Financial Engineering and Risk Management (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2023.060602
ISSN 2523-2576 Vol. 6 Num. 6

# Research on Constructing a Credit Repository for College Students Based on Big Data and Modeling Credit Risks

**Shuran Yang*, Yinghao Zhang, Yining Sun, Lulu Cheng, Shuwei Sun**

*School of Sciences, Henan University of Technology, Zhengzhou, China*
**Corresponding author*

***Keywords:*** Credit risk model, BP neural network, Grey relational analysis

***Abstract:*** College students not only constitute a desirable target customer base for credit consumption, but also represent a potential group for credit consumption. However, due to the lack of in-depth understanding of credit consumption and relevant information, default events in college student credit consumption have occurred frequently in recent years, such as “campus loans”. Therefore, assessing the current credit consumption risk of college students is particularly important. In this study, we combine the algorithmic approach of machine learning with the research methodology of questionnaire surveys to build a college student credit repository based on big data. Firstly, by collecting various influential factors about credit consumption, an analysis of individual credit risk assessment indicators of college students is conducted, and the weights of each index are determined. Then, a BP neural network algorithm-based credit consumption risk evaluation model for college students is established using a training set and a test set.

## 1. Introduction

The credit consumption of college students[1, 2] has become a hot topic in recent years. Products and services tailored to credit consumption among college students have emerged, indicating that credit consumption among college students is always an attractive business. However, college students lack stable income and asset management skills, which makes the risk of credit default relatively high. Credit risk[3], also known as default risk, is an economic loss caused by the counterparty’s inability to fulfill the obligations specified in the contract, and it is an important financial risk.

To better address real-world problems, combination models have been gradually applied to credit assessment models, and have shown good performance in corresponding research fields. Nie[4] combined the single classifiers of logistic regression and decision tree in machine learning, and applied it to a credit scoring model, effectively reducing the misjudgment rate. Zhu[5] combined convolutional neural network and feature selection algorithm Relief, and applied them to a credit scoring model using a real dataset. Experimental results showed that the combination of the two methods can achieve better performance and higher prediction accuracy. Stjepan Oreski[6] proposed using a feature selection technique based on a hybrid genetic algorithm and using a neural network as a classifier to effectively reduce the default risk of the system. Chu Lei et al.[7] proposed a credit scoring model by combining support vector machine and BP neural network, and conducted experiments using a public dataset. The experimental results showed that the model has a relatively high overall accuracy. In the 21st century, the application of grey relational analysis in the big data field has gradually increased. In 2013, researchers from Huazhong University of Science and Technology, including Yang Guang, proposed a big data-based grey relational analysis method which was detailed in the paper *Grey Relational Analysis Model Based on Big Data*[8].

Building on the studies mentioned above, this paper intends to use the combination of grey relational analysis and BP neural network to construct a credit risk assessment model.

## 2. Data Collection and Processing

### 2.1. Data Collection

After comprehensively considering the various factors that influence credit risk, an online survey questionnaire was designed to collect information from college students. A total of 500 responses were obtained, with 481 valid data points. Twenty sets of data were selected as the training set, while the remaining data were used as the test set.

### 2.2. Data Processing

Prior to the official training of the BP neural network, the collected data must be normalized according to the requirements of the model. In this study, the selected normalization method was the maximum-minimum method. Specifically, any value in a given column was divided by the range of values within that column, which was calculated as the difference between the maximum and minimum value in that column. Consequently, all the factors influencing the credit risk were scaled to a range of [0, 1].

## 3. Research Methods and Ideas

### 3.1. Grey Correlation Analysis Results

Firstly, indicators of influencing factors were quantified, as shown in Table 1.

Additionally, assuming there is a binary variable, y, which indicates the credit risk level "high" or "low", where $y=0$ indicates a "high" credit level, and $y=1$ indicates a "low" credit level. Banks and other credit agencies can use the credit level to determine whether to grant a loan or not.

