International Journal of Economics and Finance; Vol. 12, No. 6; 2020
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

A Simplified Variable Analysis of Credit Ratings for Small Chinese Enterprises Based on Support Vector Machine

Ying Chen[1], Yangkai Guo[1] & Maoguo Wu[1]

[1] SILC Business School, Shanghai University, Shanghai, China

Correspondence: Yangkai Guo, 20 Chengzhong Road, Jiading District, SILC Business School, Shanghai University, Shanghai 201899, China.

Received: March 23, 2020 Accepted: April 20, 2020 Online Published: May 10, 2020

doi:10.5539/ijef.v12n6p45 URL: https://doi.org/10.5539/ijef.v12n6p45

Abstract

Small enterprises are an important component of the national economy and valuable customers of commercial banks. Commercial banks use credit ratings, including financial and nonfinancial indices, to analyze small enterprises before committing to long-term collaborations, including loans. This paper uses a support vector machine algorithm to establish an imbalanced multi-classification model and compares the results to those of other methods. Commercial banks need simplified variable analysis credit ratings that use minimal information to rapidly and accurately obtain credit ratings and improve the efficiency of the process. Accordingly, we perform multiple tests of simplified rating systems using fewer variables.

Keywords: small enterprises, support vector machine, imbalanced multi-classification, credit rating

1. Introduction

The Chinese government has always attached great importance to the development of small enterprises and constantly builds up the financial ecological environment of small enterprises to further their development. However, it has always been very difficult for small enterprises to obtain financial support. In July 2014, Premier Li Keqiang repeatedly mentioned questions related to reducing the cost of financing for enterprises, especially small enterprises, at the state council executive meeting. The limited sources of funds and the lack of continuous financial support have become a bottleneck in the sustainable development of small enterprises. Therefore, it is urgent to solve the financing problem of small enterprises by facilitating an effective financial support system to promote their development.

At the same time, commercial banks, as money suppliers, are also under significant pressure. After a few years of rapid development, Chinese commercial banks and the financial system, in general, have established a good financial foundation for financing enterprises. Under the macro-background of the liberalization of interest rates and financial innovations, which are increasingly advanced by the Internet, market competition between commercial banks has become increasingly fierce with increases in the cost of deposits and the interest rates on loans. Therefore, this has resulted in the compression of the profits from the interest balance between deposits and loans. With fierce competition in the credit market, commercial banks need to obtain customer resources to increase their market share, especially in the form of small enterprises. Small enterprises have therefore become important customers for commercial banks. The credit ratings for small enterprises are key to solving the problem of financing small enterprises. The practical problem is finding an efficient and accurate method for obtaining a credit rating for new customers. This includes how to effectively identify and analyze the performance of small enterprises and how to eliminate small enterprises with poor performance.

With the increasing demand from commercial banks to identify promising small enterprises, credit managers need help in quickly and efficiently analyzing the performance of enterprises. Decisions concerning loan applications are required before small enterprises and commercial banks can establish long-term collaborations. Commercial banks select small enterprises according to various types of information. As a result, credit managers require an effective, scientific method of determining the credit rating, especially for small enterprises. Therefore, it is necessary to use different types of methods, including machine-learning algorithms based on data mining, to make decisions.

2. Related Literature

2.1 Credit Ratings for Enterprises

There are two categories of credit rating methods: quantitative and qualitative. Qualitative evaluation methods are called artificial expert analyses and are also known as classic credit analysis methods. At present, Chinese commercial banks still primarily use this method. However, a few quantitative credit rating methods have also been used. Initially, Altman (1977) used multiple discriminated analysis, and Zhihui and Meng (2005) and Zhang (2010) used logistic model analysis. Credit metrics were used by J.P. Morgan in the United States in 1997 (Morgan, 1997) This is a value-at-risk model that estimated the risk value of loans and other assets. McKinsey & Company designed the Credit Portfolio View model (McKinsey & Company, 1998), which is based on credit metrics. Their model added factors from the macro-economic cycle and established a relationship between macro-economic indicators, such as the economic growth rate, interest rate, and government expenditures, and the transition matrix of the credit rating. In addition, this model uses the Monte Carlo method to simulate changes in the transitional probability of the rating with the degree of cyclical factors. The Credit Monitor model, developed by KMV Ltd. in the United States (Chen, 2014), estimates the probability of loan defaults. The Credit Risk, issued by the financial products development department of the Swiss Credit Bank (Basel Committee on Banking Supervision, 2001), calculates the probability of defaults. There are many other artificial intelligent methods used for credit ratings, such as integer programming, artificial neural network, genetic algorithm, and support vector machine algorithm.

