output [1000]. This level shows that e-commerce enterprises are running well, and there's no risk at all, so it can be operated without anxiety, but also should be aware of sudden risks.

Risk level 2 shows the normal state, and the overall grading is in the range of (50,75], expected output [0100]. The level shows that e-commerce enterprises are operating normally. There is no possibility of risk. It should be operated boldly, but also should be alert to the possibility of potential credit risk and sudden risk.

Risk level 3 is for slight risk status, and the overall grading is in the range of (25, 50], expected output [0010]. It shows that operating loss of the e-commerce enterprises is small and the risk consequence various, but it will not exert great effect on the e-commerce activities. It needs to strengthen the prevention and control of credit risk.

Risk level 4 is for severe risk status, and the overall grading is in the range of (0, 25], the expected output [0001]. The risk level suggests that the operating loss of e-commerce enterprises is serious or even severe. It must take effective precautionary measures and prepare control plans as well as emergency measures. Besides, it would better to eliminate the factors lead to key risks and reduce the loss.

Data processing

According to the combined treatment method in 2.4, conduct standardizing transformation of the original data, including converting backward indicators to foreword ones and do normalization processing of the data, and then it will obtain standardized e-commerce credit risk matrix. With KPCA method, make linear dimension reduction of the inputted elgenvalues, till the cumulative contribution rate reaches 100%. And then use MPSO to optimize the data, and ANN to classify the credit risk of the enterprises.

CONCLUSION

In order to improve the credit risk prevention capacity of e-commerce enterprises, this paper proposes to use KPCA-MPSO-ANN to classify the e-commerce credit risk, and analyze the related sample data. The research proved that KPCA-MPSO-ANN combination is effective for the classification of the e-commerce credit risk, and compared with other models, the mode is simple in parameters, fast in computing speed and sound in model fitting. Due to difficulties in sample collection, the data applied in the paper is quite limited. However, this paper still proves that the classification of e-commerce credit risk can be effective in guiding the modern economic construction, and thus provide certain help for the government decisions.

REFERENCES

[1] Yu Le'an, Wang Shouyang; *Stock Categorization Based on KPC Clustering methodology*, Systems Engineering – Theory & Practice., **1,** 29 **(2009)**.

[2] Min Dan; *The Application of self-adaptive Resonance Model in Classification of Stocks* J.Taxation and Economy., **52,** 3 **(2006)**.

[3] Yang MengZhao, Zhao Chunyang,Gu Zeyuan; *On the Stock Classification Based on the PCA and Rough Set*[J], Science Technology and Engineering., **1092,** 4 **(2009)**.

[4] Yu Le'an, Wang Shouyang; *Least Square Fuzzy Support Vector Machine Methodology with Variable Penalty Factors for Credit Classification and Its Application Based on KPCA,* System Science & Math Science., **1311,** 29 (2009).

[5] Li Yunfei, Hui Xiaofeng; *The Classification Model for Stock Investment Value Based on SVM*[J], China Soft Science., **135,**1 **(2008)**.

[6] L.Yu, S.Y.Wang, K.K.Lai; *Credit Risk Assessment with a Multistage Neural Network Ensemble Learning Approach,* Expert Systems with Applications., **1434,** 34 **(2008)**.

[7] M.C.Chen, S.H.Huang; *Credit scoring and rejected instances reassigning through evolutionary computation techniques,* Expert Systems with Applications., **433,** 24 **(2003)**.

[8] F.Glove; *Improved linear programming models for discriminant analysis,* Decision Science., **771,** 21 **(1990)**.

[9] Wu Tiebin, Liu Yunlian, Li Xinjun, Yin Yongsheng; *KPCA-fuzzy Weighted LSSVM Prediction Method and Its Application,* Computer Measurement & Control., **617,** 20 **(2012)**.

[10] Tian Zhongda, Gao Xianwen, Li Kun; *Networked Control System Time-delay Prediction method Based on KPCA and LSSVM,* Systems Engineering and Electronics., **1281,** 35 **(2013)**.

[11] Zhang Dan, Han Shengju, Li Jian, Nie Shangyu; *On BP Algorithm Based on MPSO, Computer Simulation.，* ***147,*** *28* ***(*****2011)**.

[12] Su Zhi, Fang Ming, Li Zhigang; *STAR& ANN Model: on Nonlinear dynamic Characteristics of Securities Price and its Forecast,* Chinese Journal of Management Science..**9,** 16 **(2008)**.

[13] Zhao Chengbo, Mao Chunmei; *Forecast of Intensity of Carbon Emission of China based on ARIMA and BP Combination Model,* Chinese Journal of Oceanology and Limnology., **667,** 21 **(2012)**.

[14] Wang Xinhui; *On the Internetal E-commerce Credit Risk Forewarning Model Based on BP Neural Network*[D], Shenyang University of Te16chnology., **(2008)**.