Hindawi
Complexity
Volume 2018, Article ID 2067065, 9 pages
https://doi.org/10.1155/2018/2067065

Research Article

Empirical Study on Indicators Selection Model Based on Nonparametric K-Nearest Neighbor Identification and R Clustering Analysis

Yan Liu[1], **Zhan-jiang Li**[1] **and Xue-jun Zhen**[2]

[1]*College of Economics and Management, Inner Mongolia Agricultural University, Hohhot 010010, China*
[2]*Huachen Trust Limited Liability Company, Hohhot 010010, China*

Correspondence should be addressed to Zhan-jiang Li; lizhanjiang582@163.com

Received 13 September 2017; Revised 19 February 2018; Accepted 26 February 2018; Published 30 April 2018

Academic Editor: Enzo Pasquale Scilingo

The combination of the nonparametric K-nearest neighbor discriminant method and R cluster analysis is used to construct a double-combination index screening model. The characteristics of the article are as follows: firstly, the nonparametric K-nearest neighbor discriminant method is used to select the indicators which have significant ability to discriminate the default loss rate, which makes up the shortcomings of the previous research that only focuses on the indicators with significant ability to discriminate default state. Additionally, the R cluster analysis applied in this paper sorts the indicators by criterion class, rather than sorting the indicator by the whole index system. This approach ensures that indicators which are clustered in one class have the same economic implications and data characteristics. This approach avoids the situation where indicators that are clustered in one class only have the same data characteristics but have different economic implications.

1. Introduction

The existing research on the influencing factors of credit risk in microenterprises is divided into the following two categories.

(1) Existing Studies on Credit Evaluation Indicators System. Reusens and Croux (2017) think that the government debt, GDP growth rate, inflation, and other macroeconomic factors play a significant role in promoting corporate credit, so they cite these variables to build a credit evaluation index system [1]. Anand et al. (2016) think that indicators such as profitability, liquidity, firm size, and credit rating have an influence on the stability of the firm and play a vital role in credit evaluation. So these indicators should be included into the credit evaluation index system [2]. Jones et al. (2015) built a corporate credit rating system using financial indicators such as total assets. In addition to the above financial variables, Jones also cited the market variable such as enterprise scale and years of establishment into the index system [3]. Doumpos et al. (2015) mainly examined the impact of financial indicators on corporate credit and built a credit evaluation index system including asset returns, interest income, solvency, long-term debt leverage, and the size of the company [4].

(2) Existing Studies on Indicators Selection Methods. Many existing researches establish a classifier from the perspective of fuzzy to solve the credit evaluation problem [5–7]. Sohn et al. (2016) use the fuzzy logic regression method to establish the credit rating equation [8]. Abiyev (2014) develops fuzzy logic and neural network methods to extract important credit risk assessment information [9]. Ju and Sohn (2014) established a credit rating equation to pick up appropriate funding beneficiaries [10]. Elliott et al. (2014) screen out the true information which could reflect the credit state of a company based on a double hidden Markov model (DHMM) [11]. Abellán and Mantas (2014) construct the ensembles of classifiers for bankruptcy prediction and credit scoring based on random subspace method. Experimental studies show that decision tree packaging solutions provide the best results for bankruptcy forecasts and credit scores [12]. Bijak and Thomas

use (2015) improved Bayesian analysis techniques to deal with the problem of loss from bad loans [13]. Gorzałczany and Rudziński (2016) are more concerned about the supervision and division of customer credit ratings than other scholars, which helps banks make better lending decisions [14]. Jones et al. (2015) predict the variation tendency of customer credit levels and determine the credit threshold through the binary classifier [3].

The defects of the existing research are as follows. First, most of the existing research constructs indicators system from the perspective of default and nondefault, which lack the research from the perspective of the default loss rate. Second, some of the existing researches cannot classify the indicators from the perspective of the economic sense of the indicators when using R cluster analysis, so that the existing research cannot remove the indicators which have redundant information.

Contributions of This Paper. First, this paper implements nonparametric K-nearest neighbor discriminant method to remove indicators that cannot significantly distinguish samples of different default loss rate. Second, the paper classifies indicators by R clustering analysis and selects indicators which cover the largest information from each class by coefficient of variation. It ensures that the duplicate information is removed.

2. Research Principle

2.1. The Difficulty of the Problem

Difficulty 1. First difficulty is how to ensure that the selected indicators can significantly differentiate samples which have different default loss rate. In the existing study, the indicators selected by many classic methods can only distinguish different default state.

