Open Journal of Business and Management, 2019, 7, 1553-1563
https://www.scirp.org/journal/ojbm
ISSN Online: 2329-3292
ISSN Print: 2329-3284

Research on P2P Credit Risk Assessment Model Based on RBM Feature Extraction—Take SME Customers as an Example

Jianhui Yang, Qiman Li, Dongsheng Luo

School of Business Administration, South China University of Technology, Guangzhou, China
Email: liqiman@foxmail.com

How to cite this paper: Yang, J.H., Li, Q.M. and Luo, D.S. (2019) Research on P2P Credit Risk Assessment Model Based on RBM Feature Extraction—Take SME Customers as an Example. *Open Journal of Business and Management*, **7**, 1553-1563.
https://doi.org/10.4236/ojbm.2019.74107

Received: July 10, 2019
Accepted: August 4, 2019
Published: August 7, 2019

Abstract

This paper combines the nonlinear dimensionality reduction method, and the Restricted Boltzmann machine (RBM algorithm), to assess the credit risk of P2P borrowers. After screening and processing many big data indicators, the most representative indicators are selected to build the P2P customer credit risk assessment model. In addition, after comparing the advantages and disadvantages of linear dimensionality reduction algorithm and nonlinear dimensionality reduction algorithm, this paper establishes a P2P enterprise customer credit risk assessment model based on RBM feature extraction combined with contrast divergence theory. It is concluded that the effect of RBM is better than that of PCA when the same model is selected. The Logistic model performs best in the three models when the same data feature extraction method is selected.

Keywords

P2P, Credit Evaluation Model, RBM Algorithm, Credit Risk

1. Introduction

In the current boom of e-commerce, social networking, Internet finance, P2P consumer credit, consumer finance and other Internet platforms, the central bank's credit reporting has become increasingly prominent in the timeliness, comprehensiveness and hierarchy of data. How to dig deep into the massive information flow of the Internet, develop a big data risk control model based on massive indicators, comprehensively assess the credit risk status of enterprise customers, and provide a judgment basis for financial credit approval of P2P lending platform, have become the core of credit risk model system construction.

Credit risk is a difficult problem in the current P2P industry: from a macro perspective, due to the low barriers to entry of P2P, the uncontrollable macro-risk situation is getting worse. From a micro perspective, most of the P2P platform business is still in its infancy. The operating experience and risk management capabilities of platform operators are generally insufficient, and the development situation is extremely unstable. From this perspective, credit issues remain the cause of large-scale risks in the P2P industry in the future. As China's P2P industry has developed rapidly in recent years, the theoretical research on P2P network lending by domestic and foreign scholars is still closely surrounding the development of Internet platform operations. There is little discussion in the academic community on risk management, security prevention, and industry regulation. Especially in the quantitative assessment of credit risk of P2P enterprise customers, it is still almost empty. In view of this, this paper attempts to draw on the existing research results of credit risk assessment (such as the credit model of traditional commercial banks). After analyzing the credit risk characteristics of P2P industry, the credit risk of P2P borrowers is evaluated by using artificial intelligence method. The credit risk is actually through the machine learning method, by learning the borrower's historical data, to assess its future repayment ability and default risk, and obtain a P2P enterprise credit risk assessment model suitable for China's current national conditions.

