ISSN: 2688-8653

Enterprise Credit Risk Quantification and Clustering

Chen Huang

School of Department of Electrical Engineering, North China Electric Power University, Baoding 130600, China

3497483644@qq.com

Abstract

In this paper, the credit risk of enterprises is analyzed according to the C question of national college students' mathematical modeling. Firstly, several factors related to enterprise credit risk are constructed from the attachment, and then principal component analysis is used to extract the principal components. On this basis, the credit risk of each enterprise is quantified, and the risk level value F is calculated. Then, according to the obtained F value, the Q-type cluster analysis is carried out. According to the clustering results, the 123 enterprises in the attachment can be divided into four categories, for which the bank provides different loan policies.

Keywords

Principal Component Analysis; Clustering Analysis; Enterprise Credit Risk.

1. Introduction

Credit assets risk management monitoring is based on a set of benefit indicators to evaluate its quality and determine the quality of credit assets in the management process. As an important place for capital turnover, the bank is the pillar of the development of many small and medium-sized enterprises. In practice, due to the relatively small scale of SMEs and the lack of collateral assets, it is difficult for enterprises to finance, so Banks are usually based on credit policy, enterprise trade instrument information and the influence of the upstream and downstream enterprises, the strength, stable supply and demand of the enterprise to provide loans. The bank evaluates the credit risk of small, medium and micro enterprises according to their strength and reputation.

2. Quantification of Value of Credit Risk

2.1. Analysis of Factors

According to the data provided in the attachment, we select 7 variables to judge the credit risk value of an enterprise. These seven variables can be roughly divided into three categories: the strength of the enterprise, the stability of the supply and demand relationship and the level of credibility.

(1) Strength: business condition is evaluated by the total value of price and tax in the output invoice information X1; Profitability is calculated by profit X2, where X2 is the difference between output price tax and input price tax. Tax rate X3 is the quotient of output tax and total output amount.

(2) In terms of supply and demand stability: firstly, I calculated the number of valid invoices of each company every six months according to the data in the attachment. The number of invoices of input and output can represent the supply value and demand value of the company respectively. The standard deviation of the ratio between the number of output invoices recorded by different companies and input invoices recorded every half year is taken as the

ISSN: 2688-8653

parameter of the stability of the supply and demand relationship of the enterprise, which denoted as X4.

(3) Reputation: credibility X5, which is related to the input invoices of various enterprises. First, I preprocess the data to remove the invalid invoices and negative invoices and screen out the valid invoices. In other words, valid and positive invoices for each company are counted, and their proportion in the total number of votes is set as the value of the company's credit rating X5. Rating score X6, according to the credit rating companies must determine, in this to quantify it, I will be between 0 and 1 grade evaluation, because of the problem set that "Banks on credit ratings as D in principle not to lend", so we set A grade for 1 minute, B is 0.5 points, C as 0.25, D is 0 points. Default status X7 is evaluated in line with whether the enterprise has breached the contract or not. Default status is 0 and non-default status is 1.

2.2. Model Building

Considering the different contributions of each variable to the evaluation system and the large number of the 7 variables. In order to simplify and deal with the correlation degree, I use the principal component analysis method to analyze them and get the comprehensive evaluation value.

The steps are as follows:

Step1: Standardize the raw data. Assume that there are m index variables for principal component analysis, which are $x_1, x_2 \ldots, x_m$, there are n evaluation objects, and the value of the j index of the i evaluation object is a_{ij}. Convert each index value a_{ij} into the standardized index value $\tilde{a}_{ij}$, then can get

$$\tilde{a}_{ij} = \frac{a_{ij} - \mu_j}{s_j}, (i = 1,2,..n; j = 1,2,\ldots,m)$$

Where $\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$, $s_{ij} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \mu_j)^2}$, $j = 1,2,\ldots\ldots,m.$ That μ_j, s_j for the first j index of the sample mean and the standard sample. Correspondingly, denote:

$$\tilde{x}_j = \frac{x_j - \mu_j}{s_j}, j = 1,2,\ldots,m$$

is the standardized indicator variable.

Step2: Calculate the correlation coefficient matrix R, the correlation coefficient matrix R= $(r_{ij})_{m\times n}$.

$$r_{ij} = \frac{\sum_{k=1}^{n} \tilde{a}_{ki} \cdot \tilde{a}_{kj}}{n-1}, i = 1,2,\ldots,n, j = 1,2,\ldots,m$$

Where $r_{ii} = 1, r_{ij} = r_{ji}$. r_{ij} is the correlation coefficient between the index i and the index j.

