An Optimized BP Neural Network Credit Risk Prediction Model based on Cluster Analysis

Nianci Wang¹,* Wen Ma²

¹Shandong University of Science and Technology, Jinan, Shandong 250000, China; ²School of Electronic Information Engineering, Yangtze Normal University, Chongqing 408000, China.

*Corresponding Author

Abstract

Banks often determine the credit risk of such enterprises based on their creditworthiness, strength, and supply-demand relationship for small, medium-sized and micro enterprises that lack collateral assets, so as to give the most optimized credit strategy. Due to manual subjectivity and discrete data to reflect credit The ambiguity of risk, this article establishes a suitable mathematical model to quantify credit risk and determine credit strategy to solve this problem.

Keywords

Grey Correlation Analysis; 4-3-2-1 Cluster Scoring Model; Optimized BP Neural Network Model.

1. Introduction

Credit business is the main business of a bank, and the most important risk that a bank needs to face is credit risk. Credit risk has played an important role in the stability and healthy development of the national economy. Therefore, using the results of credit risk quantification and credit policies will enable the country’s economy to develop stably and sustainably. At present, my country’s bank credit risk evaluation system is not perfect. Therefore, how to carry out quantitative analysis and how to formulate credit policy is an important issue for credit risk research.

Small and medium-sized enterprises are smaller in scale, and correspondingly, their mortgage capacity is lower. Therefore, banks need to conduct credit risk analysis on the strength, reputation and influence of each enterprise, and determine whether to provide loans to enterprises and provide credit strategies such as the amount, interest rate and term according to credit policies.

The theoretical research on credit risk quantification is helpful for banks to rationally assess the credit risk of various small, medium and micro enterprises. Therefore, we need to establish a mathematical model for banks to conduct in-depth research on how to establish a scientific commercial bank credit risk assessment system.

Is given in order to affect the enterprise credit risk and the main factors of bank credit policy, we carried on the qualitative and quantitative research, first of all, using the EXCEL software in attachment a selection can be factors in the data visualization processing, and on the basis of credit analysis method of 5 c eight factors to gather the data for the grey correlation analysis to provide data support. The eight subjective factors were quantitatively analyzed by using grey correlation analysis and MATLAB software. Finally, with the correlation degree of 0.9 as the boundary, the two influencing factors of enterprise type and the ratio of purchasing quantity to revenue were deleted. Credit rating, the number of invalid and negative invoices, the total
amount of payment, the total profit of the enterprise in the regional time, the stability of the total profit of the enterprise, and the number of enterprises of the buyer are selected as six main influencing factors. The weight of each factor was obtained through the correlation degree of factors in the grey correlation analysis. K-means clustering analysis was used to cluster each factor prime data and the data was discretized in combination with the 4-3-2-1 scoring rule. Then the enterprise credit risk score (C) equation was established to calculate the score. The higher the risk of a bank lending to the enterprise. Then, four categories of high/high/low/low credit risk are obtained by secondary clustering of enterprise credit risk scores. The loan amount is determined according to the percentage of enterprise scores in all enterprise scores, and the annual loan interest rate is determined according to the relationship table between score clustering and enterprise annual interest rate.

2. Data preprocessing

2.1. Subtraction of irrelevant values
On the premise of guarantee the information correctly, as Banks on credit ratings as D in principle not to lend, so the evaluation grade D class enterprise for credit strategy does not affect, invalid invoice with negative invoice after statistical number has nothing to do with influencing factors, according to the principle of data reduction, in the end, we have to cut this kind of information.

2.2. Data normalization processing
In order to facilitate data capture and dimensionality elimination in modeling, min-max standardization is used to normalize data. The specific method is as follows: For each attribute, \( \min A \) and \( \max A \) are set as the minimum value and maximum value of attribute A respectively. 

\[
x = \frac{x - \min Q}{\max Q - \min Q}
\]

Thus, standardized index values are obtained.

3. Methods

In this question, credit information is used for quantitative analysis of credit risks and prediction of banks’ credit strategies. Since the 302 enterprises in Annex II lack credit rating compared with that in Annex I, BP neural network is used to predict the credit rating of 302 enterprises in Annex II. When there are many credit ratings of banks, subjective judgment is made manually. BP neural network takes 101 enterprise data in Annex I as training set and 22 enterprise data as test set. After continuous data feedback and update, the results of credit rating are finally trained. The 4-3-2-1 scoring mechanism based on K-means clustering in question 1 is adopted to analyze the data in Annex 2, and finally the quantitative results and credit strategies of 302 enterprises are obtained by introducing the enterprise credit risk scoring formula.