2. Literature Review

Many scholars have studied credit risk measurement and evaluation models and adopted a variety of methods. Based on the traditional credit risk measurement model and Ronalce model, Rosenberg & Gleity [1] constructed a new P2P credit risk measurement model, and through simulation, the neural network model can be used to obtain better results. On the basis of the traditional credit risk measurement model, Huang [2] combined with the support vector machine and empirical research on loan default, which shows that the new metric model combined with support vector machine can get better result than the metric model combined with neural network. Puroetal [3] takes multiple factors as independent variables, including the borrower's loan amount, credit rating, current overdue loan amount, debt yield, loan interest rate, etc., constructing a logistic regression model for testing, and obtaining good results. Jiang Wei [4] replaced the training algorithm in BP neural network with improved particle swarm optimization algorithm, and constructed BP neural network algorithm model with improved particle swarm optimization, combined with credit evaluation index system, and finally realized based on improved PSO-BP neural network. The personal credit evaluation model establishes a BP neural network credit evaluation model to quantitatively evaluate the credit of the lender and improve the automation of personal credit evaluation. Liu Chang and Xu Zhuoting [5] analyzed the causes of P2P online loan risk, and established the risk prediction model with the loan data of Lending Club, the world's largest P2P company, and gave the prediction accuracy, in order to provide credit risk management method for domestic P2P companies.

The nonlinear dimensionality reduction method used in this paper, the Restricted Boltzmann machine (RBM algorithm), comes from the field of unsupervised learning, a multi-layer limited Boltzmann proposed by Professor Hinton [6] A deep belief network model composed of machines (RBM)-DBN model. It first learns through a multi-step unsupervised neural network, then adjusts the parameters of the supervised learning, and finally trains the discriminant classifier model. It has the advantage that the traditional neural network can't compete with it—for the initialization of the parameter, it can greatly improve the fitting speed of the multi-stage neural network, thus strengthening the neural network's construction ability. An important component of DBN is the Restricted Boltzmann Machine (RBM) for each layer. At present, the application scenarios of DBN are mainly in the recognition of handwritten fonts, information retrieval, text mining, target object recognition, machine learning and machine translation.

The research on applying DBN and RBM to credit risk assessment has Chen Yanwu [7] proposed an engineering model algorithm for feature extraction using constrained Bozman machines. It can rely on expert experience to Information dimensionality reduction and feature extraction in the database to improve the accuracy of the application credit scoring model. The empirical results show that RBM is an efficient feature extraction and data dimension reduction method. Applying it to the personal credit application scoring model can significantly improve the accuracy of the model algorithm. Zhang Yanxia [8] further optimized the multi-step iterative operation for two different sparse self-encoding RBM algorithms: Sp-RBM and Log-Sum-RBM, combined with the improved idea of Polyak Averaging. From the perspective of empirical results, the accuracy and accuracy of the sparse RBM model are higher than the original RBM model, and Log-Sum-RBM has better characterization ability than Sp-RBM. [9] The author also analyzes the application of different RBM models in the field of credit risk assessment.

3. Research Method

In the classical neural network algorithm theory, Professor Hinton sees the restricted Boltzmann machine (RBM algorithm) as a typical undirected graph, as shown in **Figure 1**. v defined as the visible layer, it represents the input data set in the P2P customer credit risk assessment study. Next, we define h as a hidden layer and apply it to our credit evaluation research, which is a feature extractor. In other words, it is the dimension reduction process. In the middle of the visible and hidden layers, we use W as the neighboring weight between the layers. For the most classic RBM models, all visible neurons and hidden neurons are generally binary variables, that is $\forall i, j, v_j \in \{0,1\}, h_j \in \{0,1\}$ [10].

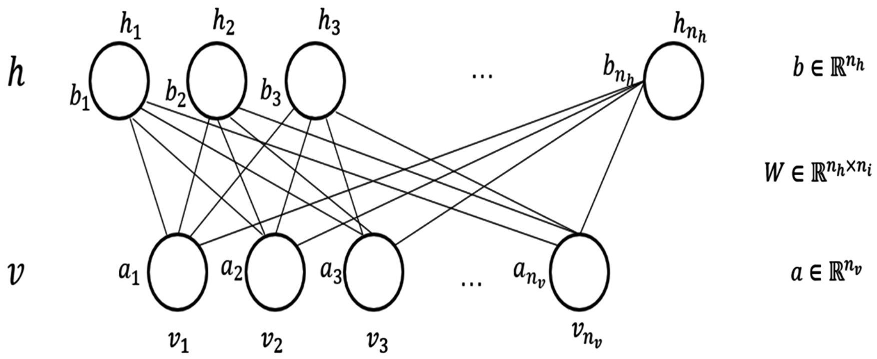

Figure 1. RBM model under classical binary variables.

