Bank Failure Analysis Based on BP Neural Network and C-Means Clustering

Yingjian Liang[*,#], Jipeng Guo[#]

School of Applied Mathematics, Beijing Normal University, Zhuhai, Guangdong, 519000, China
*Corresponding author
#These authors contributed equally

Abstract: *Banks are an important financial support for national economic development. It can adjust the market economy, adjust the industrial structure, raise funds for national construction, and provide loans to enterprises and individuals who are short of funds. With the downward pressure of economic pressure, some banks' liabilities are higher than assets. With the accumulation of time, liabilities cannot be repaid, and finally they can only end up in bankruptcy. Compared with domestic banks, the phenomenon of international bank bankruptcy is more serious. In order to reduce the occurrence of such events, we need to analyze the causes and current situation of bank bankruptcy.*

Keywords: *Entropy right method; Gray correlation; Blur comprehensive evaluation; Neural network model; Cluster analysis*

1. Introduction

First of all, the appropriate input-output data are sorted out from the 64 index data of each bank. According to the sorted data, the efficiency of each bank is evaluated and the dividing line of bank failure efficiency is drawn. Then, analyze the causes of bank failure, find out the most important five indicators, and calculate the weight of these five indicators[1]. Compare and analyze the results of bank failure, and establish an accurate failure risk prediction model. Then select the most representative 20 existing banks and 20 failed banks from the bank data in 2021, and use the data of these banks to predict and analyze the failure risk of other banks. Finally, by referring to relevant literature, we can theoretically analyze the data of banks from 2017 to 2021, judge which data may come from the same bank, and predict which banks will fail in combination with the time series data of the same bank [2].

By preprocessing the data, we find that there are many incomplete values in the table. If we delete the incomplete values directly, it will have a great impact on the integrity of the data, so we need to deal with the incomplete values. To deal with the incomplete value, we need to use SPSS software to carry out correlation analysis, eliminate abnormal indicators, put the data of a certain year into SPSS data processor, take an indicator as a unit, and compare between each two units. According to the correlation law: the larger the correlation coefficient is, the stronger their relationship is. Therefore, we choose the number with larger correlation coefficient between the two indicators as the prediction value, and the indicator data is fitted in pairs, So as to preliminarily determine the indicators [3]. Then, through reading the literature, we get the traditional indicators commonly used by banks [4]. Considering the preciseness of the results, we need to make a quantitative analysis of the above indicators. Grey correlation analysis is used to determine which indicators have an impact on bank failure. The relevant data of them are calculated and plotted by MATLAB, and the tightness of the relationship is determined according to the similarity of the graphics, so as to determine the specific indicators [5]. Finally, establish a fuzzy comprehensive evaluation model, calculate the entropy value of each index by using the entropy weight method, determine its corresponding weight, and calculate the efficiency index of each bank, so as to get the dividing line of bank failure efficiency [6].

We need to select 64 index data, select the five most important index data, and calculate their weights. First, we need to establish a principal component analysis model to reduce the dimension of the data, calculate the correlation coefficient matrix, eigenvalue and eigenfunction through the algorithm, and synthesize the above data to get the first five indicators that affect bank failure. Then, calculate the comprehensive evaluation value of the five indicators (including the information contribution rate and cumulative contribution rate of the five indicators) [7]. Finally, calculate the comprehensive score to get its corresponding weight.

The efficiency of each bank is obtained. The 64 index data have an impact on whether the bank fails, and the dividing line of bank failure is determined. On this basis, the interference data is excluded, and the more important five index data are presented concretely. In order to obtain the specific value, we obtain a range of the five comprehensive evaluation values, and find that the lower the comprehensive evaluation value is, the closer it is to collapse; The higher the comprehensive evaluation value, the farther away from bankruptcy. Then, a new data set is constructed by establishing BP neural network model, in which 64 index data are used as the input set of the data set, and the comprehensive evaluation scores of each bank are used as the output set. Finally, according to the data obtained by the algorithm, the bank composite index is predicted, and the relative error between the actual value and the predicted value is calculated to obtain a more accurate failure risk prediction model.

We establish a c-means clustering model for bank data. Then we need to cluster the bank data, classify the data with similar conditions into one class, and determine the class centers of existing banks and failed banks respectively. (basis: the class center represents the characteristic value of a class of data, which is the closest value of similarity in a class of data, so determining the class center is to determine the most representative bank.) Finally, the neural network model is used to predict the risk of bank failure [8].