Next, standardization was performed on the collected data. Firstly, the comparative sequence and reference sequence were selected. Assuming that $X_i$ is the system factor and $k$ represents $X_i(k)=1,2,\cdots,n$ . Let $X_0=Y=\{X_i(k)=1,2,\cdots,n\}$ be the reference sequence, $X_i=\{X_i(k),k=1,2,\cdots,n,i=1,2,\cdots,m\}$ the comparative sequence.

Secondly, standardization was performed. In grey relational analysis[9], it is necessary to nondimensionalize data for different sequences with dimensional differences. Generally speaking, this process can be divided into initial value processing, average processing, interval processing and other methods. In this study, normalization was conducted using the regularization method with the following formula 1:

$$X'=\frac{x-\min(x)}{\max(x)-\min(x)}, \tag{1}$$

Table 1: Quantification of indicators

| Indicator Quantification Standard of Influencing Factors of College Students' Credit Risk | | |
|---|---|---|
| Types | Name | Variable quantification |
| Financial factors | Monthly income | Below 1500=1 |
| | | 1501-2500=2 |
| | | 2501-3500=3 |
| | | Above 3500 = 4 |
| | Consumption elasticity | Numeric data |
| | Consumption difference at the beginning and end of the month | 30% or above = 1 |
| | | 30% or below = 2 |
| | | 20% or below = 3 |
| | | 10% or below = 4 |
| | | No difference =5 |
| Consumption level | Methods of credit consumption | One method=1 |
| | | Two methods=2 |
| | | Three methods=3 |
| | Maximum credit consumption limit | Living expenses within half a month = 1 |
| | | Living expenses within a month = 2 |
| | | Living expenses within two months = 3 |
| | | Living expenses within three months = 4 |
| | Time of payment | Within 1 month = 1 |
| | | Within 3 months = 2 |
| | | Within 6 months = 3 |
| | | Within 1 year = 4 |
| | | Other = 5 |
| Other | Attitudes towards credit consumption | I believe it is normal for college students to have credit expenditure: From strongly disagree to strongly agree: 1-5 points |
| | | I will use credit spending when needed: From strongly disagree to strongly agree: 1-5 points |
| | Credit risk awareness factors | I will repay the loan before the due date: From strongly disagree to strongly agree: 1-5 points |
| | | I am aware of the negative impact of loan delinquency: From strongly disagree to strongly agree: 1-5 points |
| | Improvement consciousness factors | I will become thrifty when faced with repayment pressure: From strongly disagree to strongly agree: 1-5 points |
| | | I will not seek help even if the loan exceeds my ability to repay: From strongly disagree to strongly agree: 1-5 points |

The results are shown in Table 2:

Table 2: Data standardization processing

| S/N | 1 | 2 | 3 | ... | 480 | 481 |
|---|---|---|---|---|---|---|
| Level of risk (Y) | 0.946556 | 0.118927 | 0.9856 | ... | 0.9562 | 0.4329 |
| Gender X1 | 0.60183 | 0.589105 | 0.5687 | ... | 0.8745 | 0.6945 |
| Age X2 | 0.102067 | 0.207681 | 0.3459 | ... | 0.2956 | 0.7952 |
| Educational background X3 | 0.917566 | 0.383695 | 0.9052 | ... | 0.8546 | 0.4529 |
| Consumption source X4 | 0.420073 | 0.240212 | 0.8412 | ... | 0.7745 | 0.8631 |
| Steady monthly living expenses X5 | 0.647051 | 0.270906 | 0.5978 | ... | 0.9542 | 0.8472 |
| Consumption elasticity X6 | 0.558383 | 0.320896 | 0.8074 | ... | 0.8753 | 0.2986 |
| Consumption fluctuation range X7 | 0.237621 | 0.799351 | 0.8945 | ... | 0.9543 | 0.3749 |
| The amount spent more at the beginning of the month than at the end of the month X8 | 0.096019 | 0.487318 | 0.7236 | ... | 0.7821 | 0.8459 |
| Methods of credit consumption X9 | 0.634326 | 0.420653 | 0.3425 | ... | 0.9362 | 0.5968 |
| Amount of credit consumption X10 | 0.010786 | 0.697055 | 0.5946 | ... | 0.7246 | 0.9874 |
| Amount of credit X11 | 0.107756 | 0.462582 | 0.2987 | ... | 0. 2397 | 0.8521 |
| Loan repayment time X12 | 0.357093 | 0.523187 | 0.6425 | ... | 0.7129 | 0.7638 |
| Attitude towards credit consumption risk X13 | 0.784861 | 0.921274 | 0.6675 | ... | 0.9125 | 0.8976 |
| Awareness of credit consumption risk X14 | 0.409286 | 0.683766 | 0.4587 | ... | 0.7584 | 0.8169 |