In different practical applications, the problem we are more concerned with is the distribution of visible neurons v defined by the RBM parameters $P(v|\theta)$.

$$P_\theta(v) = \sum_k P_\theta(v,h) = \frac{1}{Z_\theta}\sum_h \mathrm{e}^{-E_\theta(v,h)} \tag{1}$$

Similarly, applying the pattern of visible neurons to hidden neurons, we have:

$$P_\theta(v) = \sum_k P_\theta(v,h) = \frac{1}{Z_\theta}\sum_v \mathrm{e}^{-E_\theta(v,h)} . \tag{2}$$

In order to find the specific situation of the $P(v|\theta)$ distribution, here we need to solve the normalization factor Z_θ, and estimate it, roughly 2^{n+m} times calculations. In view of this, even if we can obtain the parameters $\omega_{i,j}$, a_j and b_j through the training of the model, we still cannot accurately calculate the unique distribution determined by these parameters.

Of course, it is worth mentioning here that, due to the special structure of RBM neurons, we know that when determining the state of local visible neuron states, in this case the activation states of each hidden neuron are conditionally independent [11].

We record the vector obtained by digging the binning variable h_k at h as $h_{-k} = (h_1, h_2, \cdots, h_{k-1}, h_{k+1}, \cdots, h_{nh})^{\mathrm{T}}$, use the following formulas (3) and (4)

$$a_k(v) = b_k + \sum_{i=1}^{n_v} w_{k,i} v_i . \tag{3}$$

$$\beta(v, h_{-k}) = \sum_{i=1}^{n_\gamma} a_i v_i + \sum_{\substack{j=1 \\ j \neq k}}^{n_h} b_j h_j + \sum_{i=1}^{n_v} \sum_{\substack{j=1 \\ j \neq k}}^{n_h} h_j w_{j,i} v_i \tag{4}$$

we got

$$E(v,h) = -\beta(v, h_{-k}) - h_k a_k(v) \tag{5}$$

Here, $h_k a_k(v)$ and $-\beta(v, h_{-k})$ represent the $E(v,h)$ formula, respectively, and the subscript is equivalent to one side of the *k* and non-*k* variables.

For the hidden neurons at the *j*th, the activation probability formula is as shown in (6) below.

$$\begin{aligned}
&P\left(h_k=1\mid v\right)\\
&=\left(h_k=1\mid h_{-k},v\right)=\frac{\left(h_k=1,h_{-k},v\right)}{P\left(h_{-k},v\right)}=\frac{\left(h_k=1,h_{-k},v\right)}{\left(h_k=1,h_{-k},v\right)+\left(h_k=0,h_{-k},v\right)}\\
&=\frac{\frac{1}{Z}\mathrm{e}^{-E\left(h_k=1,h_{-k},v\right)}}{\frac{1}{Z}\mathrm{e}^{-E\left(h_k=1,h_{-k},v\right)}+\frac{1}{Z}\mathrm{e}^{-E\left(h_k=0,h_{-k},v\right)}}=\frac{\mathrm{e}^{-E\left(h_k=1,h_{-k},v\right)}}{\mathrm{e}^{-E\left(h_k=1,h_{-k},v\right)}+\mathrm{e}^{-E\left(h_k=0,h_{-k},v\right)}}\\
&=\frac{1}{1+\mathrm{e}^{-E\left(h_k=0,h_{-k},v\right)+E\left(h_k=1,h_{-k},v\right)}}=\frac{1}{1+\mathrm{e}^{\left[\beta\left(v,h_{-k}\right)+0*\alpha_k(v)\right]+\left[-\beta\left(v,h_{-k}\right)-1*\alpha_k(v)\right]}}\\
&=\frac{1}{1+\mathrm{e}^{-\alpha_k(v)}}=sigmoid\left(\alpha_k\left(v\right)\right)=sigmoid\left(b_k+\sum_{i=1}^{n_v}w_{k,i}v_i\right)
\end{aligned}\tag{6}$$