Difficulty 2. Second difficulty is how to delete the indicators which have the problem of information overlap and redundancy.

2.2. The Method to Solve the Difficulty

The Method to Solve the Difficulty 1. The nonparametric K-nearest neighbor discrimination method will screen out the indicators which have significant discrimination ability on samples that have different default loss rate. If there are h indicators, then h identified accuracy will be calculated. The h identified accuracy is compared with the accuracy of all the indexes, and the accuracy difference between h index and all indicators is obtained. If the accuracy difference between a certain indicator and all indicators is greater than or equal to 0, then delete the index; if the accuracy difference between a certain indicator and all indicators is less than 0, then retain the index. After the above steps, the indicators which have significant discrimination ability on different default loss rate will be selected.

The Method to Solve the Difficulty 2. According to R cluster analysis, the indexes were screened again and the collinearity was excluded. By means of the R cluster analysis, the above indexes were screened out by the nonparametric K-nearest neighbor discrimination method and were reclassified according to criteria layer. The indicators which have largest coefficient of variation of each category of each criteria layer will constitute the final indicator system, and the final indicator system will not cause the problem of information redundancy.

3. Construction of Indicator System

3.1. Indicators' First Selection by Nonparametric K-Nearest Neighbor Discrimination Method

3.1.1. Selection of the Optimal K Value. In this paper, the optimal K value will be selected by error balance method (Xing and Tingjin, 2014) [15]. At the same time, set a constraint for the error balance method. Compared with the method of generalized cross validation, the error balance method can not only get the optimal K value but also reduce the computational cost greatly.

Error balance method makes the K value increase from 1 and combines the test error of all the samples to draw the trend of test error. Finally, according to the trend, determining an optimal K value ensures that the test error is minimum. This method not only specifies the direction of the optimal K value selection, but also ensures that the optimal K value is chosen within the reasonable K value range. This paper combines Góra and Wojna's thought (Góra and Wojna, 2002) with the error balance method to find the best K value [16].

Assume that E_i is test error of the ith type sample; M_i is the number of ith type samples misjudged into other class samples; l_i is the number of actual ith type samples ($i = 1, 2, 3$).

$$E_i = \frac{M_i}{l_i}. \tag{1}$$

Assume that E is the test error of the all sample; E_1 is the test error of the high default loss rate sample; E_2 is the test error of the low default loss rate sample; E_3 is the test error of the nondefault sample; n_1, n_2, and n_3 are the sample size of high default loss rate sample, low default loss rate sample, and nondefault sample.

$$E = E_1 \times \frac{n_1}{n} + E_2 \times \frac{n_2}{n} + E_3 \times \frac{n_3}{n}. \tag{2}$$

The meanings of formulas (1) and (2) are as follows: the ratio of the number of misjudgments to the actual sample size represents the test error, and the weighted average of test errors of the three types sample is the total sample test error.

Assume that E_1 is test error of the high default loss rate sample; E_2 is test error of the low default loss rate sample; E_3 is test error nondefault sample; K is the number of nearest neighbors; n is sample size; n_1, n_2, and n_3 are the sample size of high default loss rate sample, low default loss rate sample, and nondefault sample.

$$\begin{aligned} \min \quad & \left[E_1 \times \frac{n_1}{n} + E_2 \times \frac{n_2}{n} + E_3 \times \frac{n_3}{n}\right] \\ \text{s.t.} \quad & k < \sqrt{n}. \end{aligned} \tag{3}$$

The Meaning of (3). According to Góra and Wojna's theory, the optimal K value should be in the range of $1 \sim \sqrt{n}$. Under the above constraints, the optimal K value is the value that minimizes the total sample test error.

3.1.2. The Process of Index Screening through Nonparametric K-Neighbor Identification Method

(1) Calculate the Prior Probabilities. Assuming that $\widehat{p}_i$ is the prior probability of each class, where $i = 1, 2, 3$, $\widehat{p}_i \geq 0$, and $\widehat{p}_1 + \widehat{p}_2 + \widehat{p}_3 = 1$. n_i is the sample amount of each class. n is the sum of the sample sizes per class (Ganjiang, 2007) [17]:

$$\widehat{p}_i = \frac{n_i}{n}. \tag{4}$$

The Meaning of (4). Calculate the prior probability of each class through calculating the ratio between the sample number of each class and total samples. The smaller the result, the smaller the likelihood that the sample will be classified into the class.