Step3: Calculate eigenvalues and eigenvectors. Calculate the eigenvalues of the correlation matrix R $\lambda_1 \geq \lambda_2 \geq \ldots. \geq \lambda_m \geq 0$ and the corresponding eigenvector $\mu_1, \mu_2, \ldots, \mu_m$, where $\mu_j = [\mu_{1j}, \mu_{2j}, \ldots, \mu_{mj}]^T$, m new index vectors are composed of feature vectors.

ISSN: 2688-8653

$$F_1 = \mu_{11}\tilde{x}_1 + \mu_{21}\tilde{x}_2 + \ldots + \mu_{m1}\tilde{x}_m$$
$$F_2 = \mu_{12}\tilde{x}_1 + \mu_{22}\tilde{x}_2 + \ldots + \mu_{m2}\tilde{x}_m$$
$$\ldots\ldots$$
$$F_n = \mu_{1n}\tilde{x}_1 + \mu_{2n}\tilde{x}_2 + \ldots + \mu_{mn}\tilde{x}_m$$

In the equation: F_1 is the first principal component, F_2 is the second principal component, ..., F_m is the m principal component.

Step 4: Select p(p≤m) principal components to calculate the comprehensive evaluation value.

(1) Calculate the information contribution rate and cumulative contribution rate of the eigenvalue value $\lambda_j (j = 1,2,\ldots,m)$. We call

$$b_j = \frac{\lambda_j}{\sum_{j=1}^{m} \lambda_k}, j = 1,2,\ldots m$$

as the information contribution rate of the main component F_j. At the same time,

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k}$$

is the main component $F_1, F_1, \ldots, F_p$ cumulative contribution rate. When α_p is close to 1, select the first p indicator vectors $F_1, F_1, \ldots, F_p$ as the p principal components instead of the original m index vectors, to conduct a comprehensive analysis of p principal components.

(2) Calculate the comprehensive score:

$$Z = \sum_{j=1}^{p} b_j F_j$$

where b_j is the information contribution rate of the jth principal component, which is evaluated according to the comprehensive score value.

2.3. Solving and Analyzing the Model

With the help of SPSS software, the correlation coefficient matrix R was calculated, as shown in Table 1.

Table 1. Correlation coefficient matrix

	State of the economy	profits	rate	Supply and demand stability	credibility	Credit rating	Whether the default
State of the economy	1.000	-0.247	0.219	-0.050	-0.029	0.256	0.134
profits	-0.247	1.000	0.054	0.017	0.007	0.042	0.065
rate	0.219	0.054	1.000	-0.278	0.151	0.228	0.049
Supply and demand stability	-0.050	0.017	-0.278	1.000	-0.006	0.076	0.098
credibility	-0.029	0.007	-0.151	-0.006	1.000	0.008	0.032
Credit rating	0.256	0.042	0.228	0.076	0.008	1.000	0.637
Whether the default	0.134	0.065	0.049	0.098	0.032	0.637	1.000

ISSN: 2688-8653

Based on the principal component model, the variance percentage and cumulative contribution rate of eigenvalues from X1 to X7 were solved, as shown in Table 2:

Table 2. Total variance interpretation table

composition	total	Percentage of eigenvalue	cumulative	total	Sum of load square percentage	cumulative
1	1.829	26.127	26.127	1.829	26.127	26.127
2	1.374	19.632	45.759	1.374	19.632	45.759
3	1.170	16.719	62.477	1.170	16.719	62.477
4	1.003	14.333	76.810	1.003	14.333	76.810
5	0.711	10.156	86.966			
6	0.583	8.328	95.294			
7	0.329	4.706	100.00			

The main idea of the main component is to reduce the dimension, to explain the majority of the differences in the original data with fewer variables, and to find the main components associated with the explained variables, so it is usually chosen to contribute to the cumulative variance contribution of more than 85% of the k value, which is:

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq 85\%$$

According to the cumulative contribution rate, we should choose k for 5. In order to further determine the optimal number of main components, the debris of each component is shown below as Fig 1:

