ISSN: 2688-9323

Research on Credit Decision Making for Micro, Small and Medium Enterprises based on Deep Learning

Fan Chen*, Shiying Wang and Yunjie Zeng

College of Sciences, Beijing Jiaotong University, Beijing, China

Abstract

In 2018, General Secretary Xi pointed out at the symposium on private entrepreneurs that private enterprises play an important role in stabilizing economic growth, promoting innovation, increasing employment and improving people's livelihood, and small, medium and micro enterprises are the backbone of them. However, due to the problems of small scale and low collateral, financing has become an important factor hindering their development. Although the government has issued several policies to avoid their credit risks, it still cannot solve the information asymmetry problem between banks and enterprises at the root. How to establish a credit risk assessment system for MSMEs with high accuracy and generalizability has been the focus of discussion in the financial and mathematical academic circles. Deep learning, as a representative technique of artificial intelligence, can significantly improve the results of prediction-based research in finance. Based on the data set-German Credit and 123 enterprises with credit records, this paper uses principal component analysis to solve the problem of "small amount of data and large missing values" for MSME credit and constructs a credit evaluation model for MSMEs through multilayer neural network. The model outline was constructed based on the research of Zhou, Yaihua, and Wang, Sunan, and the credit rating scoring system and unexpected event risk warning system were introduced, and the loss function (loss), optimizer and other parameters were optimized based on TensorFlow. The credit evaluation system with a fitting accuracy of 88.6% and a prediction accuracy of 83.7% is obtained. This paper aims to use big data processing and deep learning to evaluate the credit risk of lending enterprises, explore the solution of information asymmetry between banks and enterprises, and effectively prevent the credit risk of MSMEs. It also introduces real data of a bank to verify its feasibility in reality, promotes the cross-fertilization research of deep learning, statistics and finance, and provides reference for the effective implementation of promoting the economic development of enterprises in China.

Keywords

MSMEs; Credit Risk; PCA; Multilayer Perceptron Neural Networks; Deep Learning; Rating Systems.

1. Introduction

1.1. Research Background and Significance

Small and micro enterprises is the collective name for small enterprises, micro enterprises, and home-based enterprises. The specific criteria are based on indicators such as the number of employees, business income, and total assets, combined with the characteristics of the industry. Due to their small size and flexible operation, they are an indispensable part of China's transition economy. However, due to various reasons, small and micro enterprises in China have the dilemma of "high cost and high tax burden; difficult to employ and difficult to finance", in which the credit issue is particularly important [1].

Orgler first proposed the use of linear regression analysis in scoring in 1970. Later, nonlinear methods also began to be applied to credit scoring. in 1980, Wiginton applied logistic regression to credit scoring and studied the model effects analytically. The logistic regression method has almost no assumptions about the premises, and because of its relatively high stability, this makes it one of the most widely used assessment methods in the field of credit scoring. After 1980, non-statistical methods also began to be used in the field of credit evaluation. With the rapid development of computer technology, artificial intelligence methods such as neural networks and genetic algorithms were also introduced by researchers in the field of credit evaluation [2].In 2000, West.D research found that least squares support vector machines also have good computational results in credit assessment, but the most widely used is the nonlinear classification method: neural network model, as it provides competitive prediction functions.In 2013, research showed that due to the self-organization, strong robustness, and self-adaptability, making it more preferable to other linear statistical methods by practitioners [3].In 2014, the literature noted that both the output and application of credit assessment are non-unidirectional and contain multiple dimensions. This inherent complexity makes higher demands on the effectiveness of algorithms for accuracy, fitness, error distribution, and output distribution [4]. Artificial intelligence methods such as neural networks and SVMs have excellent performance in automated credit evaluation.

The outbreak of the new crown epidemic in 2019 and the secondary effects of the epidemic have had a direct impact on the survival and development of micro and small enterprises. It is found that banks and MSMEs currently lack an effective avoidance and risk warning system in dealing with the outbreak, and there is no reliable credit scoring for lending enterprises and methods to calculate the probability of enterprise credit failure.

