Empirical Study on Indicators Selection Model Based on significant Discrimination and R Clustering Analysis

Lingling Gong [1, a], Guotai Chi [1], Baofeng Shi [2, b] and Wei Yao[3]
[1] *Dalian University of Technology, Dalian, Liaoning 116024, China*
[2] *Northwest A&F University, Yangling, Shaanxi 712100, China*
[3] *Dalian Vocational Technology College, Dalian, Liaoning 116024, China*
[a]*sjmechelle@yeah.net,* [b]*fengbei_wuyu@163.com*

Abstract:

Small enterprises play the important role in pushing China's economic progress, but keeping on facing the difficulty in financing the loans. Establishing a reasonable credit evaluation indicators system is one of the keys to implement accurate credit evaluate to small enterprises. Regardless of the evaluation method being used, with unsuitable indicators system, it is impossible to obtain reasonable credit evaluation results. By the application of logistic regression significant discrimination and R clustering analysis, a small enterprises credit evaluation indicators system is established. The credit evaluation system established in this paper is capable of significantly discriminating default samples from non-default ones and can effectively avoids duplicate information. The result of empirical study shows that the credit evaluation indicators system established in this paper is able to reflect 83.47% of original information with 22.22% of original indicators.

Keywords: credit evaluation; indicators selection; logistic regression; R clustering analysis.

Introduction

Most of the small enterprises in China are keeping on facing the bottleneck of financing the funds. One of the key reasons is that commercial banks are afraid of lending loans to small enterprises, since there is lack of applicable credit evaluation system. Most of existing credit evaluation indicators system, which are suitable to large and medium enterprises, normally require sophisticated analysis of financial information, are not applicable to small enterprises. Because most of small enterprises are lack of systematic retaining of financial records and their financial information is normally not easy to obtain; non-financial factors such as the business operation and willingness on loan repayment are significantly affected by owners' personal factors; the performance and profitability of the enterprise is easily affected by regional and sector's economic status [1].

Establishing a reasonable credit evaluation indicators system is the key to make reasonable credit evaluation for small enterprises.

Existing Study on Credit Evaluation Indicators System. Credit evaluation indicators systems used by world famous authorized institutions include: Five Cs and Camels used by lots of banks, and credit evaluation system used by Moody, Standard & Poor and Fitch Ratings etc. [2]. Typical small enterprises credit evaluation systems used by commercial banks in China include: the small enterprises credit risk evaluation indicators system introduced by Industrial and Commercial Bank of China, Agricultural Bank of China, and merchants credit risk evaluation indicators system used by Postal Savings Bank of China etc. [3].

Examples of Credit evaluation indictors system established by academic literatures includes: Shi et al. (2014) establishes farmers' micro-finance credit rating indicators system which consists of 13 indicators such as age, marriage status and GDP growth rate [4]. Hai et al. (2013) establishes farmers' credit rating indicators system which consists of 15 indicators such as aim of loan, Engel coefficient [5]. Lugovskaya(2010) conducts empirical study on default risk of Russian small and medium enterprises' by discriminate analysis and demonstrates that liquidity and profitability are key factors of forecasting small and medium enterprises' default status [6].

Existing Study on Indicators Selection Methods. Examples of indicators selection methods for complex evaluation systems are as follows: Xie et al.(2012) uses three quantitative analysis methods including membership, correlation and discrimination, establishes small and medium enterprises technology innovation capability evaluation indicators system [7]; Zhou et al.(2010) establishes a comprehensive human development evaluation indicators system by using R clustering analysis [8].

Examples of credit evaluation indicators selection methods are as follows: Shi et al. (2013) uses correlation analysis and probit regression significant discrimination method establishes merchant credit rating indicators selection model, and make empirical study on 2157 merchant samples in China [9]. Kim et al. (2012) carries out study on credit scoring to loan enterprises by indictors such as owners' equity, sales revenue, total liabilities, average sales turnover to employee ratio and cash flow to total assets ratio [10].