Finally, the correlation degree was calculated.

Let the correlation coefficient between the reference sequence $Y'$ and the comparative sequence $X'$ at point *k* be represented by the following formula 2:

$$\varepsilon_i(k) = \frac{\min_i \min_k |Y' - X_i| + \max_i \max_k |Y' - X_i|}{|Y' - X_i| + \varepsilon \max_i \max_k |Y' - X_i|} \tag{2}$$

According to the calculation formula 3, the correlation coefficients between the 15 influencing factors and the credit risk level are arranged in ascending order and shown in Table 3 as follows:

Table 3: Correlation degree ranking

| Variables | The degree of correlation between Y and each influencing factor |
|---|---|
| Consumption elasticity X6 | 0.88 |
| Amount of credit X11 | 0.82 |
| Educational background X3 | 0.76 |
| Attitude towards credit consumption risk X13 | 0.72 |
| Remedy consciousness X15 | 0.69 |
| Steady monthly living expenses X5 | 0.65 |
| Consumption fluctuation range X7 | 0.58 |
| The amount spent more at the beginning of the month than at the end of the month X8 | 0.56 |
| Awareness of credit consumption risk X14 | 0.55 |
| Amount of credit consumption X10 | 0.52 |
| Age X2 | 0.5 |
| Loan repayment time X12 | 0.49 |
| Gender X1 | 0.49 |
| Methods of credit consumption X9 | 0.46 |
| Source of consumption X4 | 0.42 |

The research findings reveal a certain degree of correlation between credit risk and several influencing factors, while the correlation between individual factors remains relatively low. In this study, a threshold of 0.5 was chosen for the degree of correlation, such that when the degree of correlation exceeds 0.5, it is considered a significant influencing factor, while when the degree of correlation is below 0.5, it is regarded as a weak influencing factor. Therefore, 11 indicators were ultimately selected, which include: consumer elasticity, amount of credit, educational background, attitude towards credit consumption risk, remedy consciousness, stable monthly living expenses, consumption fluctuation range, the amount spent more at the beginning of the month than at the end of the month, awareness of credit consumption risk, credit consumption amount, and age.

## 3.2. BP neural network model based on grey correlation analysis

### 3.2.1. BP Neural Network Model Training

The Back Propagation (BP) algorithm, also known as the Error Back Propagation Algorithm, was proposed by Rumelhart and MoCelland in 1986[10]. Since the training of multi-layer feedforward networks often adopts the backpropagation algorithm, people commonly refer to multi-layer feedforward networks as BP networks. The topological structure of a BPNN mainly comprises three layers: the input layer, several hidden layers, and the output layer. The topology of a BPNN is shown in Figure. 1.