In this context, the credit problem of MSMEs is becoming more and more important. The mathematical community has been paying more and more attention to micro and small credit-related topics and trying to build models or systems using mathematical methods to study and explore feasible solutions. For example, Question C of the 2020 Mathematical Modeling National Competition noted the credit risk of MSMEs scale and required the participating teams to study the credit strategy for MSMEs and the credit adjustment strategy under the influence of unexpected factors by building a mathematical model based on the data information.

If a high-precision credit risk assessment system can be constructed mathematically for the data characteristics of MSMEs, the credit risk of MSMEs can be effectively prevented. It will not only have theoretical value in the related fields of finance and statistics, but also have its practical significance in promoting the economic development of national enterprises.

1.2. National Related Policies

In order to improve economic vitality and ensure the stability of the country's economic structure, the State Council has introduced a series of policies since 2012 to support the healthy development of small and micro enterprises.

In 2015, China implemented the policy of starting point for small and micro enterprises and individual entrepreneurs and the policy of reducing the income tax of small and micro enterprises by half, reducing tax and duty by nearly 100 billion yuan.

In 2019, Premier Li Keqiang said at the second session of the 13th National People's Congress in the Great Hall of the People in Beijing that the financing cost for small and micro enterprises should be reduced by another 1 percentage point from last year's level.

In 2020, under the epidemic, to help small and micro enterprises to tide over the difficulties, the State Council has introduced multiple policies one after another, involving rent, taxes, financing, operating costs and other aspects.

In 2021, the government's work report said that the continuation of the loan deferment policy for general-purpose micro and small enterprises, the extension of the small and micro enterprise financing guarantee fee reduction award policy, large commercial banks general-purpose micro and small enterprise loans increased by more than 30%, appropriate reduction in payment processing fees for small and micro enterprises, this year it is imperative to achieve more convenient financing for small and micro enterprises, comprehensive financing costs steadily decreasing.

The research based on the credit problems of small and micro enterprises, in line with the development trend of the times, is an academic subject conducive to the development of the country and the people's livelihood, and has a certain practical and application value.

2. Introduction to Related Theories

2.1. MLP Neural Network

MLP,Multilayer Perceptron, also called an artificial neural network, is a forward-structured artificial neural network that maps a set of input vectors to a set of output vectors [Defry.Neural Networks 1: Multilayer Perceptron - MLP.2019-02-10]. The neuronal model in neural networks can be analogized to biological neurons, and by comparing the functions of biological neurons and neuronal models of neural networks, we find that the structural properties of biological neurons make it a perfect fit for what we hope to achieve in machine learning-namely, to pass through a series of regular stimulus delivery processes to finally achieve the correct output of the decision outcome [6].

Non-parametric analysis methods represented by neural networks are widely used to analyze the early warning of financial crisis of various enterprises, and ANNs credit risk model stands out with its strong advantage of approximating nonlinear functions, however, the single use of ANNs to evaluate often cannot get particularly desirable conclusions, so this paper introduces MLP to study the risk of small and micro enterprises [7].

In the MLP model, x_j denotes the input value from the jth "dendrite", w_{ji} denotes the connection power (only one unique power on each fixed input), u_i denotes the linear combination of all input signals on that neuron i, and the coefficients are the corresponding weights, i.e.

$$u_i = \sum_j w_{ji} x_j \tag{1}$$

θ_i is the threshold value of this neuron i. The intermediate value v_i is obtained by a simple summation of u_i and θ_i.

$$v_i = u_i + \theta_i \tag{2}$$

and f( ) denotes the activation function and y_i denotes the output of that neuron i, i.e.

$$y_i = f(v_i) \tag{3}$$

The perceptron consists of two layers of neurons, the input layer receives external input signals and passes them to the output layer, and the output layer is the M-P neuron, also known as the threshold logic unit [8].