2.2 Support Vector Machine

The support vector machine method was first proposed by Cortes and Vapnikin in 1995. It is easily combined with other methods and has subsequently been popularized. Because it has the advantage of solving problems using nonlinear, high-dimensional classification and regression with relatively high accuracy, it is widely used in the fields of disease diagnosis, handwritten font and text recognition, face recognition and image retrieval, analysis and application in engineering technology, and evaluation and prediction in economic and management fields. Qifeng et al. (2005) selected more than 1000 sample data points for enterprises in the light industry from a certain commercial bank in 2003. The data included the ratios of the debt payment, profitability, operational management, and the output results of Grades AAA, AA, A, and A−. Empirical studies using support vector machine achieved an overall test accuracy of 83.15% with a faster learning speed than that of the neutral network method. This formed a suitable credit rating method for commercial banks. Zhou et al. (2009) explored how to select the credit score parameters using support vector machine and produced good results with two real-world credit datasets. Kim et al. (2012) compared support vector machine and other artificial intelligence methods for multi-class problems and obtained an improved performance. Harris (2015) used a clustered support vector machine to solve the binary classification credit score problem and obtained better results compared with that of previous research. Ping-Feng (2015) proposed a new type of decision tree support vector machine that combined rough set theory and support vector machine to solve multi-class problems.

At present, there are a large number of studies that have been conducted based on support vector machine. Existing studies on the credit ratings of enterprises have mostly focused on problems of binary classification and less on the problem of multiple classifications. For binary classification problems, the same amount of sample data is generally chosen in the normal and control groups, and there have been few studies concerned with imbalanced classifications. In credit rating index systems for enterprises, indices are generally selected from financial statements that reflect historical information, and few qualitative indicators are used. Therefore, a support vector machine model for determining the credit ratings of enterprises could be further applied to play a more important role in practice.

3. Problem and Concept Analysis

According to the regulations and the collected sample data, this paper defines small enterprises as having an owner equity of more than 6 million Yuan. In addition, the number of employees is low, approved financial reports by completely audited or third-party agencies cannot be provided, and the applied loan amount is below 30 million Yuan. There needs to be an increase in the qualitative indices combined with the quantitative indices in the credit rating index system for small enterprises. Commercial banks could initially evaluate small enterprises using a decision system, and then, credit managers could analyze and judge whether to give loans to small enterprises using an artificial expert method. During this period, commercial banks would still need to regularly measure the risk exposure of older customers, paying close attention to the development of small enterprises, and reduce the possibility of bad debts as much as possible. Therefore, it is necessary to establish a model based on data mining and a machine-learning algorithm to provide information to the credit manager in the credit

management and risk management department according to the requirements and regulations of the new Basel agreement and China's banking regulatory commission.

In general, there are only a few enterprises that default in each commercial bank's database. According to existing data in the databases of commercial banks, models were selected, which identified possible defaulting customers in existing customers, using the structure of an imbalanced classification problem. An imbalanced classification problem is equivalent to the binary imbalanced classification and multiple imbalanced classification problems.

At present, there are many studies focusing on the binary imbalance classification problem. Multiple imbalanced classification problems generally refer to problems with classification categories. However, there are more than two types of categories, and there are significant differences between the numbers in each group sample, especially because there is only a small amount of sample data in certain individual groups in multiple classification problems. In addition, with many users using many different types of methods, it is difficult to fully learn the characteristics of each group, which leads to a decrease in the accuracy of the classifications. The credit rating problem for enterprises in commercial banks is a typical multiple imbalanced classification problem. First, there are 10 distinguished credit rating grades, namely, AAA, AA, A, BBB, BB, B, CCC, CC, C, D, and some commercial banks even add A+ and A− to make 12 credit rating grades, based on the different characteristics of management and the performance of the enterprises. Second, there are different numbers of enterprises in each sample data category in the commercial bank databases. For example, the vast majority of enterprises are above Grade A, and a few are below Grade BBB, even if there are still large differences between the Grade AAA, AA, and A sample groups. Third, the amount of sample data in a certain category might be zero. Because enterprises with low grades are rejected by commercial banks, there are no sample data for some categories (such as grade D) in the customer databases of commercial banks. In this paper, we use a credit rating system with 10 grades, including grades AAA, AA, A, BBB, BB, B, CCC, CC, C, and D, for a small commercial bank. Table 1 shows the number of enterprises in each grade for 164 enterprises in the customer database of the commercial bank.

Table 1. The number of enterprises in each grade for the 164 enterprises in the database

Grade	AAA	AA	A	BBB	BB	B	CCC	CC	C	D	Total
Number of enterprises	6	25	49	46	20	16	2	-	-	-	164

It can be seen from Table 1 that there are only seven grades with data and there are no sample data below Grade CCC. Classification results for 7–10 grades might exist for different years; this is different from other multiple imbalance and binary classification problems in which the quantities of the sample data are nearly the same. A basic machine-learning algorithm model cannot effectively learn the characteristics of the information from this type of sample.

Existing studies indicate that there are three methods for solving an imbalanced classification problem: (1) improvements in the algorithm, (2) improvements in the data sampling technique, and (3) the simultaneous improvement of both the algorithm and the data sampling. Algorithm improvements could include changing the inherent characteristics and the original treatment principle of the algorithm, which would allow the calculation and analysis of the model to adapt to the requirements of the problem. Techniques for improving the data sampling focus on the selection methods for the data and can be used independently. Over-sampling increases the number of sample data in the minority grades, and under-sampling decreases the number of sample data in the majority grades. A hybrid algorithm combines the sampling and algorithm techniques.