Contributions of this paper. First, in order to solve the problems mentioned above, this paper implements logistic regression significant discrimination to remove indicators that cannot significantly distinguish between default and non-default samples. It ensures indicators retained could effectively discriminate default status. Second, the paper classifies indicators by R clustering analysis, selects indicators embraced the largest information from each class by coefficient of variation value. It ensures that, those indicators reflect duplicate information is removed.

Construction of credit evaluation system

Establishment of extensive indicators set and indicators' initial filtering. Emphasis on high frequent indicators used by authorized institutions such as Moody's, Standard & Poor's, FICO, Industry and Commercial Bank of China and literatures, based on primary selecting, the paper constructs an extensive indicators set with 4 principle layers of 84 indicators.

Primary selection of indicators is based on the rule that the data of indicators should all be obtainable. Three Indicators, such as "$x_{1,49}$ Post loan asset-liability ratio", are removed and be label as " data unobtainable" in Table 1, column e.

Data standardization. The purpose of data standardization is to turn the data to fall into the range of [0,1] and to ensure the results of credit evaluation to be free from dimensional differences [12].

Positive indicators, such as "$x_{1,5}$ Net profit cash flow rate" are indicators that the greater value it is, the better credit status it represents. Let: v_{ij} be the original value of the the i^{th} sample of the j^{th} indicator, n be the total amount of samples, the i^{th} sample of the j^{th} positive indicator's standardized value x_{ij} could be calculated by equation (1) [12] as follows:

$$x_{ij} = \frac{v_{ij} - \min\limits_{1\le i\le n}(v_{ij})}{\max\limits_{1\le i\le n}(v_{ij}) - \min\limits_{1\le i\le n}(v_{ij})} . \tag{1}$$

Negative indicators, such as "$x_{1,8}$ Accounts payable turnover ratio" are indicators that the smaller value it is, the better credit status it represents. Let: v_{ij} be the original value of the i^{th} sample of the j^{th} indicator, n be the total amount of samples, the i^{th} sample of the j^{th} negative indicator's standardized value x_{ij} could be calculated by equation (2) [12] as follows: (1)

$$x_{ij} = \frac{\max\limits_{1\le i\le n}(v_{ij}) - v_{ij}}{\max\limits_{1\le i\le n}(v_{ij}) - \min\limits_{1\le i\le n}(v_{ij})} . \tag{2}$$

Ideal interval indicators, such as "$x_{3,3}$ Consumer price index" is the indicator that reflects ideal credit status if its value falls into a certain interval. For example, Consumer price index fall into the interval of [101,105] [2], means no inflation and deflation. Thus, the ideal interval for indicator "$x_{3,3}$ Consumer price index" is [101,105]. Let: q_1, q_2 be the left and right margin for ideal interval, meaning of other symbols are the same as they are in Equation (1), the i^{th} sample of the j^{th} ideal interval indicator's standardized value x_{ij} could be calculated by equation (3) [12] as follows:

$$x_{ij}=\begin{cases}1-\dfrac{q_1-v_{ij}}{\max(q_1-\min\limits_{1\le i\le n}(v_{ij}),\max\limits_{1\le i\le n}(v_{ij})-q_2)}, & v_{ij}<q_1 \quad (a)\\ 1-\dfrac{v_{ij}-q_2}{\max(q_1-\min\limits_{1\le i\le n}(v_{ij}),\max\limits_{1\le i\le n}(v_{ij})-q_2)}, & v_{ij}>q_2 \quad (b)\\ 1 \qquad\qquad\qquad\qquad , q_1\le v_{ij}\le q_2 \quad (c)\end{cases}. \tag{3}$$