Through the derivation of the above formula, we find the formula (7)

$$P\left(h_k=1\mid v\right)=sigmoid\left(b_k+\sum\nolimits_{i=1}^{n_v}w_{k,i}v_i\right).\tag{7}$$

For the symmetric RBM neuron structure map, when we fix the state condition of the hidden neurons, it can be clarified that the activation states of the respective visible neurons are also conditionally independent [12]. Similarly, we derive the independent activation probability of the visible neurons at the *i*th by the derivation of the formula as shown in (8) below.

$$P\left(v_k=1\mid h\right)=sigmoid\left(a_k+\sum_{j=1}^{n_h}w_{j,k}h_j\right).\tag{8}$$

Finally, the activation probabilities for different neurons are:

$$P\left(h\mid v\right)=\prod\nolimits_{j=1}^{n_h}p\left(h_j\mid v\right).\tag{9}$$

$$P\left(v\mid h\right)=\prod\nolimits_{j=1}^{n_v}p\left(v_j\mid h\right).\tag{10}$$

4. Empirical Study

4.1. Data Description

Whether the credit risk assessment model is effective or not, one of the important rating ideas is whether it can accurately identify the potential financial problems of SMEs borrowing from P2P. Therefore, the ideal sample in this section is the SMEs that have borrowed through the P2P platform. However, because the P2P platform does not disclose the borrower's specific information, and most companies that use the P2P platform to raise funds are not listed companies. For non-listed companies, they have no obligation to publish financial statements. Therefore, it is difficult to collect enough sample data to support this empirical study. In order to make this modeling idea go smoothly, the main method of this research is to find potential lending companies and representative P2P lending companies (similar to those of companies that raise funds through P2P platform) for empirical research. In the end, we chose SMEs listed on the GEM.

At the end of 2017, there were 722 SMEs listed on the GEM, excluding some samples with missing values, and there were 599 companies with complete financial data. The paper selects the 2016 annual report data of the enterprise, and the net profit of the following year is used as the label of the economic strength of the borrowing enterprise. If there is a loss in the next year's annual report, the risk of default of the enterprise is considered to be high, and it is regarded as a sample of default, marked as 0. Otherwise marked as 1. The enterprise data comes from wind data, and the executive data comes from web crawlers.

Data is collected form WIND database and WDZJ-OFFICIAL website.

4.2. Indicator Selection

This article establishes individual user portraits through six dimensions: identity information verification label, stability information label, financial application information label, important asset information label, commodity consumption information label, and media viewing information label. Then we will consider enterprise executive information below. In the case, combined with the empirical data, the indicators are embodied, and the P2P enterprise customer credit index system is established.

After fully considering the difficulty of obtaining the indicators, here is a summary of the corporate customer credit pre-selection indicators established by the three dimensions of the company's own label, the company's main executive label, and the external evaluation label.

Among the nearly 200 pre-selected indicator variables, it is necessary to screen out variables with significant effects. In this paper, the credit risk assessment of P2P enterprise customers, the primary problem is to discretize the continuous variables to facilitate the next data grouping and WOE coding, and to solve the IV value.

In the traditional machine learning model, if the data set is improperly discretized, the accuracy of the trained model classification will be greatly reduced. In order to discretize the continuous variables, after considering each model, we finally choose the entropy-based discretization method.