Multi-layer perceptron structures are more complex, with multiple layers of input and output layers and hidden layer neurons, which are not only much more capable of learning than a single perceptron, but also more compatible with the complexity of metrics and application levels in credit assessment models [9]. A typical multilayer M-L-P artificial neuron model is structured as follows.

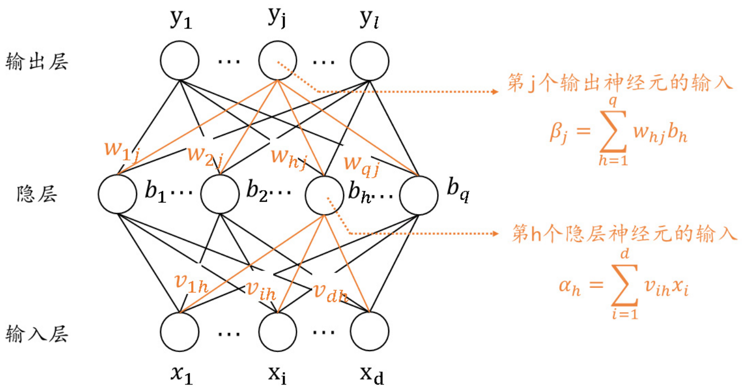

Figure 1. Multi-layer perceptron model

2.2. Credit Indicators

Credit rating index system is a general term for the evaluation elements, evaluation indicators, evaluation methods, evaluation criteria, evaluation weights and evaluation levels used by credit rating agencies in the objective and fair evaluation of the credit status of the evaluated object, and these items form a complete system, which is called credit rating index system [10][11].

With the development and improvement of information database, the credit assessment index system has also changed from a few indicators at the beginning to dozens of more complex indicators today. Early credit scoring mainly relied on experienced experts to score, and there was no set of more scientific and complete indicators for benchmarking. Nowadays, individual indicators are becoming more and more complex, and human judgment of data is often too subjective, which requires statistical tools for effective judgment and selection of indicators [12].

The following figure shows some common indicators.

Table 1. Schematic diagram of three-side measurement

Variable	Description	Impact on credit default
Financial Indicators		
Total Assets Turnover Ratio		-
Inventory Turnover Ratio		-
Current Ratio		-
Cash-to-Revenue Ratio		+
Equity Ratio		+
Profit growth rate		-
Interest coverage multiple		-
ROA		-
ROE		-
Non-financial indicators		
Duration	Duration of business existence	+
Loan History		-
Last year's record		-

Loan rejection in the past year		+
Main business		+(-)
Loan Purpose	Fixed assets (0), circulating assets (1)	+
Loan interest rate		+
Past due rate of loans at the end of the previous year		+
Average overdue loan rate		+
Gender	Female (0), Male (1)	+
Age		+(-)
Percentage of secured loans		-
Percentage of mortgage loans		-
Percentage of credit loans		+
Percentage of medium- and long-term loans		+
Macro Indicators		
GDP		-
CPI		+
Grade Unemployment Rate		-
Interest Rate		+
Generation Manager Index		-

Considering the complexity and diversity of the above information, and the small size of micro and small enterprises, which does not fully reflect the characteristics of the data, as well as the lack of strictness of banks in collecting the data, leads to some missing information in the enterprise credit evaluation of micro and small enterprises. And further leads to the bias of judgment caused by bank staff in evaluating a MSME [13]. The credit evaluation information of micro and small enterprises can be roughly divided into three categories: financial information, non-financial information, and macro indicators, which have certain correlation with each other and directly with each other, and this paper introduces principal component analysis (PCA) to fill in the missing values.

3. PCA-MLP Model

3.1. Data Structure

Micro and small enterprises are collectively referred to as micro and small enterprises, micro and small enterprises, and home-based enterprises. Therefore, the credit behavior of micro and small enterprises is somewhat similar to that of individuals, and in order to improve the accuracy of the model evaluation as much as possible, the team members chose to use a dataset that fits the characteristics of micro and small enterprises-German Credit-for training after the initial establishment of the model. Some of the data are shown in the following table.