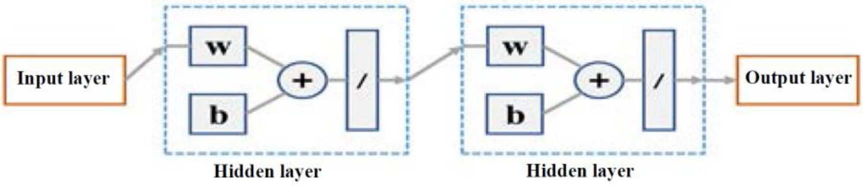


Figure 1: BPNN topology

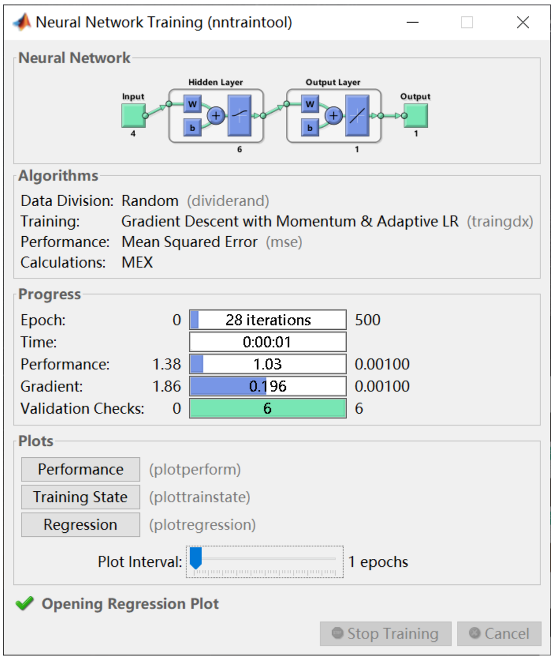


Figure 2: Structure of BP neural network

After completing the initial training of the BP neural network model, test samples were introduced to the network. It was observed that after 26 iterations, the network output error was already within the expected range, and the training time was relatively short. The mean training square error was also comparatively low, with the average prediction error being below 0.5. These results suggest that all aspects of the neural network model were constructed well, and the network architecture and learning process are shown in Figure. 2.

The prediction error of BP network is shown in Figure. 3:

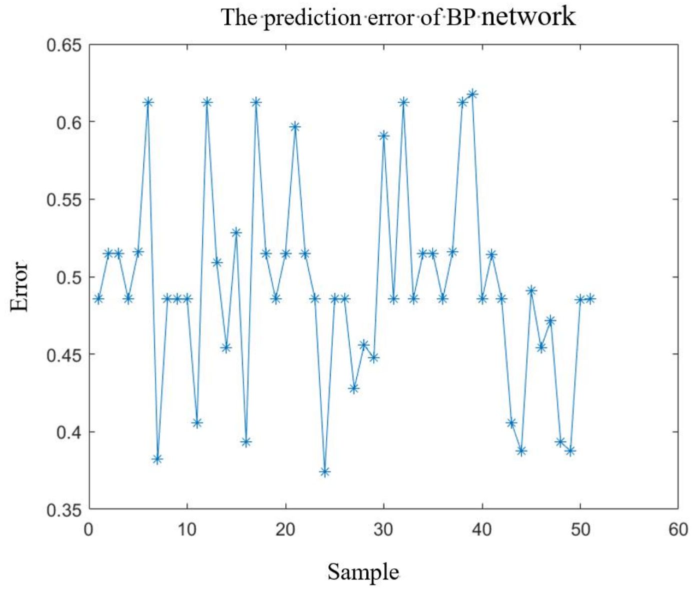


Figure 3: BP prediction error

A fitting analysis of the training samples, test set, and all data was conducted, and the results are presented in Fig. 4. The fitting goodness of the training samples was 99.855%, that of the test set was 99.788%, and that of all data reached 99.842%. These fitting goodness results demonstrate that the network training effect was effective and highly accurate. From a practical perspective, this model can explain approximately 99% of the data samples. After excluding the effects of individual data samples, it can be concluded that this model can achieve accurate judgments of credit risk levels.