First, we need to observe the data from 2017 to 2021, find out the data of important indicators, and compare the data indicators of the previous two years. Then, excel is used to analyze the matching degree of these data, so as to find the banks with high similarity, which can be regarded as the same bank. Finally, observe the score of each bank in different years, analyze the change trend of their scores, and then find out which banks have the trend of bankruptcy.

2. Model assumptions and symbols

2.1 Model assumptions

Hypothesis 1: the changes of indicators of each bank will not affect the changes of indicators of other banks.

Hypothesis 2: bank efficiency and bankruptcy have no systematic impact.

2.2 Symbols

The following Table 1 shows the description of symbols.

Table 1: Symbol Description

Symbols	Symbol Description
PCA	Principal component analysis
FHCE	Fuzzy hierarchical comprehensive evaluation
NNM	Neural network model
FCM	C-means clustering algorithm

3. The establishment and solution of the model

3.1 The dividing line of bank failure efficiency

SPSS software is used to analyze the correlation of data. Take 2017 as an example, as shown in Table 2.

At the 0.05 level (two tailed), the correlation is significant.

At the 0.01 level (two tailed), the correlation is significant.

When solving the predicted value, it is necessary to select two series, namely, the series with missing indicators and the series with large correlation coefficient. The data of the two series are fitted in pairs according to the following: the indicators with larger correlation coefficient have closer relationship. Take the data of X1 and X3 in 2017 as an example, select 30 banks, draw a scatter diagram, and get the fitting curve [9]. As shown in figure 1.

Table 2: Partial data of correlation analysis in 2017

|		VAR00010	VAR00011	VAR00012	VAR00013
VAR00010	Pearson correlation	1	0.387	-0.233	-0.411
	Significance (double tailed)		0.391	0.615	0.360
	Number of cases	7	7	7	7
VAR00011	Pearson correlation	0.387	1	-0.812	-0.930
	Significance (double tailed)	0.391		0.026	0.002
	Number of cases	7	7	7	7
VAR00012	Pearson correlation	-0.233	-0.812	1	0.875
	Significance (double tailed)	0.615	0.026		0.010
	Number of cases	7	7	7	7
VAR00013	Pearson correlation	-0.411	-0.930	0.875	1
	Significance (double tailed)	0.360	0.002	0.010	
	Number of cases	7	7	7	7
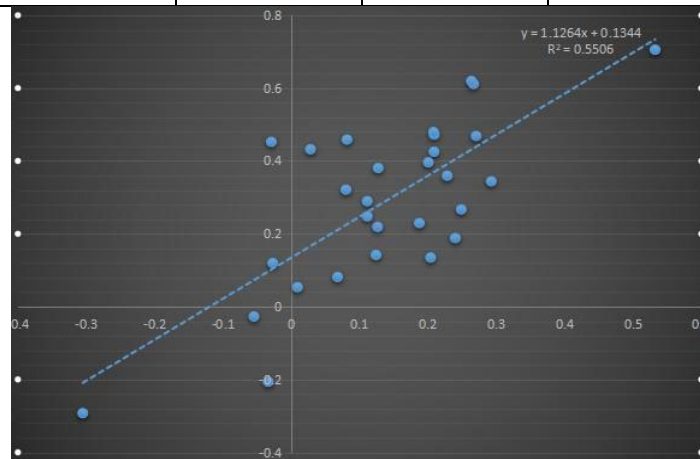
Figure 1: Data fitting

Through reading the literature, we can get the commonly used input-output evaluation indicators of banks: fixed capital / total assets (x38), operating expenses / short-term liabilities (x33), net profit / total assets (x1), total liabilities / total assets (x2), total assets / total liabilities (x17), profit from operating activities / total assets (X22), net profit / sales (x23), sales / short-term liabilities (X63), sales / fixed assets (x64), etc. Considering the preciseness of the results, we use the grey correlation analysis method for quantitative analysis. Grey correlation analysis refers to the method of quantitative analysis and comparison of a dynamic process. Its method is to determine the comparison parameters, judge whether the geometric similarity between the parameter series and the comparison series is close, and reflect the correlation degree of data. This method is usually used to analyze the influence of various factors on the results [10].

Specific steps:

Step 1: determine the analysis sequence, which is divided into reference sequence and comparison sequence. Let the reference sequence be $Y=\{Y(k)|k=1\ldots n\}$, and the comparison sequence be $X=\{Xi(k)|k=1\ldots n, i=1\ldots m\}$.