(2) Using K-Nearest Neighbor Estimation to Obtain the Probability Density Function. Assuming: $\widehat{f}_i(x)$ is the probability density function of each class, where $i = 1, 2, 3$. k_i is the k number which belongs to ith type neighbors, and $k_1 + k_1 + k_1 = k$. n_i is the sample amount of each class. $An(k, x)$ is the volume which contains k neighbors on interval $[x - a, x + a]$ (Ganjiang, 2007) [17]:

$$\widehat{f}_i(x) = \frac{k_i}{n_i An(k, x)}. \tag{5}$$

The Meaning of (5). The probability of x falls within the established range.

(3) Calculate Posterior Probability. Assuming that $P(G_i \mid X)$ is the posterior probability of a known category. $\widehat{p}_i$ is the prior probabilities of each class, $\widehat{p}_i \geq 0$, and $\widehat{p}_1 + \widehat{p}_2 + \widehat{p}_3 = 1$. $\widehat{f}_i(x)$ is probability density functions of each class. $\sum_{i=1}^{3} \widehat{p}_i \widehat{f}_i(x)$ is the sum of the product of the probability density function and the prior probability of each class (Ganjiang, 2007) [17]:

$$P(G_i \mid X) = \frac{\widehat{p}_i \widehat{f}_i(x)}{\sum_{i=1}^{3} \widehat{p}_i \widehat{f}_i(x)}. \tag{6}$$

If $P(G_1 \mid X)$ is the largest of three, then the sample should be sent to the class which is high default loss rate; if $P(G_2 \mid X)$ is the largest of three, then the sample should be sent to the class which is low default loss rate; if $P(G_3 \mid X)$ is the largest of three, then the sample should be sent to the class which is nondefault, where 1 minus the error rate equals the accuracy.

(4) Measure the Identification Accuracy of the Default Loss Rate. Assuming that $A_j$is the accuracy of the jth type sample; N_j is the number of jth type samples judged by the nonparametric K-nearest neighbor discriminant method. l_j is the actual number of jth type samples. Then A_j is

$$A_j = \frac{N_j}{l_j}. \tag{7}$$

The Meaning of (7). The larger the calculated value, the better the nonparametric K-nearest neighbor discriminant method which is used to identify different classes of samples.

Assume that A is the identification accuracy of all the sample; there are

$$A = A_1 \times \frac{n_1}{n} + A_2 \times \frac{n_2}{n} + A_3 \times \frac{n_3}{n}. \tag{8}$$

The Meaning of Formula (8). The discrimination accuracy A of all the samples is equal to the weighted average of the discrimination accuracy A_1 of the high default loss rate sample, the discrimination accuracy A_2 of the low default loss rate sample, and the discrimination accuracy A_3 of the nondefault sample. The higher the A, the higher the accuracy of discrimination of all samples.

(5) Calculate the Degree of Influence of the ith Indicator on the Discrimination Accuracy. Assume that I_i is the degree of influence of the ith indicator on the accuracy of the discrimination; A_i is the identification accuracy of the residual indicator after eliminate the ith index; A_0 is the identification accuracy of all the indicators. Then I_i is

$$I_i = A_i - A_0. \tag{9}$$

Formula (9) reflects the degree of influence of the ith index on the accuracy of the discriminant.

(6) Three Criteria of Indicator Screening Based on Nonparametric K-Nearest Neighbor Discriminant

Criterion 1. If the discrimination accuracy A_i of the residual indicators after the ith index is excluded is larger than the discrimination accuracy A_0 of all the indicators, that is to say $I_i > 0$, it means that the accuracy of the discrimination after deleting the index is improved so the index should be removed. Mark the standard as standard one. All the indicators that meet Criteria 1 should be removed.

Criterion 2. If the discrimination accuracy A_i of the residual indicators after the ith index is excluded is equal to the discrimination accuracy A_0 of all the indicators, that is to say $I_i = 0$, it means that the accuracy of the discrimination after deleting the index does not change, so the indicator should be removed. Mark the standard as standard two. All the indicators that meet Criteria 2 should be removed.

Criterion 3. If the discrimination accuracy A_i of the residual indicators after the ith index is excluded is smaller than the discrimination accuracy A_0 of all the indicators, that is to say $I_i < 0$, it means that the accuracy of the discrimination after deleting the index decreases so the index should be retained. Mark the standard as standard three. All the indicators that meet Criteria 3 should be retained.

3.2. Indicators' Second Selection by R Clustering Analysis. R-type clustering, also known as variable clustering, is a method of clustering variables. In order to reflect the characteristics of things and ensure the uniqueness of each index, R-type clustering method should be used to further cluster these variables selected by nonparametric K-nearest neighbor discriminant method to delete information redundancy.