Table 1 The Extensive Indicators Set of Small Enterprises Credit Evaluation

No.	(a) Principle	(b) Name of indicators		(c) Result of	(d) Name of indicators		(e) Result of
1	x_1 Internal financial factors	$x_{1,1}$ Current ratio	1	**Retained**	$x_{1,10}$ Net operating activities cash flow current liabilities cover ratio	1	Removed by R clustering analysis
2		$x_{1,2}$ Capital immobilized ratio	2		...	...	
3		$x_{1,3}$ Cash ratio	1		$x_{1,20}$ Retained earnings growth rate	1	
4		$x_{1,4}$ Operating margins	1		$x_{1,21}$ Total assets cash recovery rate	1	Removed by logistic regression significant analysis
5		$x_{1,5}$ Net profit	1		...	...	
6		$x_{1,6}$ Inventory turnover ratio	1		$x_{1,48}$ Total assets growth rate	1	
7		$x_{1,7}$ Total assets turnover ratio	1		$x_{1,49}$ Post loan asset-liability ratio	1	data unobtainable
8		$x_{1,8}$ Accounts payable turnover rate	2		$x_{1,50}$ Post loan operating net cash flow rate	1	
9		$x_{1,9}$ Cash turnover cycle	1		$x_{1,51}$ Post loan total asset cash recovery rate	1	
10	x_2 Internal non-financial	$x_{2,1}$ Years of experiences in related industry	4	**Retained**	$x_{2,13}$ Auditing status	4	Removed by logistic

11	factors	$x_{2,2}$ Patent status	4		...	...	regression significant analysis
12		$x_{2,3}$ Enterprises set up a date	4		$x_{1,48}$ Total assets growth rate	1	
13		$x_{2,4}$ Product sales range	4		$x_{2,6}$ Representative automotive and real estate value	1	Removed by R clustering analysis
14		$x_{2,5}$ Living condition	4		...	...	
15					$x_{2,12}$ Legal disputes	4	
16	x_3 External macro economic factors	$x_{3,1}$ Industry prosperity index	1	**Retained**	$x_{3,4}$ Engel coefficient	2	Removed by R clustering analysis
17		$x_{3,2}$ Urban and rural residents per capita savings balance	1		$x_{3,5}$ Urban per capita disposal income	1	
18		$x_{3,3}$ Consumer price index [101,104]	3		$x_{3,6}$ GDP growth rate	1	Removed by logistic regression significant analysis
19	x_4 pledge status	$x_{4,1}$ Mortgage, pledge and guarantee status	4				

Note: 1- Positive indicator, 2- Negative indicator, 3-Ideal interval, 4- Qualitative indicator.

By rational analysis, qualitative indicators that cannot conduct data standardization are normalized according to certain reasonable rule as shown in Table 2, column b-c.

Indicators' First selection by logistic regression significant discrimination method. (1) Establishment of logistic regression model. Let: Y be the dependent variable, represent default status of small enterprises' loan sample. Y=1 means default and Y=0 means non-default. Let: $P(Y=1|x_1,x_2,\ldots x_j)$ be the probability of default while conducting credit evaluation by indicators $x_1,\ldots,x_j$; x_j be the j^{th} indicator, β_0 be the constant number, $\beta_1, \beta_2,\ldots,\beta_j$ be the coefficients in Equation of logistic regression, illustrate the logistic regression model [9] as follows:

$$P(Y=1|x1,x2,\ldots,xj) = \frac{exp(\beta_0 + Z)}{1+exp(\beta_0 + Z)} = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j)}{1+exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j)}. \quad (4)$$

In Equation (4), $Z=\beta_1 x_1+\beta_2 x_2+\ldots+\beta_j x_j$.

(2)Selecting Indicators by logistic regression significant discrimination. The original assumption: if the j^{th} indicator has no effect on small enterprises' loan default status, the coefficient of logistic regression of the j^{th} indicator $\beta_j=0$. The alternative assumption: if the j^{th} indicator significantly effect on small enterprises' loan default status, the coefficient of logistic regression of the j^{th} indicator $\beta_j\neq 0$. Establishing the Wald statistics and judging: whether the coefficients $\beta_1, \beta_2,\ldots,\beta_j$ in Equation (4) is 0 or not. In another words, it is to make judgment on whether the j^{th} indicator would significantly effect on small enterprises' loan default status. If yes, it means the j^{th} indicator would affect the default status significantly, then the j^{th} indicator's coefficient in logistic regression β_j is 0, and this indicator should be retained. If no, it means the coefficient of the j^{th} indicator β_j is not equal to 0 [13].