Step 2: dimensionless processing of data, formula:

$$x_i(k) = \frac{X_i(k)}{X_{i(l)}} \qquad (1)$$

Step 3: calculate the correlation coefficient between data.

Step 4: calculate the correlation degree. Because the correlation coefficient obtains the correlation degree between series, it cannot obtain an accurate value. Therefore, in order to obtain an accurate value and exclude the influence of too much information on the result, we need to concentrate the correlation

coefficient at each time into a specific value and calculate its average value. Formula:

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \varepsilon_i(k)(\text{k} = 1\dots\text{n}) \tag{2}$$

Step 5: sort the obtained correlation degree from large to small.

After the grey correlation analysis, we get the input-output indicators that have an impact on the efficiency of banks.

Based on the above indicators, we establish a fuzzy comprehensive evaluation model, use the entropy weight method to calculate the entropy value of each indicator, determine its corresponding weight, and calculate the efficiency index of each bank, so as to get the dividing line of bank failure efficiency. Principle: entropy can be used to judge the degree of dispersion between indicators. The smaller the entropy, the greater the degree of dispersion of indicators, and the greater the impact on comprehensive evaluation.

Specific steps:

Step 1: construct the judgment matrix of 64 indicators from 2017 to 2021.

Step 2: normalization processing judgment matrix.

Step 3: get the target entropy.

Step 4: give weight to entropy.

Step 5: calculate weight.

3.2 Calculate the comprehensive evaluation value

We use the method of establishing principal component analysis model to calculate. Principle of principal component analysis: among many variables, mining related variables, and dividing variables into related and unrelated categories. The correlation shows that there is information overlap. The role of principal component analysis is to delete redundant variables, reduce irrelevant variables, get new variables, and reduce the dimension of data[11].

Specific steps:

Step 1: standardize the index data with SPSS, taking 2017 as an example (only some data are intercepted here) in Table 3.

Table 3: Descriptive statistics

	N	Minimum value	Maximum value	Average value	Standard deviation
X42	7027	-701.630000	2156.800000	0.26363779	27.071123329
X64	6993	0.000010	294770.00000	208.73195011	5140.7088038
X63	6997	0.000015	1042.200000	8.12685230	19.996418769
X62	7027	0.00000	25016000.000	4763.2019547	310783.48359
X58	7027	-0.004194	1108300.0000	158.74086224	13221.235589
X61	7005	0.000016	21110.000000	15.64222762	261.55453366
X60	6892	0.000047	2137800.0000	432.88295800	26128.019222
X59	7026	-327.970000	119.580000	0.27782866	6.339149050
X57	7026	-315.370000	126.670000	0.193224296	4.344045841
X56	7027	-1108300.000	1	-157.73673851	13221.247567
X54	6993	-122.030000	180440.00000	99.63600216	2799.5866192
X55	7027	-800470.0000	4398400.0000	8855.6932941	72475.270699
X53	6993	-130.470000	180440.00000	98.72057901	2785.3674071

Step 2: judge the correlation of index data.

Step 3: determine the number of principal components m, m=5, and the observed value is m, m=64.

Step 4: calculate its eigenvalue and eigenvector to get the expression of Fi after principal component analysis.

Step 5: five principal components are obtained.

3.3 Prediction model of bankruptcy risk

Through references, we believe that neural network model can predict the risk of bank failure.

Working principle of neural network model: neural network is derived from neurons, and neural network model is derived from the mathematicization of neurons. It gathers a large number of similar indicators, layers them from the indicators, refines the data, inputs eigenvalues, and finally outputs the target value through correlation calculation. Among them, the input index is the index data of 64 banks, and the output index is the comprehensive score of each bank. Let the comprehensive score of each bank be y, and the comprehensive score obtained after neural network model processing is X. in order to make the data more convincing, we need to calculate the relative error[12]. Formula for calculating relative error R:

$$R = \frac{|y-x|}{y} * 100\% \tag{3}$$

Taking 2017 as an example to calculate the relative error of some data values, as shown in Table 4.