(1) R-Type Clustering Analysis Based on the Method of Squared Sum Method. The R cluster analysis of the indexes in the same criterion layer is carried out by the squared sum method.

Assume that S_i is the sum of square deviation of ith type indicators ($i = 1, 2, 3, \ldots, l$); the n indexes are divided into class l; n_i is the number of the ith type indicator; $X_i^{(j)}$ is the standardized sample value vector ($j = 1, 2, \ldots, n_i$) of the jth indicator in the ith class; $\overline{X}_i$ is the average vector of the ith class of indicators:

$$S_i = \sum_{j=1}^{n_i} \left(X_i^{(j)} - \overline{X}_i\right)' \left(X_i^{(j)} - \overline{X}_i\right). \tag{10}$$

Assume that S is the sum of square deviation of all types of indicators ($i = 1, 2, 3, \ldots, l$):

$$S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_i^{(j)} - \overline{X}_i\right)' \left(X_i^{(j)} - \overline{X}_i\right). \tag{11}$$

Step 1. Treat n indicators as n classes.

Step 2. Combine any two of indicators in those n indictors into one class, no change on indicators left. There are $n(n-1)/2$ kinds of combination. According to (10), calculate each class of indicators' sum of square deviation S_i.

Step 3. Calculate total sum of squares of deviations as to the indicators in all of the classes by (11), and reclassify the indicators in the way of indicators' combination that would minimize the total sum of squares of deviation. k sorts total sum of squares of deviations.

Step 4. Repeat Step 3 until the kind of classification is l.

In the R cluster analysis, the number of reasonable categories is between 2 and 4. In order to avoid the subjective randomness of the number of categories, the nonparametric K-W test of each class after clustering is used to judge the rationality of the classification number Z. The original hypothesis of the nonparametric K-W test is that there are no significant differences in the numerical characteristics of the different indicators.

If the significance level of each category sig > 0.05, then accept the original hypothesis. That is to say, there is no significant difference between the indicators from the same class, and the number of classification is reasonable. On the contrary, indicators should be reclustered.

(2) Analysis of the Size of the Discriminant Force Based on the Coefficient of Variation. An indicator's coefficient of variation reflects its identification ability. The bigger an indicator's coefficient of variation is, the more information content it is contained. Therefore, the indicator with the biggest coefficient of variation within the same class should be retained.

Assume that s_j is the overall standard deviation of the jth indicator; x_j is the mean of the jth indicator; the formula of the coefficient of variation of the jth index is

$$v_j = \frac{s_j}{x_j}. \tag{12}$$

The advantage of the coefficient of variation is that the indicator which has the largest coefficient of variation has a strong ability to distinguish different information, and its role in the comprehensive evaluation is the largest, through removing the index whose coefficient of variation is small to ensure that the index system is simple and effective.

Assume that LGD_j is the default loss rate of the jth sample; L_j is receivable principal and interest of the jth sample which is not repaid now; R_j is receivable principal and interest of the jth sample.

$$\mathrm{LGD}_j = \frac{L_j}{R_j}. \tag{13}$$

4. Empirical Study

4.1. Sample Selection and Data Sources. The loan data of 860 microenterprises in this paper is derived from the credit database of a head office of a commercial bank. There are 830 nondefault customers and 30 default customers. Each sample includes 68 indicators such as Debt Asset Ratio, Acid-Test Ratio, and Operating Profit Ratio, which is shown in columns 2–69 of Table 1. The default loss rate calculated by formula (13) is set out in column 72 of Table 1. According to the type of the default loss rate, 860 customers will be divided into three categories and placed in column 73 of Table 1.

4.2. Screening of Indexes Based on Nonparametric K-Nearest Neighbor Discriminant Method. Select the optimal value of k. In this paper, the sample size is 860, and k value should be smaller than the square root of the sample size, so the value of k is less than $\sqrt{860} \approx 29.32$. k should belong to $[1, 29]$.

Find the best k value. Combined with the objective function, the best value of k can make the test error of all samples be the smallest. The test error of each value of $K = 1, 2, 3, \ldots, 29$ is used to draw the trend of test error and k value.

Table 1: Loan data of 860 microenterprise customers.