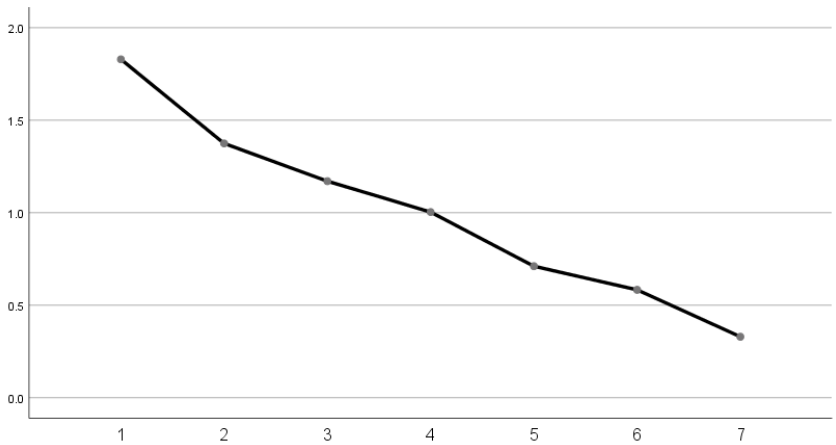

Fig 1. Rubble figure

According to the criterion of selecting the number of main components, the general selection number is analyzed by the main component of the eigenvalue greater than 1. Combined with the gravel diagram, the slope of the steep grade has a significant turning point in the fourth main component, and the cumulative variance contribution of the fourth main component has reached 76.81%, which can reflect most information. Therefore, after the comprehensive

ISSN: 2688-8653

analysis of the eigenvalue and the gravel diagram, the first four main components can explain the information contained in the original index variable.

Using the data in the component matrix, the corresponding coefficient is obtained by the corresponding characteristic root of the main component, and the calculated result is:

$$F_1 = 0.384\tilde{x}_1 - 0.282\tilde{x}_2 - 0.362\tilde{x}_3 + 0.0044\tilde{x}_4 - 0.0163\tilde{x}_5 + 0.216\tilde{x}_6 + 0.0229\tilde{x}_7$$

$$F_2 = 0.706\tilde{x}_1 + 0.37\tilde{x}_2 - 0.532\tilde{x}_3 + 0.304\tilde{x}_4 - 0.0051\tilde{x}_5 - 0.0043\tilde{x}_6 + 0.5724\tilde{x}_7$$

$$F_3 = -0.2616\tilde{x}_1 - 0.4179\tilde{x}_2 - 0.0555\tilde{x}_3 + 0.291\tilde{x}_4 - 0.163\tilde{x}_5 + 0.825\tilde{x}_6 + 0.808\tilde{x}_7$$

$$F_4 = 0.2137\tilde{x}_1 + 0.0609\tilde{x}_2 + 0.02\tilde{x}_3 + 0.775\tilde{x}_4 + 0.398\tilde{x}_5 + 0.067\tilde{x}_6 + 0.0369\tilde{x}_7$$

So, the F_i is the ith main component factor. $\tilde{x}_i$ is the ith parameter that xi is treated with standardized processing. According to the main component coefficient, the four main components are combined by seven indexes, each of which is a linear combination of all indexes, the greater the absolute value of the main component coefficient, the greater the correlation of the main component F_i and the index x_i, which is the more important the index is for the main component, and the negative sign is negative.

Substitute the data in the total variance interpretation graph to obtain:

$$\begin{aligned} F &= \frac{1}{5.376}(1.829F_1 + 1.374F_2 + 1.17F_3 + 1.003F_4) \\ &= 0.34F_1 + 0.2556F_2 + 0.218F_3 + 0.1866F_4 \end{aligned}$$

According to this formula, the corresponding F value of each company can be calculated. The larger the F value is, the lower the credit risk of this company is.

3. Clustering Analysis of Enterprises

3.1. Model Building

The basic idea of cluster analysis is to establish a classification method, which can automatically classify a batch of sample data according to their intimate degree in nature without prior knowledge. Since this is the classification of samples, Q-type cluster analysis in cluster analysis is selected. Here are the relevant steps for clustering:

Step1: Let the sample have n observation data, and the ith observation data is x_i. In this case, each sample can be viewed as a point on the coordinate axis, then the distance between each two points is denoted by $d(x_i, x_j)$ meets the following conditions:

$$\begin{cases} d(x_i, x_j) \geq 0, \\ d(x_i, x_j) = d(x_j, x_i) \\ d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j) \\ i, j, k = 1,2,\ldots n \end{cases}$$

and $d(x_i, x_j) = 0$ if and only if $x_i = x_j$.