Table 2. German Credit partial data set

OBS#	Checking account status	Credit term (in months)	Credit history	Credit limit	Average balance in savings account
1	0	6	4	1169	4
2	1	48	2	5951	0
3	3	12	4	2096	0
4	0	42	2	7882	0
5	0	24	3	4870	0
6	3	36	2	9055	4
7	3	24	2	2835	2
8	1	36	2	6948	0
9	3	12	2	3059	3
10	1	30	4	5234	0
11	1	12	2	1295	0
12	0	48	2	4308	0
13	1	12	2	1567	0
14	0	24	4	1199	0

3.2. PCA Model

Through practical investigation and diversified research on related literature, the credit of micro and small enterprises is characterized by "small data volume and large missing values", which refers to the clustering, grouping, deletion or truncation of data due to the lack of information in rough data. It means that the value of one or some attributes in the existing data set is incomplete. Due to the multiple and complex credit indicators used in credit assessment, omissions in the data collection process of financial institutions or other unexpected factors can lead to some missing data in credit assessment.

Too many missing values will not only cause bias in credit evaluation. Moreover, when pre-processing the data, the simple deletion method is prone to waste of information due to the small sample size, and the calculation of the weight deletion method is too difficult.

In this paper, the missing values are repaired for the characteristics of large missing values of data in micro and small enterprise credit research. Considering the problems such as the type of missing values is not well determined and there are difficulties in comparing the effects, this paper researches the principal component analysis algorithm (PCA) based on the feature of partial correlation between credit indicators, tests the feasibility and compares it with other methods, and finally selects the publicly available German Credit dataset for analysis and testing, and experiments prove that this PCA residual value filling method works better than other The PCA residual value filling method is proved to be better than other algorithms [14].

The principal component analysis algorithm (PCA) is the most commonly used linear dimensionality reduction method, and its goal is to map high-dimensional data into a low-dimensional space by some linear projection and expect the maximum amount of information (maximum variance) of the data in the projected dimension, thus using fewer data dimensions while retaining the characteristics of more original data points.The purpose of PCA dimensionality reduction is to ensure as much as possible that The purpose of PCA is to reduce the dimensionality of the original features while ensuring that "no information is lost", that is, to project the original features to the dimension with the maximum projected information as much as possible, so that the loss of information is minimized after the dimensionality reduction [15].

Some data in German Credit were randomly deleted, and then the missing data set was processed using deletion method, mean repair method, multiple repair method and PCA method respectively, and finally the prediction accuracy and test set accuracy of MLP credit assessment model were used to compare the effect of different missing value processing methods, and some variables (v3, v4, v7, v16 and v28) were cited for comparison.

Table 3. Missing value processing effect comparison

Method		V3	V4	V7	V16	V28	Prediction accuracy	Test set accuracy
Complete Data	Standard deviation	1.17259	0.41559	0.44448	0.28329	0.44043	0.8665	0.752
	Variance	1.37498	0.17272	0.19756	0.08026	0.19398		
Missing data	Standard deviation	1.07818	0.42277	0.44753	0.28951	0.36158	0.7027	0.7
	Variance	1.16246	0.17874	0.20028	0.08381	0.13074		
Deletion Method	Standard deviation	1.088984	0.413399	0.455822	0.264827	0.367267	0.9162	0.6786
	Variance	1.185886	0.170898	0.207828	0.070133	0.134885		
Mean Repair	Standard deviation	1.054741	0.410766	0.436883	0.28262	0.351311	0.7583	0.6588
	Variance	1.112479	0.168729	0.190867	0.079874	0.123419		
Multiple Patching	Standard deviation	1.07241	0.422109	0.448508	0.286182	0.357071	0.8465	0.756
	Variance	1.150064	0.178176	0.201159	0.0819	0.1275		
Principal Component Analysis	Standard deviation	1.07241	0.422109	0.448508	0.286182	0.357071	0.8505	0.78
	Variance	1.150064	0.178176	0.201159	0.0819	0.1275		