(1) Serial number	Standardized data x_{ij} (2) Debt Asset Ratio	…	(69) Arrived pledge score	(70) Receivables are not received	(71) Receivable principal and interest	(72) Default loss rate	(73) Type of default loss rate
1	0.384246284	⋯	0.1	236825.96	236825.96	1	2
…	…	…	…	…	…	…	…
25	0.497509115	⋯	0.1	4045109.5	4890783	0.827088	1
…	…		…	…	…	…	…
860	0.809515108	⋯	0.348824889	0	1635339.76	0	0

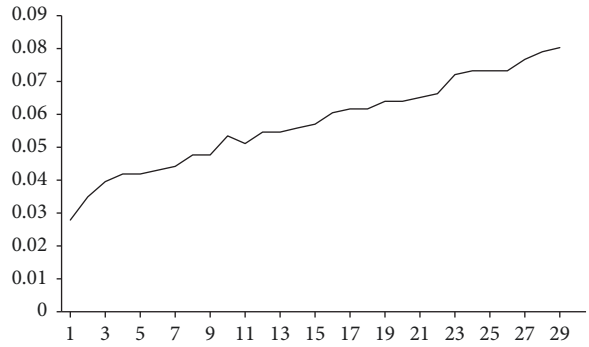

Figure 1: The trend of k value.

Determine the optimal k value. It can be seen from Figure 1 that the k value corresponding to the minimum test error is 1, so the value of k is 1.

The specific process of screening indicators is based on nonparametric K-nearest neighbor discriminant method. The indicators are placed in column 1 of Table 2.

The discriminant accuracy A_0 of 68 indices is obtained by nonparametric K-nearest neighbor discrimination.

Step 1. Calculate the discriminant accuracy HDA_0 of the high default loss rate sample. Among the 24 high default loss rate samples, the number of samples that were accurately discriminated by nonparametric K-nearest neighbor discriminant method was 13. According to formula (7), the discriminant accuracy HDA_0 of the high default loss rate sample is $\mathrm{HDA}_0 = M'/l_1 = 13/24 = 54.17\%$, placed in column 2 of Table 2.

Step 2. Calculate the discriminant accuracy LDA_0 of the low default loss rate sample. Among the 6 low default loss rate samples, the number of samples that were accurately discriminated by nonparametric K-nearest neighbor discriminant method was 0. According to formula (7), the discriminant accuracy LDA_0 of the low default loss rate sample is $\mathrm{LDA}_0 = M''/l_2 = 0/6 = 0\%$, placed in column 2 of Table 2.

Step 3. Calculate the discriminant accuracy UA_0 of the nondefault sample. Among the 830 nondefault samples, the number of samples that were accurately discriminated by nonparametric K-nearest neighbor discriminant method was 823. According to formula (7), the discriminant accuracy UA_0 of the nondefault sample is $\mathrm{UA}_0 = N/l_3 = 823/860 = 99.16\%$, placed in column 2 of Table 2.

Step 4. Calculate the discriminant accuracy A_0 of all samples. Putting $\mathrm{HDA}_0 = 54.17\%$, $\mathrm{LDA}_0 = 0\%$, and $\mathrm{UA}_0 = 99.16\%$ into formula (8), then we can obtain the discrimination accuracy A_0 of all the samples: $A_0 = (n_1/n) * \mathrm{HDA}_0 + (n_2/n) * \mathrm{LDA}_0 + (n_3/n) * \mathrm{UA}_0 = (24/860) * 54.17\% + (6/860) * 0\% + (830/860) * 99.16\% = 97.2127\%$, placed in column 2 of Table 2.

One of the 68 indicators is deleted one by one, and the discriminant accuracy A_i of the remaining 67 indicators is calculated by the nonparametric K-nearest neighbor discriminant method.

The discriminant accuracy of the high default loss rate sample, the discriminant accuracy of the low default loss rate sample, the discriminant accuracy of the nondefault sample, and the discriminant accuracy of the total sample can be obtained by using the 67 indicators after removing the index X_1, placed in the first row of Table 2. Similarly, remove the X_2 to X_{68} one by one, and calculate the discriminant accuracy of the high default loss rate sample, the low default loss rate sample, the nondefault loss rate sample, and the discriminant accuracy of the total sample, placed in the other rows of Table 2. Substitute A_i and A_0 into (9), $I_i = A_i - A_0$, and then calculate the influence degree of the ith index on the discrimination accuracy; the degree of influence is placed in column 7 of Table 2.

Screen indicators based on the degree of discrimination of different indicators.

Standard 1 (remove indicators whose $I_i > 0$). According to the degree of influence I_i of the second column of Table 3, the degree of influence I_i of X_7, X_{14}, and X_{48} is larger than 0. Discrimination accuracy can be improved if this type of indicators is eliminated and the results are placed in the corresponding row in column 3 of Table 3.