Let: W_j be the j^{th} indicator's Wald test statistics; $\hat{\beta}_j$ be the j^{th} indicator's estimated coefficient of logistic regression; $S_{\hat{\beta}_j}$ be the standard deviation of $\hat{\beta}_j$, and $W_j(j=1,\ldots, k)$ could be calculated by equation (5) [13] as:

$$W_i = \frac{\hat{\beta}_j^2}{S_{\hat{\beta}_j}^2}. \quad (5)$$

Let: the level of significance $\alpha=0.05$[20], compare W_j's test probability sig with the level of significance α. If $sig \leq \alpha$, thus $\beta_j\neq 0$, which means the j^{th} indicator effect small enterprises' loan default status significantly; if $sig \geq \alpha$, thus $\beta_j=0$, which means the j^{th} indicator does not have significant effect on small enterprises' loan default status.

Compare each $\hat{\beta}_j$ with 0, if $\hat{\beta}_j\neq 0$, the j^{th} indicator should be retained; if $\hat{\beta}_j=0$, the j^{th} indicatorshould be removed [13].

Table 2 The Standardization of Qualitative Indicators

(q) Qualitative indicators	(2) Standards	(3)Result of standardization
$x_{2,1}$ Years of experiences in related industry	(1) Relevant working experience ≥8years;	1.000
	(2) 5years ≤ relevant working experience < 8years ;	0.700
	(3) 2years ≤ relevant working experience < 5years ;	0.400
	(4) 0 < relevant working experience <2 years，or no experience.	0.000
…	…	…
$x_{4,1}$ Pledge status	(1) National treasury pledge;	1.000
	(2) Banker's acceptance ;	0.950
	…	…
	(20) Without pledge evidence .	0.000

The second indicators selection based on R clustering analysis. (1) Idea of indicators selection based on R clustering analysis method. Classifying indicators in the same principle layers by R clustering analysis and clusteringing indicators that reflect the same information into one class. By doing so, indicators in different classes reflect different data characteristics. It ensures that the information reflected by indicators selected from different classes is not duplicated and redundant.

The reason why implement R clustering analysis on indicators by principle layers, in stead of implement R clustering analysis to all the indicators as a whole is: R clustering analysis classifies the indicators merely based on the characteristic of indicators' data, but does not take the economic meaning of indicators into consideration. Implementing R clustering analysis by principle layers could ensuring the indicators clusteringed in the same class embrace the same economic meanings and data characteristics, avoiding of clusteringing indicators merely with the same data characteristics but different economic meanings into the same class.

(2)The steps of R clustering analysis. Implementing R clustering analysis to indicators based on sum of squares of deviation, by calculating sum of squares of deviation to each class of indicators, and determining sorts of clusteringing classes in the aim of ensuring the total sum of squares of deviation of indicators in all of the classes is minimized. The steps of R clustering analysis are as follows [12]:

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 $n(n\text{-}1)/2$ kinds of combination. According to Equation (5), calculate each class of indicators' sum of square deviation S_i.

If clustering n indicators into l classes, let S_i be the i^{th} class's sum of square deviation; n_i be the number of the i^{th} indicator; $X_i^{(j)}$ be the standardized sample value vector (j=1,2,…,n_i) of the j^{th} indicator in the i^{th} class; $\overline{X}_i$ be the average vector of the i^{th} class of indicators, and the i^{th} class's sum of square deviation S_i could be calculated by Equation (6) [4] as:

$$S_i = \sum_{j=1}^{n_i} (X_i^{(j)} - \overline{X}_i)'(X_i^{(j)} - \overline{X}_i) . \tag{6}$$

Step3：calculate total sum of squares of deviations as to the indicators in all of the classes by Equation (7), re-classify the indicators in the way of indicators' combination that would minimize the total sum of squares of deviation. k sorts of total sum of squares of deviations S could be calculated [4] as:

$$S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_i^{(j)} - \overline{X}_i)'(X_i^{(j)} - \overline{X}_i) . \tag{7}$$

Step4: repeat Step 3 until the kinds of classification is l.