When training sample data sets, more attention should be paid to the minority samples, and the data characteristics of minority samples should be analyzed. Support Vector Machine (SVM) based on Statistical Learning Theory (SLT) has superiority in solving classification and regression problems. Therefore, it can improve the basic algorithm of SVM to achieve the characteristics of machine learning algorithms for solving different classification and regression problems.

4. Methodology

According to the book "The theory and algorithms of support vector machines" written by Deng et al. (2005) and Shen (2004), the support vector machine method mostly solves the problems of regression and classification, which include linearly separable problems and linearly inseparable problems.

For linearly separable problems, the training sample dataset in the binary classification problems is $x_i \in R^n$, $i = 1, 2, ..., n$, and the classification of the corresponding level is $y_i \in \{-1, 1\}$, i=1, 2. The classified hyperplane $(w \cdot x_i) + b = 0$, where $w \cdot x_i$ is the dot product of w and x_i. Two types of sample data, both satisfying the

constraint $y_i[(w \cdot x_i) + b] \geq 1$, i = 1, 2, … , n and the classified margin equaling to $\frac{2}{\|w\|}$, are generated. Under the constraint of $y_i[(w \cdot x_i) + b] \geq 1$, i = 1, 2, … , n, the objective function maximizes $\frac{2}{\|w\|}$. Maximizing the classified margin is the same as minimizing $\frac{\|w\|^2}{2}$. The optimal classified hyperplane can divide the sample data and minimize $\frac{\|w\|^2}{2}$. Support vectors on the hyperplane contribute to the optimal hyperplane and the decision functions. Therefore, it is unnecessary to require all training data to be on $y_i[(w \cdot x_i) + b] \geq 1$, i = 1, 2, … , n, and the constraint conditions can be relaxed to $y_i[(w \cdot x_i) + b] \geq 1 - \xi_i$, i=1, 2, …, n, where $\xi_i \geq 0$, i = 1, 2, … , n. $\sum \xi_i$ describes the degree to which the training set is incorrectly distinguished with incorrect data. The penalty parameter C is an adjustable parameter greater than 0; a large C indicates a punishment for faulty classifications. This is a quadratic programming problem, and the following equations are used to solve the optimization problem:

$$\min_{w,b}\left(\tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right) \tag{1}$$

$$s.t.\ \ y_i\left[(w \cdot x_i) + b\right] \geq 1 - \xi_i,\ \ i = 1,\ 2, \ldots,\ n \tag{2}$$

$$\xi_i \geq 0,\ \ i = 1,\ 2, \ldots,\ n \tag{3}$$

Consequently, the parameter C is used to balance the training accuracy and the generalization ability. ξ_i indicates the slack variables used to solve the problem over a larger feasible region, w $\in$ R^n is a weight vector to explain the location of the separating hyperplane in each space, and b is the position error of the mobile hyperplane. Because this is a quadratic programming problem, the optimal solution is the following Lagrange function of saddle points:

$$L(w, b, \alpha) = \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i(w \cdot x + b) + \xi_i - 1\right] - \sum_{i=1}^{n}\beta_i\xi_i \tag{4}$$

where $\alpha_i \geq 0$ and $\beta_i \geq 0$ are the Lagrange multipliers. At the saddle point, the gradients of w, b, and ξ are zero; therefore,

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} a_i y_i x_i = 0 \Rightarrow\ w = \sum_{i=1}^{n} a_i y_i x_i \tag{5}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{6}$$

$$\frac{\partial L}{\partial \xi} = C - \alpha_i - \beta_i = 0 \tag{7}$$

Inserting Eqs. (5)–(7) into Eq. (4) and calculating the maximum of Eq. (4) on α, the dual optimization problem of Eqs. (1)–(3) are obtained as shown below:

$$\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j)\right] \tag{8}$$

$$s.t.\ \ \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{9}$$

$$0 \leq \alpha_i \leq C \qquad i = 1,\ 2, \ldots, n \tag{10}$$

To solve the above equations, α needs to satisfy $\alpha_i = 0$, $0 < \alpha < C$ and $\alpha = C$. If $0 < \alpha < C$ and $\alpha = C$, the corresponding x_i is the support vector. In the support vector machine method, the corresponding x_i of $\alpha = C$ is on the border and is known as the bound support vector. In addition, the corresponding x_i of $0 < \alpha < C$ is in the interval and is known as the normal support vector. According to the Karush–Kuhn–Tucker (KKT) conditions, at the optimum points, the Lagrange multiplier and the constraint conditions both equal 0:

$$\alpha_i \left\{ y_i \left[(w \cdot x) + b \right] - 1 + \xi_i \right\} = 0, \quad i = 1, 2, ...n \tag{11}$$

$$\beta_i \xi_i = 0, i = 1, 2, ...n, \quad i = 1, 2, ...n \tag{12}$$

For a normal support vector $(0 < \alpha < C)$, it is known that $\beta_i > 0$ from Eq. (7) and Eq. (12) and $\xi_i = 1$. However, for any normal support vector,

$$y_i \left[(w \cdot x) + b \right] = 1 \tag{13}$$

Therefore, b is

$$b = y_i - (w \cdot x) = y_i - \sum_{x_j \in J} \alpha_j y_j (x_i \cdot x_j), x \in JN \tag{14}$$

Here, JN is the set of normal support vectors, and J is the set of support vectors. The constraints of Eqs. (2) and (3) limit w and b and make the empirical risk of error equal to 0. At the same time, they minimize $\|w\|$ to minimize the VC dimension. Therefore, the optimization of Eq. (1) embodies the principle of structural risk minimization and has a good generalization ability. This method could therefore solve linearly separable problems very well.