Standard 2 (remove indicators whose $I_i = 0$). According to the degree of influence I_i of the second column of Table 3, the degree of influence I_i of X_1, X_{21}, X_{27}, and X_{68} is equal to 0. Discrimination accuracy will not change if this type of indicators is eliminated and the results are placed in the corresponding row in column 3 of Table 3.

Standard 3 (retain indicators whose $I_i < 0$). According to the degree of influence I_i of the second column of Table 3, the degree of influence I_i of X_{51}, X_{64}, and X_{65} is less than 0. Discrimination accuracy will decrease if this type

Table 2: The result of nonparametric K-nearest neighbor discriminant model.

Serial number	(1) Index	(2) The accuracy of all the indicators	The identification accuracy after removing the ith index (3) Nondefault customer's judgment accuracy UA_i	(4) Low default loss rate customer's judgment accuracy LDA_i	(5) High default loss rate customer's judgment accuracy HDA_i	(6) All customer's judgment accuracy A_i	(7) The degree of influence on the accuracy of the default state $I_i = A_i - A_0$
1	X_1 Debt Asset Ratio		99.16%	0.00%	54.17%	97.2127%	0%
...	...		...	...	...	...	...
7	$X7$ Equity Ratio		99.16%	0.00%	54.17%	97.2129%	0.0002%
...	...		...	...	...	...	...
14	X_{14} total liabilities net cash flow		99.28%	0.00%	54.17%	97.3285%	0.1158%
...	...		...	...	...	...	...
21	X_{21} EBITDA		99.16%	0.00%	54.17%	97.2127%	0%
...	...	$HDA_0 = 54.17\%$	...	...	...	...	...
27	X_{27} liquidity turnover rate	$LDA_0 = 0.00\%$	99.16%	0.00%	54.17%	97.2127%	0%
...	...	$UA_0 = 99.16\%$	...	...	...	...	...
48	X_{48} marital status	$A_0 = 97.2127\%$	99.16%	0.00%	58.33%	97.3287%	0.116%
...	...		...	...	...	...	...
51	X_{51} gender		99.16%	0.00%	50.00%	97.0963%	−0.1164%
...	...		...	...	...	...	...
64	X_{64} GDP growth rate		99.04%	0.00%	54.17%	97.0968%	−0.1159%
65	X_{65} CPI		99.04%	0.00%	54.17%	97.0968%	−0.1159
...	...		...	...	...	...	...
68	X_{68} arrived pledge score		99.16%	0.00%	54.17%	97.2127%	0%

Table 3: The result of indicator screening based on nonparametric K-nearest neighbor identification model.

	(1) Index	(2) Influence level I_i	(3) Result	(4) Standard
1	X_1 Debt Asset Ratio	0%	Delete	2
...	...	...	...	...
7	$X7$ Equity Ratio	0.0002%	Delete	1
...	...	...	...	...
14	X_{14} total liabilities net cash flow	0.1158%	Delete	1
...	...	...	...	...
21	X_{21} EBITDA	0%	Delete	2
...	...	...	...	...
27	X_{27} liquidity turnover rate	0%	Delete	2
...	...	...	...	...
48	X_{48} marital status	0.116%	Delete	1
...	...	...	...	...
51	X_{51} gender	−0.1164%	Reserved	3
...	...	...	...	...
64	X_{64} GDP growth rate	−0.1159%	Reserved	3
65	X_{65} CPI	−0.1159	Reserved	3
...	...	...	...	...
68	X_{68} arrived pledge score	0%	Delete	2

Table 4: The retained indicator after the first indicator filters.

(1) Number	(2) Indicator
1	X_9 net assets and year-end loan balance ratio
2	X_{19} Operating Profit Ratio
3	X_{20} costs net profit margin
4	X_{23} net cash flow from operating activities
5	X_{28} fixed assets turnover rate
6	X_{29} shareholders' equity turnover rate
7	X_{37} whether to audit
8	X_{51} gender
9	X_{64} GDP growth rate
10	X_{65} CPI

of indicators is eliminated and the results are placed in the corresponding row in column 3 of Table 3.

Indicator Screening Results. 58 indicators were excluded from the 68 indicators, and 10 indexes were retained. Table 4 shows the retained indicators by nonparametric K-nearest neighbor discrimination.

4.3. Indicators' Second Selection by R Clustering Analysis

4.3.1. R Clustering Analysis Based on the Method of Squared Sum Method. The 10 indicators reserved from the previous screening are classified according to the criteria layer. The classification results are shown in Table 5.