Table 4: Relative error analysis

	Comprehensive score of each bank (y)	Processed value (x)	Error (R)
1	159.52	161.54	-1.27%
2	137.87	139.39	-1.10%
3	24.67	26.66	-8.07%
4	89.765	92.43	-2.97%
5	290.03	292.22	-0.76%
6	200.64	203.23	-1.29%
7	51.17	53.31	-4.18%
8	257.07	254.66	0.94%
9	236.55	234.25	0.97%
10	115.13	114.32	0.70%
11	132.71	135.43	-2.05%
12	121.70	120.66	0.85%
13	101.75	102.39	-0.63%
14	98.33	95.26	3.12%
15	205.42	207.11	-0.82%

Observing the table data, it is found that the error value is not much different, so the neural network model can be used to predict the bank failure risk.

3.4 Establishing C-means clustering model

C-means clustering algorithm clusters the data by optimizing the objective function, so as to obtain the class center with prominent features [13].

Specific steps:

Step 1: according to this question, we divide the data of 2021 into two categories: existing banks and failed banks, and establish a c-means clustering model.

First, determine the initial class center. Then, the Euclidean distance formula is used to calculate the distance from each index data to the class center [14].

$$D_e(x, y) = \sqrt{\sum_{i=1}^{n} |X_i - Y_i|^2} \tag{4}$$

Finally, calculate the average value of the distance from each index data to the class center in order to obtain a new class center [15].

Step 2: calculate the class centers of existing banks and failed banks respectively through MATLAB. There are 20 class centers.

Step 3: get 20 representative existing banks and 20 representative failed banks.

Step 4: put the obtained data into the neural network model established in question 3 to predict the comprehensive value of bank failure risk.

4. Discussion and reflection

This paper analyzes the causes and current situation of bank bankruptcy, and uses fuzzy comprehensive evaluation model, principal component analysis model, BP neural network model, c-means clustering model and clustering analysis method to infer the banks that may have the trend of bankruptcy.

References

[1] Liu Ting, Berlin. Based on the logical decision map — the deployment spare parts of the main component analysis method determine [J]. Aviation engineering progress, 2012.

[2] Gao Jie; Tian Yunfei; Xu Ying; et al. The application of the main component analysis method in the UUV cluster's anti -water search [J]. Digital ocean and underwater offense and defense, 2022,5 (01).

[3] Xi Xiaoqian; Li Hong; Wang Ruijun; et al. The aspirational evaluation of the corn variety and the gray correlation analysis of the agronomic traits [J]. Agricultural Journal, 2022,12 (04).

[4] He Na; Wang Baoqing; Yan Zhuangzhuang. Comprehensive evaluation and analysis of badminton skills and tactics based on entropy value method and gray association analysis method [J]. Sports science and technology literature notification, 2022, 30 (04).

[5] Gao Han; Fang Panpan; Dou Jingjing; et al. Empirical research based on AHP+fuzzy comprehensive analysis method [J]. Today's wealth (China Intellectual Property), 2020, (07).

[6] Chen Wenjin; Zhu Feng; Zhang Tongyan; et al. Green supply chain evaluation based on fuzzy comprehensive analysis method [J]. Hebei Enterprise, 2019, (12).

[7] Nie Yuanxiu. Photovoltaic power prediction method based on AFSA-BP neural network [J]. Zhejiang Power, 2022,41 (04).

[8] Wu Jiajun; Liu Baomin; Liu Nian. Forecasting of foundation pit based on BP neural network [J]. Journal of Shaoyang College (Natural Science Edition), 2022,19 (01).

[9] Wang Qiuping. The connection between the vague comprehensive judgment and AHP [J]. Journal of Xi'an University of Technology, 1997, (04).

[10] Ye Jianqiang; Blue Tao Chrysanthemum; Huang Zhuozhong; et al. The agronomy traits of different sources and its gray correlation analysis [J]. 2010, (08).

[11] Jin Yi. Looking back at the risk management of banks in Balin Bank's closing dialysis banking industry [J]. Audit observation, 2020, (10).

[12] Ji Fuquan; Sun Qingwen; Sun Zihan. Thinking and revelation of the bankruptcy of the contractor bank [J]. Instrument and meter user, the era of wealth, 2021, (12).

[13] Xing Tianyi; Yin Junwei. U.S. Commercial Bank's bankruptcy disposal experience and revelation [J]. Northern Finance, 2020, (09).

[14] Wang Ping; Gao Liying; Luo Zhenghua; et al. Research on the energy efficiency system based on the related coefficient-entropy power method [J]. 2021, (05).

[15] Li Jiajun; Sun Yajun; Xu Zhimin; et al. Based on fuzzy comprehensive analysis method evaluation and exploration drilling risk [J]. Mining safety and environmental protection, 2015, 42 (06).