However, linearly inseparable problems indicate that using any straight line would incorrectly distinguish large amounts of data in the training set. For linearly inseparable problems, support vector machine selects a kernel function K, which is used on the sample data to map the dataset to a high-dimensional data space, transforming the linearly inseparable problem into a linearly separable problem and constructing an optimal hyperplane separating the points of the difficult nonlinear data. Different kernel functions obtain different classifiers, and the parameters used in the selection of the kernel function are very important. Under this condition, Eq. (8) changes to the following form:

$$\max_{\alpha} [\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)] \tag{15}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{16}$$

$$0 \le \alpha_i \le C, i = 1, 2, ...n \tag{17}$$

Here, $K(x_i, x_j) = [\phi(x_i) \cdot \phi(x_j)]$ is a kernel function that solves the dual problem to determine the final decision function:

$$f(x) = \text{sgn} \left\{ \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right\} \tag{18}$$

If the kernel function $K(x_i, x)$ is appropriately selected, the linearly inseparable problem in input space can be transformed into a linearly separable problem in feature space. There were many different kernel functions that can be used in a support vector machine model. In this paper, the Gaussian radial basis function was used as the kernel function.

$$K(x, y) = \exp(-\gamma \|x - y\|^2) \quad \gamma > 0 \tag{19}$$

The sample data were mapped to a high-dimensional space using the kernel function, which was then used to solve the problem using the nonlinear relationship between the class labels and the characteristics of the data, as well as the problem of having an insufficient number of prior experiences. γ is an inherent parameter of the function that maps data to the distribution in the new feature space. C is the penalty parameter, which indicates that there are fewer errors in the support vector machine classification model when its value increases. The choice of parameters without prior knowledge was achieved via a grid search method, which is a common method used to set parameters.

The process in this paper, using an integrated learning algorithm to improve the support vector machine method, primarily included three steps: segmentation, training, and aggregation. In our sample datasets, the positive subsets were the good credit ratings for small enterprises, such as Grades AAA, AA, A, BBB, BB, B, CCC, CC, C, and D. These enterprises do not default and make up a large percentage of the sample datasets. The poorer credit ratings of small enterprises were negative subsets, and the minority dataset in the customer database, such as Grades C or D, could not easily be predicted by machine-learning methods with inadequate characteristics. The first step in the algorithm is segmentation, which reclassifies the existing sample groups to achieve nearly balanced groups. There are less data in the negative groups; therefore, they did not need to be further segmented.

Conversely, the data samples in the positive groups required detailed segmentation and classification. The sample groups could be divided into k ($k > 3$) subclasses; for example, the classification result of customer information after data cleaning is multiple subcategories. The second step is training, which consolidates the sample data in the negative classes. For example, there were only two data points for grade 7 in the sample data, and these were the negative classes. If traditional methods were directly used to eliminate negative classes, there would be no data. As a result, the sample data for the negative classes were retained, and the sample data for the positive categories were the key points that could be classified in detail. Support vector machine was used to classify the sample data after segmentation. The third step is aggregation. After training the sample data, it is necessary to integrate each individual class to form a suitable method for this type of classified problem to distinguish all different classes according to the distance between each feature vector of the sample data for each class. First, the sample data were separated into negative classes, and then, the sample data were separated into positive categories. In addition, new test data were classified into appropriate classes via the support vector machine method.

4.1 Input

The known training set was a small enterprise sample data $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where x represents the information of the different characteristics of each small enterprise and y represents the corresponding grade of the small enterprise. The positive categories in the training set, P, were the good customer datasets for the credit rating, and the negative categories in the training set, N, were the poor customer datasets for the credit rating (the sample size of P was m_1, the sample size of N was m_2, $m_1 + m_2 = L$, and $m_1 \geq m_2$). M is the number of categories in the positive sample dataset.

The positive categories of the sample dataset, P, in the algorithm could be divided into M (in this case M = 6) data subsets with V_i $(i = 1,2, \ldots, M)$. Simultaneously, the support vector C_i $(i = 1,2, \ldots, M)$ could be obtained separately to separate each sample dataset. Amalgamating the negative categories of the N sample dataset, each subset is represented by V_i. The decision hyperplane $g_i(x)$ was used to determine the formula in support vector machine and to obtain the optimal solution $D_i = [V_i, N]$. d_i indicates the distances (c_i, c_0) between the hyperplane separating each category of space and was used to complete the calculation. Here, $v_i = exp\left(\frac{d_i}{d_m}\right)$, $i = 1,2, \ldots, M$, and d_m was the median of all distances between the hyperplane.

4.2 Output

$$F(x) = y = \mathrm{sgn}\left[\sum_{i=1}^{M} v_i g_i(x)\right] \tag{20}$$

Where sgn(x) is a sign function. The final output is the specific category y that corresponds to the arbitrary input of a sample vector x. In the learning process, it is important to pay attention to the choice of M, which is not only the number of categories in the positive data sample set but also the number of classifiers in the integrated learning algorithm. Because the size and distribution of the data affect the efficiency and accuracy of the classification results of support vector machine classifiers according to the actual sample data available each year, given the value of M for classification, it is estimated that M can be used between 6 and 11.