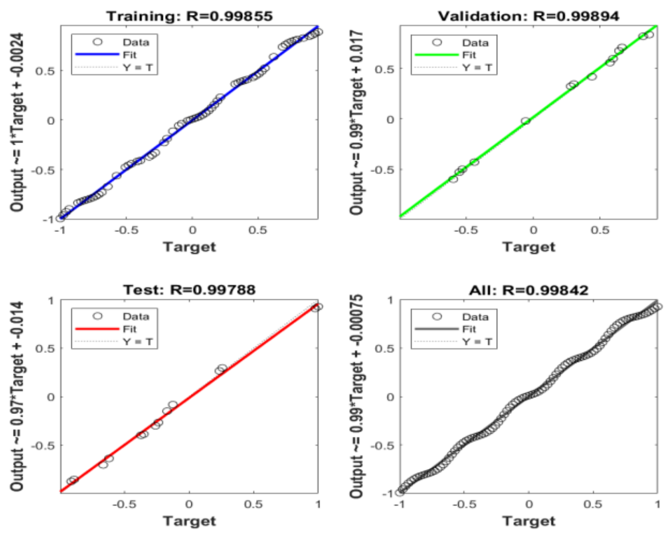


Figure 4: Goodness of fit

### 3.2.2. Testing of BP Neural Network Model

Based on the aforementioned model, predictions were made for the next 21 sets of data in the sequence. A comparison was then made between these 21 predicted data sets and their corresponding true values to assess the real-world performance of the BP neural network model. The results showed that out of the 21 data sets, 18 were correctly predicted. This indicates that the prediction accuracy reached 85.71%, which is in line with the expected level of precision for the model.

## 4. Research Conclusions

This project aims to focus on the credit consumption of college students and use grey correlation analysis to identify significant factors. Then, a BP neural network model was established for credit risk assessment through the input of the test set and training set. Ultimately, an effective credit risk assessment model was developed.

As time marches on and society continues to evolve, an increasing number of credit products are finding their way into the lives of college students. Consequently, major financial institutions can use this model to gain insight into the credit status of loan applicants and make informed decisions regarding loan approval. The establishment of this model holds great practical significance for elevating the integrity and conscientiousness of college students, fostering a culture of integrity in higher education, and advancing the development of a broader societal framework built on trust and sincerity.

## Acknowledgement

This work is supported by the Key project of Innovation and Entrepreneurship Training Program for College Students in Henan Province in 2022 (202210463068).

## References

*[1] Ding Fengjiao, Zhang Qi, You Dongmei, et al. The development prospect of online staging shopping platform in university campus-based on the field research of Nanchang college students' consumer groups. China Collective Economy, 2016, 13): 64~65.*

*[2] Xiaowen Zhu, Wei Ren, Qiang Chen, Richard Evans. How does internet usage affect the credit consumption among Chinese college students? A mediation model of social comparison and materialism. Internet Research, 2021, 31 (3).*

*[3] Liu Hao. Research on the current situation and influencing factors of college students' credit consumption risk. Zhongnan University of Economics and Law, 2019.*

*[4] Nie G. Credit card churn forecasting by logistic regression and decisiontree. Expert Systems with Applications, 2011, 38 (12): 15273-15285.*

*[5] Zhu B, Yang W, Wang H, et al. Ahybrid deep learning model for consumer credit scoring, 2018 international conference on artificial intelligence and big data (ICAIBD). IEEE, 2018: 205-208.*

*[6] Oreski S, Oreski G. Genetical gorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 2014, 41 (4): 2052-2064.*

*[7] Chu Lei. Comparative Study on Personal Credit Evaluation Based on BP Neural Network and SVM. Shanghai Normal University, 2014.*

*[8] Meng Haodong. Research on Fault Diagnosis of Transmission Box Based on Neural Network and Grey Theory. North University of China, 2005.*

*[9] Xu Xiao. Functional consumption measurement and model construction of ARM Android application based on DVFS technology. Southeast University, 2016.*

*[10] William B. Claster. Athematics and R Programming for Machine Learning. RC Press, 2020.*