(3) Indicators selection reflecting maximum information content

An indicator's coefficient of variation reflects its identification ability in the evaluation system. The bigger an indicator's coefficient of variation is the more

information content it is embraced. Therefore, based on the idea of information content maximizing, when deciding which indicator in each of the class should be retained by R clustering analysis, the one with the biggest coefficient of variation should be retained.

Let: v_j be the coefficient of variation of the j^{th} indicator; n be the number of evaluation samples; $\overline{x}_j$ be the average of the j^{th} indicator, x_{ij} be the value of the i^{th} evaluation sample of j^{th} indicator's; and v_j could be calculated by Equation (8) [12] as:

$$v_j = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2}}{\overline{x}_j}, \tag{8}$$

While, $\overline{x}_j$ is calculated by Equation (9) [12] as:

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}. \tag{9}$$

Selecting of indicators based on R clustering analysis and coefficient of variation method would prevent the credit evaluation indicators system from reflecting redundant information and ensuring the indicators with maximum information content are retained.

Reasonability judgment on credit evaluation indicators system. (1) Idea of reasonability judgment on credit evaluation indicators system. Based on the information contents reflected by variation of indicators, constructs rule of reasonability judgment to credit evaluation indicators system. The percentage of the variation ofdata of finally established indicators system's original data, to the variation of extensive indicators set's original data, is the information contribution of the established indictors system. An indicators system is considered as reasonable if the final established indicators system is able to contribute more than 80% of original information by using less than 30% of indicators in the extensive indicators set (Chi and Wang 2011) [12].

(2) Method of reasonability judgment on indicators system. Let S be the covariance matrix of indicators' data; trS be the trace of the covariance matrix; s be number of indicators in the final established indicators system; h be number of extensive indicators, and the information contribution of established indicators system toward the extensive indicators set In could be calculated [12] as:

$$In=trS_s/trS_h. \tag{10}$$

Empirical Study

Source of sample and data Standardization. This paper extracts data of small enterprises loans' from one of urban commercial banks in China. Samples with principles or interests that are not settled for more than 90 days (not include 90 days) post due date is treated as default. In the data base of the bank, from year

2004 to 2012, there are 3111 small enterprise loans have been settled, among which 2841 are default samples and 270 are non-default samples.

Standardisze the positive indicators, nagetive indicators and Ideal indicators according to the Equation (1), (2), (3). Standardize qualitative indicators according to the rules set out in Table 2. List the results in Table 3, column 3116 to 6226.

First, selection indicators based on logistic regression significant discrimination method. Process standardized data (in Table 3, column 3116-6226) by SPSS statistics software. Compute the j^{th} indicator's estimated coefficient of logistic regression $\hat{\beta}_j$ and its standard deviation $S_{\hat{\beta}_i}$, Wald test statistics W_j and W_j's test probability *sig*, via "Binary logistic regression analysis" in the SPSS statistics software. Indicators which significantly effect small enterprises' loan default status are listed in Table 4, column 1, and their corresponding value of $\hat{\beta}_i$ 、 $S_{\hat{\beta}_i}$ 、 W_i and *sig*, are listed in Table 4, column 2 to 5. As it is mentioned in part 2.3 of the paper, the indicators selection rule of logistic regression significant discrimination is: let the significance level α=0.05 (listed in Table 4, column 6. indicators with $sig \leq \alpha$, thus $\beta_j \neq 0$, should be retained, as it has significant effect on small enterprises' loan default status; while indicators with $sig \geq \alpha$, thus $\beta_j = 0$, should be removed as it does not have significant affect on small enterprises' loan default status.

For each of indicator, compare the value in Table 4 column 5 with column 6. 38 indicators such as "$x_{1,1}$ Current ratio" with $sig \leq \alpha$ (0.05), should be retained; and 43 indicators such as "$x_{1,21}$ Total assets cash recovery rate" with $sig \geq \alpha$ (0.05), should be removed. Those 38 indicators retained are labeled as "Retained" and those 43 indicators removed are labeled as "Removed by logistic regression significant analysis", in Table 1, column 3 and 5.