5. Data

The support vector machine method does not require the sample data to consist of a normal distribution and correlation tests. We collected sample data for small enterprises from the customer database of a city commercial bank in Zhejiang Province, China, in the financial year of 2017. Table 2 is a descriptive statistical analysis of the 164 enterprises and the 17 variables.

Table 2. A descriptive statistical analysis of the original sample data

Variables	N	Minimum	Maximum	Mean	Standard deviation	Variance
Working years of the manager	164	1	35	11.15	6.124	37.500
Educational background of the manager	164	0	7	2.19	1.965	3.860
Corporate lifetime	164	0	21	7.18	4.149	17.214
Investors' assets	164	2	5	4.67	0.768	0.590
Sales output ratio	164	0	3	2.34	0.840	0.705
Debt ratio	164	0.066	0.770	0.432	0.136	0.019

Owner's equity	164	7925567	590875849	68910686	65072022	$4.234*10^{15}$
Current ratio	164	0.650	6.980	1.631	0.903	0.815
Accounts receivable turnover	164	2.080	127.218	10.39	12.037	144.889
Sales growth rate	164	−0.920	14.040	0.334	1.176	1.384
Profit growth rate	164	−4.060	130.260	1.210	10.347	107.062
Return on equity	164	−0.060	38.650	0.475	3.015	9.090
Personal credit record of the manager	164	1	1	1	0	0
Industry policy	164	0	1	0.71	0.456	0.208
Local environment	164	0	1	0.77	0.423	0.179
Operating site conditions	164	0	1	0.96	0.188	0.035
Equipment utilization	164	0	1	0.87	0.335	0.112

According to the distribution characteristics of the original data, there were very few sample data that were less than zero. Most were greater than zero and uniformly distributed. To reduce collinearity between the different variables, sample data are mapped to [0, 1], which was an extreme linear model of processing in treating the sample data. The more the order moves from small to large and the array of the large data, the better the data.

Complete credit rating methods for enterprises contain a credit rating index system. Small enterprises in this paper refer to enterprises that have an owner equity above 6 million Yuan. However, the numbers of employees are low, and complete financial statements audited or recognized by third-party agencies cannot be provided. The credit rating index system for small enterprises includes quantitative and qualitative indices. To obtain more accurate results, we drew on the experience of state-owned commercial banks in China and had many discussions with experts. After these discussions, the index system was redesigned as shown in Table 3.

Table 3. The credit rating index system for small enterprises

Variables	Definitions	Description (marks)
Working years of the manager	The experience of the manager at the small enterprise.	Above 6 years: 100; 3–6 years: 50; Below 3 years: 0
Educational background of the manager	Primary school, middle school, diploma, bachelor degree, master degree.	Above bachelor degree: 100; Diploma: 50; Below diploma: 0
Corporate lifetime	The number of years the small enterprise has been operating.	More than 5 years: 100; 2–5 years: 50; Less than 2 years: 0
Investors' assets (Yuan)	The individual property of the investors.	Above 6 million: 100; 5–6 million: 80; 4–5 million: 60; 3–4 million: 40; 2–3 million: 20; Below 2 million: 0
Sales output ratio	Sales units/produced units	Above 90%: 100; 80%–90%: 50; Below 80%: 0
Debt ratio	Total liabilities/total assets	0%: 100; 0%–50%: 50; 50%–100%: 0
Owner's equity (Yuan)	Owner's equity in small enterprises	The sample data are all more than 6 million: 100
Current ratio	Current assets/current liabilities	Above 1.2: 100; 1.1–1.2: 75; 1–1.1: 50; 0.9–1: 25; Below 0.9: 0
Accounts receivable turnover	Net sales/net accounts receivables	Above 4: 100; 2–4: 50; Below 2: 0
Sales growth rate	(Net sales in this year- net sales in last year)/ net sales in last year * 100%	Above 20%: 100; 15%–20%: 75; 10%–15%: 50; 5%–10%: 25; Below 5%: 0
Profit growth rate	(Total profits in this year- total profits in last year) total profits in last year * 100%	Above 20%: 100; 15%–20%: 75; 10%–15%: 50; 5%–10%: 25; Below 5%: 0
Return on equity	Total profits/owners' equity	Above 10%: 100; 8%–10%: 75; 6%–8%: 50; 4%–6%: 25; Below 4%
Personal credit record of the manager	Personal credit records of managers at local banks	The sample data are all good: 100
Industry policy	Industry policy in the local area	Good: 100; Normal: 50; Limited: 0
Local environment	Environment policy in the local area	Good: 100; Normal: 50; Limited: 0
Operating site conditions	Operating site: owned or leased	Self-owned: 100; Leased: 0
Equipment utilization	The ratio of operating equipment	High: 100; Medium: 50; Low: 0

6. Empirical Results

The result of the experiment was the predicted precision ratio:

$$The\ accuracy\ rate\ of\ precision = \frac{the\ accurate\ numbers\ of\ sample\ data}{the\ numbers\ of\ all\ sample\ data} \tag{21}$$