To solve the IV indicator, we need to calculate the WOE (Weight Of Evidence) value in the first step [13]. Combined with the P2P enterprise customer credit risk assessment model to be established in this paper, the dependent variable here is the case that the enterprise loan is overdue and the normal loan is repaid. In fact, WOE is a measure of the proportion of defaults when estimating the value of an independent variable in a particular dimension. If the value of WOE is larger, it means that the dimension is more important.

$$woe_i = \ln\left(\frac{b_i/b_T}{g_i/g_T}\right)\times 100 \tag{11}$$

$$\text{IV} = \sum_i\left(\left(\frac{b_i}{b_T}-\frac{g_i}{g_T}\right)*\ln\left(\frac{b_i/b_T}{g_i/g_T}\right)\right) \tag{12}$$

It can be seen from the above formula (12) that the IV value is calculated by the weighted average of the WOE values, which represents the feature size of the dimension information. In other words, the IV index is the independent variable and the dependent variable result. An associated metric. From the structure of Equation (12), we know that the IV value is always greater than 0, so we can sum the IV values corresponding to the entire group to calculate the overall IV value (Table 1).

Take the operating income indicator as an example to illustrate and explain how to calculate the cabinet (the attributes of the variables) WOE and IV, see Table 2.

The IV value of the operating income index = 0.36 > 0.3, which is an indicator with strong forecasting ability. In the actual data analysis, sometimes the variable with the IV value between 0.01 and 0.02 is still significant in the use of Logistic regression. Therefore, this paper adopts a conservative approach and only excludes variables with IV less than 0.01 (Table 3).

Table 1. Predictive ability of value indicators.

Range of IV	Predictive power
<0.02	NO
0.02 - 0.10	WEAK
0.1 - 0.3	MEDIUM
>0.3	Strong

Table 2. WOE and IV calculations after the optimal segmentation of operating income indicators.

	Binning range	Number of default samples	Normal sample number	WOE	IV
1	[14747804.5, 12933539309]	12	156	0.2771371	0.0243930
2	[12933539309, 25852330812]	13	154	0.3700832	0.0450963
3	[25852330812, 38771122316]	6	108	−0.0482852	0.0004343
4	[38771122316, 51689913820]	2	148	−1.4619785	0.2936793
total	-	33	566	−0.8630433	0.3636032

Table 3. P2P corporate customer credit indicators screened according to IV values.

The registered capital	Assets balance	Balance of current liabilities	Operating income
The total number of employees	Notes payable	Total current liabilities	Operating cost
Major shareholders holdings	Accounts payable	Total illiquid liabilities	Operating profit
The top 10 shareholders holding together	Deferred revenue	Long-term accounts receivable	Non-operating income
Other liquid assets	Remuneration payable	The total amount of comprehensive income	Non-business expenses
Net profit/business revenue	Current assets/total assets	The total profit year-on-year growth rate	Total liabilities year-on-year growth rate
Operating profit/business revenue	Non-current assets/total assets	Net profit growth rate	Total assets year-on-year growth rate

Continued

Total operating costs/business revenue	Operating cycle	Year-on-year growth rate of net assets	Cash flows to year-on-year growth rate
Return on equity	Long-term capital debt ratio	Inventory turnover days	Inventory turnover
Return on total assets	Long-term capital fit ratio	Annualized return on equity	Earnings per share
Return on invested capital	Sales net interest rates	Annualized rate of return on total assets	Net assets per share
Rate of return on labor input	Sales gross profit margin	Annual net interest rate of the total assets	EBITDA per share
Profit total	The cost of sales ratio	Taobao purchase index	1688 industry index
Executives at age	Executives of gender	Executive level of education	Executives marital status
Executives phone number use fixed number of year	Social networking sites active number of fans	Senior management years in the industry	Senior management years in this profession

4.3. RBM Feature Extraction

Analysis of the sample data found that there were only 33 samples of the GEM < 0 in 2017. Normal sample: Default sample = 566:33, approximately 17:1. Such data sets can easily fall into the trap of uneven learning. Therefore, we hope to influence the unbalanced sample set by the change of sampling method to obtain a more balanced distribution of data samples.