As can be seen from the table, the amount and precision of data obtained by different processing methods vary, and the mean, standard deviation, and variance of data obtained after the deletion method are all very different from those between the original data and the missing data, and the prediction accuracy is low. Although the data obtained according to the mean repair method and the multiple repair method are close to the mean value of the original data, the standard deviation and the variance are much different. From the comparison of the above analysis results, it can be seen that the standard deviation and variance of the data are not much different from the original data after the PCA algorithm mends the missing values, and the prediction accuracy is also similar to the complete data, which indicates that the PCA mending method is suitable for repairing the missing values of micro and small enterprise credit data, so this paper decided to use the PCA algorithm to process the original data set [16].

3.3. MLP Neural Network Model

Considering the availability, flexibility and visualization of the python-based open-source package TensorFlow, the team members used TensorFlow to optimize the programming and train the data related to 123 enterprises with credit records, simulating the problems of missing data and low data volume of micro and small enterprises and trying to solve [Favre. Detailed explanation of the principle of multi-layer perceptron & Python and R implementation. 2018-05-07], and obtained the results with higher training accuracy and fitting accuracy.

The model outline is constructed based on the literature "Study on credit risk of micro and small enterprises based on multilayer perceptron neural network", and some parameters are improved and optimized to make it more suitable for fitting a small amount of data. The five-layer neural network used in this paper for model construction 18]. The structure is as follows.

Optimizer: ADam (learning_rate=0.001, beta_1=0.9, beta_2=0.999) The formula is as follows.

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \tag{4}$$

$$m_t{}' = \frac{m_t}{1-\beta_1{}^t} \tag{5}$$

$$v_t = \beta_1 \cdot V_{step-1} + (1-\beta_1) \cdot g_t{}^2 \tag{6}$$

$$v_t{}' = \frac{v_t}{1-\beta_2{}^t} \tag{7}$$

$$\eta = lr \cdot m_t{}'/\sqrt{v_t{}'} \tag{8}$$

$$w_{t+1} = w_t - \eta \tag{9}$$

Loss function: sparse_categorical_crossentropy is formulated as follows.

$$H(y\_, y) = -\sum y\_ \times \ln(y) \tag{10}$$

y_ denotes the predicted result, y denotes the actual result.

Layer 1 Dense: number of neurons: 80; activation function: relu; regularization function: L2 (REGULARIXER=0.04).

Layer 2: Pooling layer (Dropout: 0.5)

Layer 3: Dense: number of neurons: 40; activation function: relu

Layer 4: Pooling layer (Dropout: 0.5)

Layer 5: Dense: number of neurons: 2; activation function: sigmoid

The formula for the regularized L2 function is as follows.

$$loss' = loss + REGULARIZER * loss_{l2}(w) \tag{11}$$

$$loss_{l2}(w)=\sum|w_i|^2 \tag{12}$$

$$REGULARIZER = 0.04 \tag{13}$$

The activation functions relu and sigmoid are formulated as follows.

$$f(x) = \max(0, x) \tag{14}$$

$$f(x) = \frac{1}{1+e^{-x}} \tag{15}$$

4. Model Testing

4.1. Tensorboard Visualization

TensorBoard, as a built-in visualization tool in TensorFlow, can be used to show the network graph, the metric change of the tensor, the distribution of the tensor, etc. Especially when training the network, we can set different parameters (e.g., weight W, bias B, number of convolutional layers, number of fully connected layers, etc.), and it makes the understanding, debugging and optimization of TensorFlow programs more efficient and intuitive by visualizing the information in the log files output from the TensorFlow programs.

We visualize the training loss value (epoch_loss), test loss value (val_loss) and classification animation and run the results as follows.