For example, C was selected in the range of 100–10,000 for the experiment and was increased by 10^n. γ was selected from a range with an increasing speed of 10^{-n}. According to the results of the convergence and the accuracy of the precision, a gradual narrowing of the scope should produce a higher classification accuracy. If C was small and the actual testing accuracy was low, then C was gradually increased and approached the optimal value range for support vector machine. Every trial required approximately 10 min. Owing to time and energy limitations, after repeated testing and analyses, parameter combinations in the classification model were ruled out if they would lead to the results being divergent, not convergent. This was found for the following divergent parameters C = (1800, 1900, 2000, 2100, 2200, 2500, 2700) and γ = (0.001, 0.003, 0.005, 0.007, 0.009). The test results in this range were better than those for other parameters. The different combinations of C and γ were composed of several different classification models of the support vector machine model. The average precision accuracy of each classification was measured 20 times. There were 35 different testing results for the analysis, as shown Table 4, after the use of the integrated support vector model.

Table 4. The classified results for 17 variables

γ \ C	1800	1900	2000	2100	2200	2500	2700
0.001	0.7560	0.7560	0.7548	0.7548	0.7524	0.7560	0.7583
0.003	0.7690	0.7726	0.7738	0.7714	0.7726	0.7726	0.7702
0.005	0.7702	0.7643	0.7655	0.7655	0.7655	0.7631	0.7643
0.007	0.7631	0.7631	0.7667	0.7667	0.7607	0.7595	0.7595
0.009	0.7619	0.7607	0.7607	0.7607	0.7595	0.7583	0.7583

Other methods were also used to classify the same sample data of the 164 enterprises to compare to the results of the support vector machine classification method. Table 5 shows the results.

Table 5. The comparison results for the model classifications of different methods

Method	Accuracy of classification
Two-step clustering method	Silhouette measure of cohesion and separation indicates that the cluster quality is poor
K mean clustering method	27.8%
System clustering method	43.9%
The radial basis function neural network method	24.5%
The multi-layer perceptron neural network method	29.8%

It can be seen that the accuracy of the experimental results using other methods is low. After interviews with the credit managers at commercial banks, the credit managers cooperated with the risk assessment manager to identify the potential risks of loans to given enterprises. In practice, the possible amount of variable data needed to analyze small enterprises rapidly and accurately needs to be as small as possible. Therefore, the credit rating variables were reduced and eliminated to see if ideal results could be obtained. Because there were 17 collected variables, there were numerous different possible combinations of variables, all of which could not be tested. Using factor analysis and principal component analysis, dimensionality reduction was found to be unsuitable for credit rating analyses in commercial banks. The principal components and factors calculated were not stable, and it was difficult to interpret the results. Therefore, we could only delete variables according to a correlation analysis, which was based on changes in the accuracy rate, to determine combinations of variables.

First, 16 variables were selected from the 17 variables of the valid sample data for the 164 small enterprises. After standardization of the sample data, the owner equity variable was equal to 1; therefore, this variable was eliminated. The new index system included 16 variables: the working years of the manager, educational background of the manager, corporate lifetime, investors' assets, sales output ratio, debt ratio, current ratio, accounts receivable turnover, sales growth rate, profit growth rate, return on equity, personal credit record of the manager, industry policy, local environment, operating site conditions, and equipment utilization. The test results are shown in Table 6.

Table 6. Classification results for 16 variables not including owner equity

C / γ	1800	1900	2000	2100	2200	2500	2700
0.001	0.7418	0.7409	0.7409	0.7436	0.7455	0.7482	0.7509
0.003	0.7664	0.7673	0.7718	0.7745	0.7745	0.7727	0.7727
0.005	0.7636	0.7591	0.7591	0.7564	0.7582	0.7527	0.7609
0.007	0.7527	0.7573	0.7573	0.7582	0.7609	0.7627	0.7591
0.009	0.7618	0.7655	0.7673	0.7664	0.7664	0.7682	0.7700

Table 6 indicates that the accuracy rate of the classification with 16 variables is lower than that with 17 variables, showing that support vector machine is more effective for high-dimensional classification problems with higher accuracy.

Next, 16 variables were again selected from the 17 variables of the valid sample data. After standardization, the personal credit records of the manager variables were all good in the sample data, and the standard was 1. As a result, 16 variables were used for testing, including the working years of the manager, educational background of the manager, corporate lifetime, investors' assets, sales output ratio, debt ratio, owner's equity, current ratio, accounts receivable turnover, sales growth rate, profit growth rate, return on equity, industry policy, local environment, operating site conditions, and equipment utilization. The test results are shown in Table 7.