Considering the sample imbalance problem in the actual P2P enterprise customer credit risk assessment research, here we use the undersampling method, divide the sample set into 17 groups, respectively classify and return with the default sample set, and finally the total classifier model is obtained.

Before using the RBM algorithm for feature extraction of indicators, it is necessary to standardize the index values for Z-Score, $\frac{x-\overline{x}}{\sigma}$ [14] [15]. In addition, we must also establish the number of neurons in the hidden layer. Given that there is no universal standard for establishing the number of hidden neuron nodes in the academic world, the approach we have here is the suggestion of Professor Hinton in the 2012 paper: Depending on the data category of the different samples, the number of hidden neuron nodes must be at least less than the most basic number of bytes in the sample set. The sample set used in this paper has a total data volume of 599 and its logarithm $\log_2(599)=9.226$, so at least 10 hidden neuron nodes should be selected. Based on the reconstruction error after the single-step iteration, we made the following attempt in **Table 4**, and finally selected 40 hidden neuron nodes with the smallest single-layer reconstruction error to train the RBM model.

4.4. Model Comparison

After undergoing the above data preprocessing, we put P2P enterprise data into the machine learning model for classification and prediction, such as SVM, Logistic, KNN and so on. As can be seen:

1) The effect of RBM is better than that of PCA when the same model is selected.

2) The Logistic model performed best in the three models with the same data feature extraction method selected. In general, the RBM-Logistic model has the best classification, with an accuracy rate of 74.87% (**Figure 2**).

5. Result

According to the P2P SME customer credit index system screened by IV value, it can be seen that in the P2P SME credit risk assessment, corporate financial information (such as operating income, net assets growth, current liabilities total ratio, etc.) is a very important ring. This also reflects the full understanding of the financial situation of SMEs, while focusing on the debt situation of enterprises, which is of great benefit to the construction of the credit evaluation index system.

Table 4. Reconstruction errors of different hidden neuron nodes under single-step iteration.

Hidden node	Single layer reconstruction error
10	30.4052
15	29.8549
20	29.548
25	29.5532
30	29.6088
40	**29.0978**
45	29.3006
50	29.5142
55	29.4811
60	34.5148

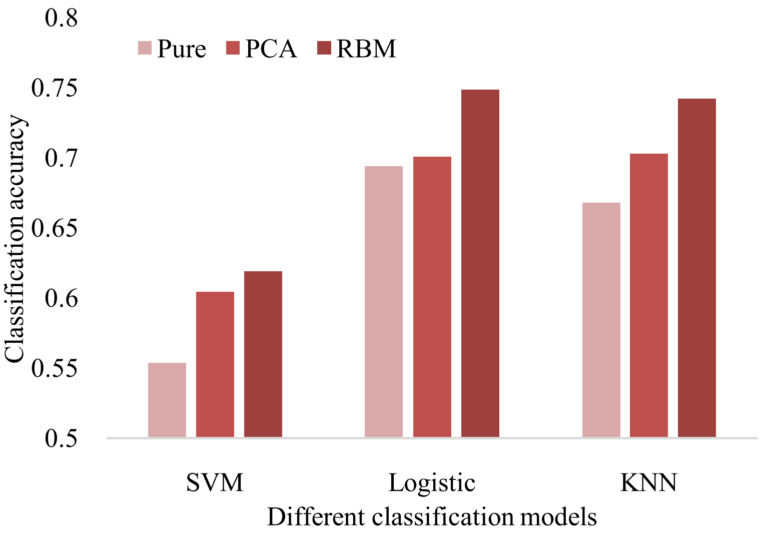

Figure 2. Comparison of different model classification accuracy results.

A good credit risk model can not only reduce the burden of qualification review for SMEs for the P2P credit platform, but also reduce the risk of lending, while also speeding up the financing process for SMEs, which has many benefits for both parties. Therefore, for the P2P online lending platform, it is very important for the P2P business to construct a scientific and reasonable credit evaluation model.