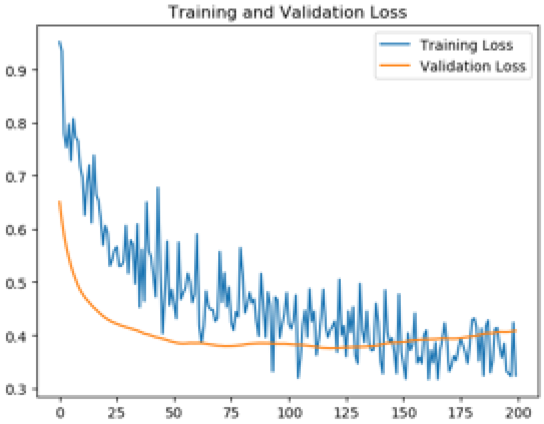

Figure 2. Visualization of loss values and classification (a)

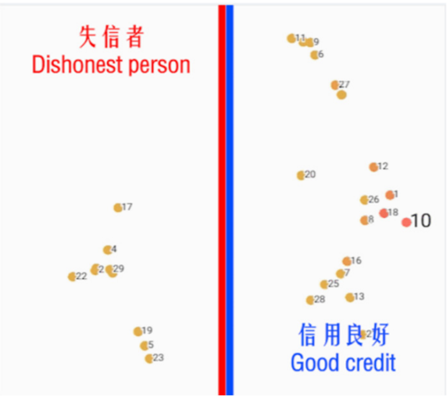

Figure 3. Visualization of loss values and classification (b)

The final credit evaluation system with a fitting accuracy of 88.26% and a prediction accuracy of 83.78% was obtained.

4.2. Comparison of Credit Assessment Models

4.2.1. Logistic Regression

Regression analysis is a statistical method for predicting one or more response variables from a set of independent variables, and can also be applied in the field of credit assessment to predict the effect of variables on response variables. Logistic regression is a classification model that first assumes that the data obeys a certain distribution and then uses a great likelihood estimate to do the estimation of the parameters. Logistic regression methods make few assumptions about the premises, and because of their relatively high stability, this makes them one of the most widely used assessment methods in the field of credit scoring.

Assuming that defaults may follow a logical distribution. y=0 means that the customer will not default and y=1 means that the customer will default, the principle of logistic regression is to use the existing data of the customer for modeling and analysis to predict the probability p of a customer default. For the kth customer, the information vector is: xk = (x1k, x2k, ⋯, xmk), which is modeled as follows.

$$\ln\left(\frac{p_k}{1-p_k}\right) = \beta_0 + \beta_1 x_1^k + \cdots + \beta_m x_m^k = \beta_0 + X_k\beta \tag{16}$$

$$\beta = (\beta 1, \cdots, \beta m)' \tag{17}$$

The value of the parameter β′can be calculated by the great likelihood estimation method.

4.2.2. Support Vector Machines

Data from micro and small enterprises often have a large number of missing, noisy points and singular values, and the data samples that can be used as valid information input are relatively limited. In contrast, accuracy and precision can be well improved by building a support vector machine model.

Support vector machines, a class of linear classifiers that perform binary classification of data by supervised learning, have a decision boundary that is the maximum margin hyperplane solved for the learned samples.

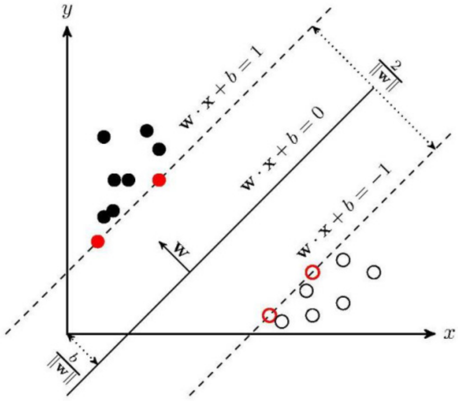

Figure 4. Support vector machine model

4.2.3. Comparison of Results

In order to establish a credit evaluation model with higher accuracy, the group members conducted preliminary experiments on the existing credit evaluation model and tested the fit of the model with the credit investment data of MSMEs. Some of the simulation results and the comparative and experimental results of the fit are shown in Table.