The second indicators selection based on R clustering analysis. (1)The steps of R clustering analysis. After the first step "logistic regression significant discrimination based indictors' selection", there are 20 indicators retained in principle layer "x_1 Internal financial factors". Conduct R clustering analysis on indicators' data in each of the principle layer via SPSS statistics software by taking R clustering analysis on the principle layer x_1 as an example. Substitute all of the 20 indicators' data in principle layer x_1 (in Table3, line 1-20, column 3116-6226) into Equation (6)-(7), via R clustering analysis (by "sum of square deviation") in SPSS statistics software, classify principle layer x_1 into 9 classes, name of each class as 1, 2…9 in Table5, column 4. Likewise, classify the principle layer x_2 into 5 classes and principle layer x_3 into 4 classes. Principle layer x_4 only has one indicator being one class, and with no need of clusteringing analysis.

Table 3 Original and standardized data of small enterprises credit evaluation indicators

(1)	(2)	(3)	(4)	Original data v_{ij}	Standardized

No.	Principle Layer	Names of Indicators	Indicator Type	(5) Sample 1	…	(3115) Sample 3111	(3116) Sampl e 1	..	(6226) Sampl e
1		$x_{1,1}$ Current	Positive	1.312	…	1.102	0.077	..	0.065
…		…	…	…	…	…	…	..	…
5		$x_{1,5}$ Net profit	Positive	2424789.0	…	3.865	0.124	..	0.760
…		…	…	…	…	…	…	..	…
9	x_1 Internal financial factors	$x_{1,9}$ Cash turnover cycle	Positive	-2.546		7.500	0.502		0.505
…		…	…	…	…	…	…	..	…
20		$x_{1,20}$ Retained earnings	Positive	1.809		1.251	0.522		0.528
…		…	…	…	…	…	…	..	…
25		$x_{1,25}$ Debt to assets ration	Negative	0.334	…	0.603	0.657		0.369
…	…	…	…	…	…	…	…	..	…
76	x_3 External macro-economic factors	$x_{3,3}$ Consumer price index	Ideal interval [101,104	100.600	…	104.900	0.976	..	1.000
…	…	…	…	…	…	…	…	..	…
81	x_4 Pledge status	$x_{4,1}$ Mortgage,	Qualitati ve	0.350	…	0.570	0.350	..	0.570

Table 4 Significance test on credit evaluation indicators based on logistic regression

(1)Name of indicators	(2) $\hat{\beta}_i$	(3) $s_{\hat{\beta}_i}$	(4) w_i	(5)*sig*	(6)Significance level α
$x_{1,1}$ Current ratio	-3.968	1.723	5.303	0.021	0.050
$x_{1,2}$ Capital immobilized ratio	-2.582	1.073	5.791	0.016	
…	…	…	…	…	
$x_{2,1}$ Years of experiences in	-3.361	0.621	29.613	0.000	
$x_{2,2}$ Patent status	2.419	1.119	4.675	0.031	
…	…	…	…	…	
$x_{3,1}$ Industry prosperity index	-8.889	2.022	19.333	0.000	
$x_{3,2}$ Urban and rural residents	21.633	5.618	14.829	0.000	
…	…	…	…	…	
$x_{4,1}$ Mortgage, pledge and	-2.407	0.794	9.196	0.002	

Table 5 Indicators selection based on R clustering analysis

(1) No.	(2) Principle layers	(3) Name of indicators	(4) Name of Classes	(5) coefficient of variation	(6) Result of selection
1	x_1 Internal financial factors	$x_{1,9}$ Cash turnover cycle	1	0.452	**Retained**
2		$x_{1,10}$ Net operating activities cash	1	0.212	Removed
3		$x_{1,12}$ Net operating activities cash	1	0.223	Removed
4		$x_{1,19}$ Capital accumulation rate	1	0.099	Removed
5		$x_{1,20}$ Retained earnings growth rate	1	0.241	Removed
…		…	…	…	…
19		$x_{1,7}$ Total assets turnover ratio	8	1.057	**Retained**
20		$x_{1,8}$ Accounts payable turnover rate	9	0.442	**Retained**
21	x_2 Internal non-financial factors	$x_{2,1}$ Years of experiences in related	1	0.443	**Retained**
…		…	…		**Retained**
32		$x_{2,5}$ Living condition	5	0.749	**Retained**
33	x_3 External macro economic factors	$x_{3,1}$ Industry prosperity index	1	0.175	**Retained**
34		$x_{3,4}$ Engel coefficient	1	0.096	Removed
…		…	…		…
37		$x_{3,3}$ Consumer price index	3	0.044	**Retained**
38	x_4 Pledge status	$x_{4,1}$ Mortgage, pledge and guarantee	1	0.581	**Retained**