Table 7. Classification results for 16 variables not including personal credit records

C / γ	1800	1900	2000	2100	2200	2500	2700
0.001	0.7727	0.7718	0.7700	0.7691	0.7709	0.7709	0.7709
0.003	0.7782	0.7845	0.7873	0.7891	0.7864	0.7909	0.7873
0.005	0.7855	0.7836	0.7800	0.7773	0.7718	0.7764	0.7736
0.007	0.7755	0.7755	0.7736	0.7764	0.7755	0.7736	0.7718
0.009	0.7718	0.7736	0.7709	0.7709	0.7691	0.7664	0.7636

A third test used eight variables, including the debt ratio, current ratio, sales growth, sales growth rate, return on equity, corporate lifetime, industry policy, investors' assets, and sales output ratio and omitting the working years of the manager, educational background of the manager, owner's equity, accounts receivable turnover, profit growth rate, personal credit record of the manager, local environment, operating site conditions, and equipment utilization. The test results are shown in Table 8.

Table 8. The classification results for the first test using eight variables

C / γ	1800	1900	2000	2100	2200	2500	2700
0.001	0.6298	0.6286	0.6369	0.6381	0.6357	0.6524	0.6607
0.003	0.6976	0.6917	0.6881	0.6905	0.6881	0.6929	0.6929
0.005	0.6964	0.7012	0.7000	0.6976	0.7060	0.7095	0.7095
0.007	0.7107	0.7107	0.7095	0.7190	0.7202	0.7286	0.7310
0.009	0.7262	0.7274	0.7321	0.7298	0.7298	0.7286	0.7286

A second test also used eight variables, including the debt ratio, current ratio, sales growth rate, return on equity, corporate lifetime, industry policy, investors' assets, and sales output ratio and omitting the educational background of the manager, working years of the manager, owner's equity, current ratio, accounts receivable turnover, profit growth rate, return on equity, personal credit records of the manager, industry policy, local environment, operating site conditions, and equipment utilization. The test results are shown Table 9.

Table 9. The classification results for the second test using eight variables

C / γ	1000	1200	1400	1600	1800	2000	2200
0.001	0.6286	0.6190	0.6238	0.6250	0.6298	0.6369	0.6357
0.003	0.6690	0.6869	0.6905	0.6976	0.6976	0.6881	0.6881
0.005	0.6964	0.6917	0.6857	0.6929	0.6964	0.7000	0.7060
0.007	0.6857	0.6940	0.6988	0.7071	0.7107	0.7095	0.7202
0.009	0.6952	0.6988	0.7107	0.7143	0.7262	0.7321	0.7298

Different combinations of the variable parameters, C and γ, together constitute a support vector classification machine. As a result, there were 35 support vector classification machines in one table, all operated multiple times, which built more than a thousand classifiers of the support vector machine. The result of the operations was the average accuracy rate after 20 repetitions, which enhanced the precision and robustness of the classification. From Table 10, it can be seen that the accuracy results with the other methods are low.

Table 10. Comparison results of the classifications

The name of model	The accuracy rate of classification
Two-step clustering method	Silhouette measure of cohesion and separation indicates that the cluster quality is poor
K mean clustering method	35.98%
System clustering method	15.24%
The radial basis function neural network method	31.9%
The multi-layer perceptron neural network method	49%

7. Conclusions

In China, it is necessary for commercial banks to identify the credit ratings of small enterprises. This paper uses a suitable ensemble support vector machine method to analyze sample data from a customer database in a commercial bank. After many tests and analyses, the index system of the variables was gradually adjusted. Because the characteristics of support vector machine are suitable for high-dimensional nonlinear classification problems, more features were included in the variable indices. Therefore, the attained accuracy was higher. The support vector machine method does not require the sample data to have a normal distribution nor does it need correlation tests to solve this type of imbalanced multi-classification problem and to enhance the precision and robustness of the classification.

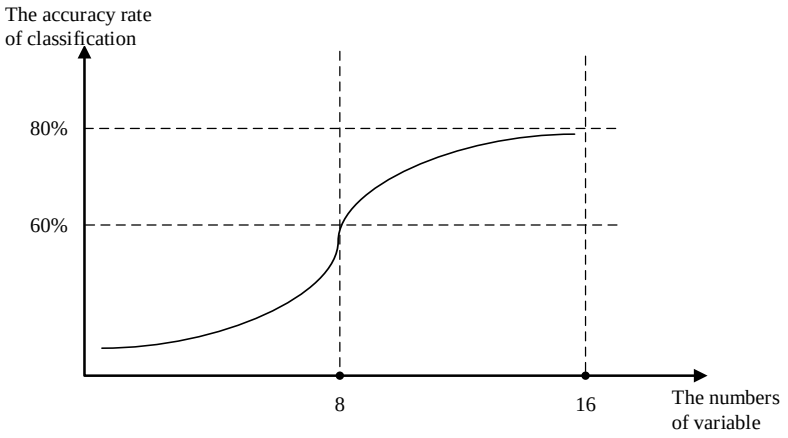

Figure 1. A schematic diagram of the relationship between the number of variables and the accuracy rate of the classification

From Figure 1, it can be seen that there were 15–17 variables in the index system sample data and that the accuracy rate of the classification was close to 80%. Decreasing the number of variables in the index system caused the accuracy rate of the classification to decrease. The accuracy rate for an index system of eight variables was over 62%. For some combinations of parameters, it was above 70%, which is a relatively good evaluation

accuracy. When the number of variables decreased to seven, the accuracy quickly fell below 60% (Figure 1). Therefore, eight variables were selected as the simplified credit rating index system for small enterprises.