After comparing the advantages and disadvantages of the linear dimensionality reduction algorithm and the nonlinear dimensionality reduction algorithm, combined with the contrast divergence theory, the P2P enterprise customer credit risk assessment model based on RBM feature extraction is established. Finally, it is concluded that the effect of RBM is better than that of PCA when the same model is selected. The Logistic model performs best in the three models when the same data feature extraction method is selected. Therefore, the P2P network lending platform can consider constructing the RBM-Logistic model with the highest accuracy when conducting credit risk assessment for SMEs.

Acknowledgements

The authors wish to acknowledge the valuable support of everyone who helped in the actual administration of the questionnaire and all the other participants of the study.

Fund

Project supported by the Ministry of education humanities and social sciences research project: Research on credit risk evaluation and supervision of P2P (Y9150040).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gleit, R.A. (1994) Quantitative Methods in Credit Management: A Survey. *Operations Research*, **42**, 589-613. https://doi.org/10.1287/opre.42.4.589

[2] Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H. and Wu, S. (2004) Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study. *Decision Support Systems*, **37**, 543-558. https://doi.org/10.1016/S0167-9236(03)00086-1

[3] Puro, L., Teich, J.E., Wallenius, H. and Wallenius, J. (2010) Borrower Decision aid for People-to-People Lending. *Decision Support Systems*, **49**, 52-60. https://doi.org/10.1016/j.dss.2009.12.009

[4] Jiang, W. (2018) Research on Personal Credit Evaluation Model and Algorithm Based on Improved PSO-BP Neural Network. Master's Thesis, University of Electronic Science and Technology, Chengdu.

[5] Liu, C. and Xu, Z.T. (2018) Research on Credit Risk of P2P Network Loan—An

Empirical Analysis of the Lending Club Platform. *Rural Economy and Technology*, **29**, 102-103.

[6] Hinton, G.E. (2002) Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, **14**, 1771-1800. https://doi.org/10.1162/089976602760128018

[7] Chen, Y.H. (2016) Internet Loan Application Scoring Model Based on Feature Extraction of Restricted Pozmann Machine. Master's Thesis, Shanghai Normal University, Shanghai.

[8] Zhang, Y.X. (2016) Deep Learning Model Based on Restricted Boltzmann Machine and Its Application. Master's Thesis, University of Electronic Science and Technology, Chengdu.

[9] Freund, Y. and Haussler, D. (1994) Unsupervised Learning of Distributions of Binary Vectors Using Two Layer Networks.

[10] Li, X., Pang, J., Mo, B., *et al.* (2016) Deep Neural Network for Short-Text Sentiment Classification. In: *International Conference on Database Systems for Advanced Applications*, Springer International Publishing, Berlin, 168-175. https://doi.org/10.1007/978-3-319-32055-7_15

[11] Roux, N.L. and Bengio, Y. (2008) Representational Power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Computation*, **20**, 1631-1649. https://doi.org/10.1162/neco.2008.04-07-510

[12] Tenenbaum, J.B., Silva, V. and Langford, J.C. (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, **290**, 2319-2323. https://doi.org/10.1126/science.290.5500.2319

[13] Pope, D.G. and Sydnor, J.R. (2011) What's in a Picture? Evidence of Discrimination from Prosper.com. *Journal of Human Resources*, **46**, 53-92. https://doi.org/10.1353/jhr.2011.0025

[14] Klafft, M. (2009) Online Peer-to-Peer Lending: A Lenders' Perspective. *SSRN Electronic Journal*, **2**, 81-87. https://doi.org/10.2139/ssrn.1352352

[15] Larrimore, L., Li, J., Larrimore, J., *et al.* (2011) Peer to Peer Lending: The Relationship between Language Features, Trustworthiness, and Persuasion Success. *Journal of Applied Communication Research*, **39**, 19-37. https://doi.org/10.1080/00909882.2010.536844