(2) Indicators selection reflecting maximum information content. According to the theory mentioned in above, the bigger an indicator's coefficient of variation is, the more information content it is able to reflect. Therefore, retain the one with the largest coefficient of variation value in each of the class. Calculate each indicator's coefficient of variation by taking "$x_{1,9}$ cash turnover cycle" as an example. Substitute standardized data of indicator $x_{1,9}$ (in Table 3, line 9, column 3116-6226) into Equation (9) and calculate $\overline{X}_i$, the average value of $x_{1,9}$, is 0.424. Substitute the same data and $\overline{X}_i$ =0.424 into Equation (8), the coefficient of variation value of $x_{1,9}$ could be calculated as:

$$v_j = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\overline{x}_j\right)^2}}{\overline{x}_j} = \frac{\sqrt{\frac{1}{3111}\sum_{i=1}^{3111}\left(x_{ij}-0.424\right)^2}}{0.424} = 0.192/0.424 = 0.452,$$

List the results in Table5, line1, column 5. Likewise, calculate coefficient of variations of indicators in Table 5, line 2-4 using corresponding data in Table 3, column 3116-6226.

Sorting the indicators (listed in Table 5), according to their coefficient of variation value. Among those 5 indicators, "$x_{1,9}$ Cash turnover cycle" with the biggest coefficient of variation 0.452 should be retained, and the remaining indicators should be removed. Likewise, calculate the coefficient of variation for each of the indicators in Table5. Sort each class of indicators according to their coefficient of variation values, and retained the indicator with the biggest

coefficient of variation in the class, and remove the remaining indicators in each class.

The result of R clustering shows that, among those 38 indicators retained in the first selection, 20 of which reflect duplicate information, should be removed (label as "Removed" in Table 5, column 6 and Table 1, column 3 and 5) and 18 indicators with the maximum information contents are retained (label as "Retained" in Table 5, column 6 and Table 1, column 3 and 5).

Reasonability judgment on credit evaluation indicators system. Substitute corresponding data (in Table3, column 5-3115) of 18 indicators which listed in Table6, column 3 into the numerator of Equation (10) or and all the indicator data (in Table3, column 5-3115) into the denominator of Equation (10), the information contribution of established indicators system toward the extensive indicators set In could be calculated as: $In=trS_s/trS_h=3.1423\times10^{18}/3.6230\times10^{18}=83.47\%$. Percentage of retained 18 indicators to 81 extensive indicators is 18/81=22.22%. Therefore, paper establishes a credit evaluation indicators system, able to reflect 83.47% of original information with 22.22% of indicators.

Logistic regression–R clustering analysis based small enterprises credit evaluation indicators system. By the application of logistic regression significant discrimination and R clustering analysis model, the paper establishes a small enterprises credit evaluation indicators system, consists of 4 principle layers, such as "x_1 Internal financial factors" and 18 indicators such as "$x_{1,1}$ Current ratio" as listed in Table 6, column 3.