R clustering analysis is used in the same criterion layer to classify indicators, and K-W test is used to test different classification results.

(1) In the criterion layer of solvency, it can only be divided into one class and no K-W test because there is only one indicator. The criterion layers of the basic situation of legal representative and nonfinancial factors within the enterprise are similar to the criterion layer of solvency, so X_9, X_{51}, and X_{37} should be reserved.

Table 5: Indicators selection based on R clustering analysis.

Criteria layer	Index
Solvency	X_9
Profitability	$X_{19}X_{20}X_{23}$
The basic situation of legal representative	X_{51}
Operating capacity	$X_{28}X_{29}$
Nonfinancial factors within the enterprise	X_{37}
Enterprise external macroconditions	$X_{64}X_{65}$

(2) In the criterion layer of operating capacity, there are two indicators. Firstly, two indicators can be divided into one class. The result of the K-W test for these two indicators is $p < 0.05$, which indicates that the original hypothesis with the same data feature between X_{28} and X_{29} is refused, so X_{28} and X_{29} have different data feature and should be reserved simultaneously.

(3) In the criterion layer of enterprise external macroconditions, there are two indicators. Firstly, two indicators can be divided into one class. The result of the K-W test for these two indicators is $p < 0.05$, which indicates that the original hypothesis with the same data feature between X_{64} and X_{65} is refused, so X_{64} and X_{65} have different data feature and should be reserved simultaneously.

(4) In the criterion layer of profitability, there are three indicators. Firstly, three indicators can be divided into 2 classes. The result of the K-W test for the two indicators among 3 indicators is $p < 0.05$, which indicates that the original hypothesis with the same data feature is refused, and

Table 6: Comparative analysis of 3 models.

Model	Judgement accuracy			
	Nondefault customer's judgment accuracy	Low default loss rate customer's judgment accuracy	High default loss rate customer's judgment accuracy	All customer's judgment accuracy
	UA_i	LDA_i	HDA_i	A_i
Combined model which preserved in this paper	100%	100%	100%	100%
neural network model	99.8%	50%	75%	98.95%
Stepwise discriminant analysis	96.75%	100%	83.33%	96.40%

two indicators should be clustered into 2 class. In this case, there is no need to divide three indicators into one category. Finally, in this criterion layer, three indicators should be divided into three categories. So X_{19}, X_{20}, and X_{23} should be reserved simultaneously.

The classification results of indicators are as follows. (1) In the criterion layer of solvency, X_9 should be reserved. (2) In the criterion layer of the basic situation of legal representative, X_{51} should be reserved. (3) In the criterion layer of nonfinancial factors within the enterprise, X_{37} should be reserved. (4) In the criterion layer of operating capacity, X_{28} and X_{29} should be reserved simultaneously. (5) In the criterion layer of enterprise external macroconditions, X_{64} and X_{65} should be reserved simultaneously. (6) In the criterion layer of profitability, X_{19}, X_{20}, and X_{23} should be reserved simultaneously.

4.3.2. Analysis of the Size of the Discriminant Force Based on the Coefficient of Variation. R clustering analysis shows that there is no redundant information in each index layer, so there is no need to use the coefficient of variation to delete the index with weaker recognition ability. So far, the paper has completed the second index screening process.

By the application of nonparametric K-nearest neighbor discriminant method and R clustering analysis, the paper establishes a small enterprises credit evaluation indicators system, which contains 6 principle layers and 10 indicators.

4.4. Comparative Analysis. In order to reflect the superiority of combined model of the nonparametric K nearest neighbor discriminant and the R clustering proposed in this paper, the comparative analysis of the combined model with stepwise discriminant analysis and neural network model will be carried out. The superiority of an indicator screening model can be reflected in the indicators selected by the model having higher identification ability. Therefore, this article will compare the discriminatory power of the three models.

Comparative analysis includes the following two steps.

Step 1. The combined model, stepwise discriminant analysis model, and neural network model will be used, respectively, to screen indicators that have significant discriminating ability on default loss rate.

Step 2. Use the selected index system to test the discrimination ability of the model. The higher the discriminative power of the model, the greater the superiority of the model.

Table 6 shows the discriminating ability of the three models for all types of samples. The discriminatory power of the combined models is higher than the stepwise discriminant analysis model and the neural network model, no matter for the discrimination ability of some samples or the discrimination ability of all the samples. Therefore, the combination model has more superiority than the other two models: that is to say the index system screened by the combination model has stronger identification ability. In addition, the combinatorial model is also more suitable for analyzing multiclassification problems because it has higher discriminative power when dealing with multiclassification problems.