3.1. Credit rating prediction based on optimized BP neural network
Since the bank’s judgment of enterprise reputation is artificial judgment based on the financial data of enterprises, which is very subjective and accidental, we need to find a model to replace manual judgment of the credit grade of 302 enterprises by using existing data. For decision prediction, the most common models are BP neural network model, Probit decision model, cellular automata prediction model and so on. However, the latter two models more or less for the establishment of this model have a large gap. Probit model is a single-factor binary selection model based on continuous data, which is difficult to carry out multi-factor dynamic analysis. However, the rules of cellular automata state change are local in time and space, so it is difficult
to draw accurate conclusions when analyzing the whole. Although the BP neural network is essentially a static feedforward network, the accuracy of its feedback layer changes according to the data. This model has a large amount of data for machine training, and the BP neural network can serve as feedback to form a complete and reliable model.

3.2. Establishment of prediction model based on optimized BP neural network

BP neural network is composed of three layers: input layer, hidden layer and output layer. Among them, the input layer receives input information from the outside world and affects hidden layer neurons. The intermediate hidden layer neurons are responsible for processing information using a series of algorithms, rehearsing it, and finally delivering the information to the output layer. Output the final information to the outside world from the output layer.

Suppose the neuron network has N input neurons, M output neurons and P hidden layer neurons. The output of neurons is: $X_i = X_i^o = \sigma(\sum w_{ij}^h X_j + \omega_i^h), \ i = 1, 2, \ldots, p$

The output of the output layer neuron is: $y_i = \sum w_{ik} X_k + \omega_i^o, \ i = 1, 2, \ldots, n$

Since the BP neural network corrects the error by iterative approximation, if there is a deviation in the data, the continuous iteration of the calculator may lead to an exponential increase in the error value with the increase of the number of iterations. Therefore, the artificial correction layer is introduced to intervene in the computer’s prediction of credit rating.

The enterprise credit risk score formula:

$$C = 0.1696P_1 + 0.1627P_2 + 0.1701P_3 + 0.1663P_4 + 0.1698P_5 + 0.16148P_6$$

Where $P_i$ represents the input variable of the wavelet network model, that is, the data after clustering of the six influencing factors is taken as the input layer. Generation C is the forecast output of the wavelet network model, that is, the enterprise credit risk score.

When the input variable is $X_i$, the output expression of the hidden layer is as follows:

$$h(j) = h\left(\sum_{i=1}^{m} a_i X_i - b_i\right)$$

3.3. BP neural network training process

The training process of BP neural network is shown as follows:

Figure 1. BP neural network training diagram
4. Conclusion

Through the calculation model of bank’s loan quota and annual interest rate to enterprises in the question, combined with the formula.

\[
M = O \times \left( \frac{P}{\sum_{i=1}^{123} P_i} \right)
\]

Get the loan amount and annual interest rate of each enterprise and summarize them in supporting materials. The partial result list is shown as:

<table>
<thead>
<tr>
<th>Enterprise code</th>
<th>Loan amount</th>
<th>Annual interest rate</th>
<th>Loan time limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>E124</td>
<td>138002.0729</td>
<td>13.05%</td>
<td>1</td>
</tr>
<tr>
<td>E125</td>
<td>317741.9594</td>
<td>11.05%</td>
<td>1</td>
</tr>
<tr>
<td>E126</td>
<td>808351.8564</td>
<td>7.85%</td>
<td></td>
</tr>
<tr>
<td>E127</td>
<td>1000000</td>
<td>4.00%</td>
<td></td>
</tr>
<tr>
<td>E128</td>
<td>962386.4948</td>
<td>7.05%</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>E324</td>
<td>653046.2506</td>
<td>7.85%</td>
<td>1</td>
</tr>
<tr>
<td>E325</td>
<td>328751.5076</td>
<td>12.65%</td>
<td>1</td>
</tr>
<tr>
<td>E326</td>
<td>328751.5076</td>
<td>12.65%</td>
<td>1</td>
</tr>
<tr>
<td>E327</td>
<td>328751.5076</td>
<td>12.65%</td>
<td>1</td>
</tr>
<tr>
<td>E328</td>
<td>328751.5076</td>
<td>12.65%</td>
<td>1</td>
</tr>
</tbody>
</table>
References