Step2: Standardize the data.

ISSN: 2688-8653

$$y_i = \frac{x_i - \mu_i}{s_i}$$

where $\mu_i = \frac{1}{n}\sum_{i=1}^{n} x_i, s_i = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_i)^2}$, x_i represents the value of the ith index variable, μ_i represents the sample mean of the ith index variable, s_i represents the sample standard deviation of the ith index variable, and n represents the number of objects pointed to.

Step3: The distance between each sample point is calculated, and the distance matrix $(d_{ij})_{m\times n}$ is constructed. In this paper, I use Euclidean distance to calculate $d(x_i, y_i)$.

$$d(x_i, x_j) = |x_i - x_j|$$

Step 4: N classes are constructed and each class contains only one sample point, and the platform height of each class is zero. The two classes closest to each other were combined as the new class, and the distance between the two classes was taken as the platform height in the cluster diagram.

Step 5: Draw the clustering diagram and select the number and class of the decision class you need.

Table 3. Average value of F of all kinds of enterprises

Result	
categories	Mean of F
First class	-0.995046285
Second class	0.333154022
Third class	-0.465286724
Four class	0.928602721

Using SPSS software to analyze the f values of each company after the analysis of the main component, the clustering of cluster and the categories of various enterprises (Please refer to the Table 3), which are divided into four categories according to the results of the operation. The classification results are shown in Table 4.

According to the inherent risk degree, the commercial loan is divided into normal, secondary, suspicious and lost four forms. The results of the cluster analysis and the analysis of the data, the fourth class is the normal class, the second type is the secondary class, the third class is the suspicious class, the first category is the loss class. For the loss class, the bank shall not lend; Suspicious classes require reasonable credit based on the long-term situation of the enterprise; Normal and secondary loans can be lent properly.

ISSN: 2688-8653

Table 4. Classification results

First type of enterprise code					
E1	E4	E14	E45	E52	E82
E101	E102	E111	E112	E113	
Second type of enterprise code					
E2	E28	E49	E69	E90	E106
E6	E29	E50	E72	E92	E108
E8	E30	E51	E75	E93	E109
E10	E32	E53	E76	E94	E110
E12	E33	E54	E77	E95	E114
E15	E34	E58	E78	E96	E115
E17	E36	E60	E79	E97	E116
E18	E38	E62	E80	E98	E117
E21	E40	E65	E83	E99	E118
E22	E41	E66	E85	E100	E119
E24	E43	E67	E86	E104	E120
E27	E46	E68	E88	E105	E122
Third type of enterprise code					
E3	E20	E39	E59	E87	E19
E5	E23	E44	E61	E89	E37
E7	E25	E47	E63	E91	E57
E9	E26	E48	E70	E103	E84
E11	E31	E55	E73	E107	E123
E13	E35	E56	E81	E121	
Four type of enterprise code					
E16	E42	E46	E71	E74	

4. Conclusion

Through the processing of the existing index data of each enterprise, seven representative data values are calculated. Principal component analysis is carried out to select four principal components, and the credit risk evaluation value Q of each enterprise is calculated. On this basis, the cluster analysis of enterprises is carried out, which can be divided into four categories: normal, secondary, suspicious and loss category. The mean values of Q of the four categories were -0.995046285, 0.333154022, -0.465286724 and 0.928602721. Then, according to the genealogy, the category of each enterprise can be obtained. According to the classification results, the bank can make credit strategy. For the loss category, the bank will not lend; Suspicious types need to be based on the long-term situation of the enterprise to carry out reasonable credit; Normal and sub-prime can be given normal lending.

References

[1] Si Shoukui, Sun Zhaoliang.Mathematical Modeling Algorithms and Applications.

[2] Shi Jiuyu, Chai Yanyou, Wang Zhiying. Application of Principal Component Analysis in Enterprise Economic Benefit Analysis [J]. Journal of Harbin Engineering University, 2005, (05):137-140.

ISSN: 2688-8653

[3] Shang Yurong. Research on Credit Risk Evaluation of Small and Medium-sized Enterprises in Commercial Banks [D]. Xi 'an University of Science and Technology,2018.

[4] Li Qingdong. Financial Performance Evaluation and Clustering Analysis of Listed Companies [J]. Industrial Technology Economics,2005, (08):146-148.