5. Conclusion

5.1. The Main Conclusions. The credit index system which contains 10 indicators is selected by the model combination of nonparametric K-nearest neighbor discrimination method and R cluster analysis.

5.2. The Characteristics of This Article. First, the nonparametric K-nearest neighbor discrimination method is used to select the indicators which have significant discriminant ability on samples with different default loss rate. The study in this paper makes up the deficiency of previous studies which mainly focus on the default state.

Second, the R cluster analysis applied in this paper is based on the criterion layer rather than the whole index system. This will ensure that the indicators clustered into the same class have the same economic implications and data features, which avoid the clustering of indicators which mainly focus on the same data characteristics but ignore different economic implications.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of the paper.

Acknowledgments

The research is supported by the Key Project of National Natural Science Foundation of China (71731003), China Postdoctoral Science Foundation (2015M582746XB), and Natural Science Foundation of Inner Mongolia Autonomous Region of China (2016MS0714). The authors would like to show great gratitude to the organizations mentioned above.

References

[1] P. Reusens and C. Croux, "Sovereign credit rating determinants: A comparison before and after the European debt crisis," *Journal of Banking & Finance*, vol. 77, pp. 108–121, 2017.

[2] V. Anand, K. Soomro A, and K. Solanki S, "Determinants of Credit Rating and Optimal Capital Structure among Pakistani Banks," *Romanian Journal of Economic Forecasting*, vol. 19, no. 60, pp. 169–182, 2016.

[3] S. Jones, D. Johnstone, and R. Wilson, "An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes," *Journal of Banking & Finance*, vol. 56, no. 4, pp. 72–85, 2015.

[4] M. Doumpos, D. Niklis, C. Zopounidis, and K. Andriosopoulos, "Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from European listed firms," *Journal of Banking & Finance*, vol. 50, pp. 599–607, 2015.

[5] M. A. Sanchez, O. Castillo, and J. R. Castro, "Generalized Type-2 Fuzzy Systems for controlling a mobile robot and a performance comparison with Interval Type-2 and Type-1 Fuzzy Systems," *Expert Systems with Applications*, vol. 42, no. 14, pp. 5904–5914, 2015.

[6] M. Mansouri, M. Teshnehlab, and M. Aliyari Shoorehdeli, "Adaptive variable structure hierarchical fuzzy control for a class of high-order nonlinear dynamic systems," *ISA Transactions*, vol. 56, pp. 28–41, 2015.

[7] S. F. Derakhshan and A. Fatehi, "Non-monotonic robust H2 fuzzy observer-based control for discrete time nonlinear systems with parametric uncertainties," *International Journal of Systems Science*, vol. 46, no. 12, pp. 2134–2149, 2015.

[8] S. Y. Sohn, D. H. Kim, and J. H. Yoon, "Technology credit scoring model with fuzzy logistic regression," *Applied Soft Computing*, vol. 43, pp. 150–158, 2016.

[9] R. H. Abiyev, "Credit rating using type-2 fuzzy neural networks," *Mathematical Problems in Engineering*, vol. 2014, Article ID 460916, 8 pages, 2014.

[10] Y.-H. Ju and S.-Y. Sohn, "Updating a credit-scoring model based on new attributes without realization of actual data," *European Journal of Operational Research*, vol. 234, no. 1, pp. 119–126, 2014.

[11] R. J. Elliott, T. K. Siu, and E. S. Fung, "A Double HMM approach to Altman Z-scores and credit ratings," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1553–1560, 2014.

[12] J. Abellán and C. J. Mantas, "Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3825–3830, 2014.

[13] K. Bijak and L. C. Thomas, "Modelling LGD for unsecured retail loans using Bayesian methods," *Journal of the Operational Research Society*, vol. 66, no. 2, pp. 342–352, 2015.

[14] M. B. Gorzałczany and F. Rudziński, "A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability," *Applied Soft Computing*, vol. 40, pp. 206–220, 2016.

[15] W. Xing and C. Tingjin, *Nonparametric Statistics*, Tsinghua University Press, Beijing, China, 2nd edition, 2014.

[16] G. Góra and A. Wojna, "RIONA: a new classification system combining rule induction and instance-based learning," *Fundamenta Informaticae*, vol. 51, no. 4, pp. 369–390, 2002.

[17] Z. Ganjiang, *Application of Nonparametric Density Estimation in Discriminant Analysis*, Nanjing Information Engineering University, 2007.