Table 6 Small enterprises credit evaluation indicators system

(1) No.	(2) Principle Layers	(3) Name of indicators
1	x_1 Internal financial factors	$x_{1,1}$ Current ratio
2		$x_{1,2}$ Capital immobilized ratio
3		$x_{1,3}$ Cash ratio
4		$x_{1,4}$ Operating margins
5		$x_{1,5}$ Net profit
6		$x_{1,6}$ Inventory turnover ratio
7		$x_{1,7}$ Total assets turnover ratio
8		$x_{1,8}$ Accounts payable turnover rate
9		$x_{1,9}$ Cash turnover cycle
10	x_2 Internal non-financial factors	$x_{2,1}$ Years of experiences in related industry
11		$x_{2,2}$ Patent status
12		$x_{2,3}$ Enterprises set up a date
13		$x_{2,4}$ Product sales range
14		$x_{2,5}$ Living condition
15	x_3 External macro	$x_{3,1}$ Industry prosperity index

16	economic factors	$x_{3,2}$ Urban and rural residents per capita savings balance
17		$x_{3,3}$ Consumer price index
18	x_4 Pledge status	$x_{4,1}$ Mortgage, pledge and guarantee status

Conclusion

The paper implements logistic regression significant discrimination to remove indicators that can not significantly distinguish between default and non-default samples, and ensuring indicators retained could effectively discriminate default samples from non-default ones. Furthermore, it classifies the indicators by R clustering analysis, selects indicators embodied the largest information from each of the classes according to their coefficient of variation value. It ensures that those indicators reflect duplicate information are removed and also prevent the credit evaluation system from reflecting redundant information. Further more, it makes empirical study on data-set of 3111 small enterprises' information from an urban commercial bank in China. The result of the empirical study shows that the credit evaluation indicators system established by this paper is able to reflect 83.47% of original information with 22.22% of indicators.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (the grant number 71471027, 71171031), the Ministry of Education of China as Science and Technology Research Project (the grant number 2011-10), the China Banking Regulatory Commission as a Risk Management Project of Banking Information Technology (the grant number 2012-4-005) and the Bank of Dalian as Credit Rating and Loan Pricing Systems for Small Business (the grant number 2012-01). We would like to thank the organizations mentioned above.

References

[1] J. G. Wu and C.X. Liang. Empirical analysis on factors of small and medium enterprise financial difficulties [J]. Friends of Accounting, 2013(03): 64-67.

[2] B. F. Shi, G. T. Chi. A credit risk evaluation index screening model of petty loans for small private business and its application [J]. Advances in Information Sciences and Service Sciences, 2013, 5(7): 1116-1124.

[3] Postal Savings Bank of China. Postal savings bank of China form of merchant credit rating [R]. Postal Savings Bank of China, 2009.

[4] B. F. Shi, G. T. Chi. A Model for recognizing key factors and applications thereof to Engineering [J]. Mathematical Problems in Engineering, Vol. 2014, Article ID 862132: 1-9.

[5] L.P. Hai, B. F. Shi, G. R. Peng. A credit risk evaluation index system establishment of petty loans for farmers based on correlation analysis and significant discriminant [J]. Journal of Software, 2013, 8(9): 2344-2351.

[6] L. Lugovskaya. Predicting default of Russian SMEs on the basis of financial and non-financial variables [J]. Journal of Financial Services Marketing, 2010, 14(4): 301-313.

[7] X. Y. Li, S. L Chen. Evaluation index optimization based on neural network in post-appraisal of flood control planning[J]. South-to-North Water Transfers and Water Science & Technology, 2008, 6(3):112-114.

[8] J. L. Xie, P. Can. Theoretical construction and empirical test on evaluation index system of regional SMEs' technological innovation capability [J]. Technology Economics, 2012,31(3):32-37.

[9] L. B. Zhou, G. Li, G. T. Chi. R clustering analysis-coefficient of variation based comprehensive human development evaluation indicators system [J]. System Engineering, 2010, 28(12):56-63.

[10] Kyoung-jae Kim, Hyunchul Ahn. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach [J]. Computers & Operations Research, 2012, 39(8):1800-1811.

[11] G. T. Chi and W. Wang, The comprehensive evaluation method and application of theory based on the scientific development, Science Press, Beijing, 2009.

[12] M. H. Liu, L. Yu, X. L. Zhang, S. X. Guo. Cumulative logistic regression-based measurement models of road traffic congestion intensity [J]. Journal of BeiJing Jiaotong Unvesity, 2008, 32(6):52-56.