Table 4. Compare Results

	Prediction accuracy	Test set accuracy
MLP Neural Network	**88.26**	**83.78**
Logistic Regression	83.78	82.93
Support vector machine	86.49	83.55

5. Conclusion

In this paper, a PCA-MLAP multi-layer neural network credit model for micro and small enterprises with large missing values and small data volume is constructed by analyzing the Ger-Credit dataset and some micro and small enterprises' credit data, and the prediction accuracy reaches 83.78%. For a bank, the non-performing assets brought by one defaulted enterprise often need the deposit and lending spreads of dozens of enterprises to fill; for enterprises, a bank's perfect lending model can eliminate the "information asymmetry" between banks and enterprises to the greatest extent, thus reducing the loan interest rate. Therefore, for financial institutions, even a 1% increase in prediction accuracy can bring immeasurable profits. The nonlinear function of neural network is more closely related to the complex operation system of financial market, which is very feasible.

References

[1] Li, C.-E. Bank structure and SME financing [J]. Economic Research, 2002, 6:38-45.

[2] Bao Qin. Research and application of PCA-MLP credit scoring model based on data containing residual values[D]. Jinan University,2016.

[3] Saberi M, Mirtalaie M S, Hussain F K, et al. A granular computing-based approach to credit scoring modeling [J]. Neurocomputing, 2013, 122: 100-115.

[4] Zhong H, Miao C, Shen Z, et al. Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings [J]. Neurocomputing, 2014, 128: 285-295.

[5] Defry. Neural networks 1: multilayer perceptron - MLP.2019-02-10.

[6] Jaww. Machine Learning (ML) III of Multilayer Perceptron.2020-02-13.

[7] Zhou, Yaihua, Wang, Sunan. Research on credit risk of small and micro enterprises based on multilayer perceptron neural network[J]. Modern Management Science,2015,9:45-48.

[8] xholes. multilayer perceptron: Multi-Layer Perceptron.2017-11-07.

[9] Penguin. A multi-layer perceptron based neural network model application-analysis of old car auction transaction data.2018-07-20.

[10] Li Xue. Li M. G. Research on local government debt and its credit rating [M]. Beijing: Economic Science Press,2017:270-290.

[11] Chen, J. Y. Credit value assessment and credit rating evaluation system for small and medium-sized enterprises [D]. Shandong: Shandong University of Science and Technology, 2004.

[12] Qiao Wei. Construction of credit rating index system and model for small and medium-sized enterprises[J]. Journal of Kaifeng University, 2011(4):85-100.

[13] Zhang Peijun, Yang R. Research and Empirical Analysis on Credit Risk Evaluation of Micro, Small and Medium Enterprises. 2021-04-29.

[14] Yao Shangfeng. A personal credit evaluation model based on principal component analysis and BP neural network [J]. Mathematical Practice and Understanding,2007,37(21):21-24.

[15] Han Kao , Zhang Yaohui , Sun Fujun , Wang Shaohua . Method of determining index weights based on principal component analysis [J]. Sichuan Journal of Military Engineering,2012,33(10):124-130.

[16] Li Jinghua, Guo Yaohuang. Research on the method of principal component analysis for multi-indicator evaluation - principal component evaluation [J]. Journal of Management Engineering, 2002, 16(1):40-43.

[17] Favre. Detailed explanation of the principle of multi-layer perceptron & Python and R implementation. 2018-05-07.

[18] Wang Xue-Guang, Guo Yan-Bing, Qi Zhan-Qing. The effect of activation function on the performance of BP networks and its simulation study[J]. Control Theory and Applications,2002,21(4):15-19.