We chose to use the following eight variables: the working years of the manager, debt ratio, profit growth rate, sales growth rate, corporate lifetime, industry policy, investors' assets, and sales output ratio. After normalization, the enterprise owner's equity and the personal credit record of the manager were both equal to 1 and were very important evaluation indices. These two variables were not incorporated into the calculation when the classifiers were tested. However, they could not be ignored in the simplified credit rating index system. As a result, the simplified credit rating index system for small enterprises includes 10 variables: the working years of the manager, debt ratio, profit growth rate, sales growth rate, corporate lifetime, industry policy, investors' assets, sales output ratio, enterprise owner's equity, and personal credit record of the manager.

References

Altman, E. I., Haldeman, R., & Narayanan, P. (1997). ZETA analysis: A new model to identify bankruptcy risk of corporations. *J. Bank Financ., l*, 29-54. https://doi.org/10.1016/0378-4266(77)90017-6

Basel Committee on Banking Supervision. (2001). *Consultative document: The new basel capital accord.* Bank for International Settlements.

Carey, M. (2002). A guide to choosing absolute bank capital requirements, Board of Governors of the Federal Reserve System. *International Finance Discussion Papers, no. 726*. https://doi.org/10.17016/IFDP.2002.726

Chen, Y., & Chu, G. L. (2014). Estimation of Default Risk Based on KMV Model—An Empirical Study for Chinese Real Estate Companies. *Journal of Financial Risk Management, 3*, 40-49. https://doi.org/10.4236/jfrm.2014.32005

Cheng-Lung, H., Mu-Chen, C., & Chieh-Jen, W. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl., 33*(4), 847-856. https://doi.org/10.1016/j.eswa.2006.07.007

Chong, Wu., & Hang, X. (2008). Research on customer credit evaluation model based on integrated support vector machine in the electronic commerce environment (in Chinese). *Chinese Journal of Management Science.*

Committee on the Global Financial System. (2001). *A Survey of stress tests and current practice at major financial institutions.* Bank for International Settlements.

Cristianini, N., & Shawe-Taylor, J. (2005). *An introduction to support vector machine and other kernel-based learning method*. China Machine Press, English edition.

Cuihua, S. (2004). *Research on individual credit rating of loans for consumption based on support vector machine method* (in Chinese). Doctoral Dissertation in China agricultural university.

Euro-currency Standing Committee of the central banks of the Group of Ten countries. (1998). *On The Use of Information and Risk Management*. Bank for International Settlements, Basle, 1st October.

Fuyong, W., & Xiufu, S. (2009). Research on financial prediction on credit evaluation method of Listed company based on support vector machine (in Chinese). *China intelligent automation conference proceedings in 2009* (the second volume).

Gieseke, K. (2002). *Credit Risk Modeling and Valuation: An Introduction.* Humboldt-University zoo Berlin, August 19.

Global Credit Research. (2002). Loss Calc TM: Moody's Model for Predicting Loss Given Default (LGD).

Grinold, R. C., & Kahn, R. N. (1999). Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Selecting Superior Returns and Controlling Risk, McGraw-Hill.

Jorion, P. (2000). *Value at Risk: The New Benchmark for Managing Financial Risk.* McGraw-Hill.

Junni, L., Zhang, W., & Härdle, K. (2010). The Bayesian Additive Classification Tree applied to credit risk modeling. *Comput. Stat. Data Anal., 54*(12), 1197-1205. doi: https://doi.org/10.1016/j.csda.2009.11.022

KMV. (1993a). *Portfolio Management of Default Risk*. Released date: November15, 1993, Revised: May 31, 2001.

KMV. (1993b). *San Francisco, Credit Monitor Overview*. KMV Corporation.

Kyoung-jae, K., & Hyunchul, A. (2012). A corporate credit rating model using multi-class support vector

machines with an ordinal pairwise partitioning approach. *Comput. Oper. Res., 39*(8), 1800-1811. https://doi.org/10.1016/j.cor.2011.06.023

Ligang, Z., Kin, K. L., & Lean, Y. (2009). Credit scoring using support vector machines with direct search for parameters selection. *Soft Comput, 13*(2),149-155. doi: https://doi.org/10.1007/s00500-008-0305-0

Liyong, Y., & Jiehui, Z. (2004). Research on default probability prediction based on Logistic regression analysis (in Chinese). *Journal of Finance and Economics, 9*, 15-23.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. *J. Bank Financ., 1*(3), 249-276. https://doi.org/10.1016/0378-4266(77)90022-X

Morgan, J. P. (1997). *Credit Metrics*. New York, Technical Document, April 2.

Naiyang, D., & Yingjie, T. (2009). *Support vector machine: Theory, algorithm, development* (1st ed.). China Science Press.

Naiyang, D., & Yingjie, T. (2009). *Support Vector Machine-Theory.* Algorithms and Extensions, Science Press.

O'Connor, R., Golden, J. F., & Rack, R. (n. d.). A Value-At-Risk Calculation of Required Reserves For Credit Risk In Corporate Lending Portfolios.

Pai, P. F., Tan, Y. S, & Hsu, M. F. (2015). Credit rating analysis by the decision-tree support vector machine with ensemble strategies. *Int. J. Fuzzy Syst., 17*(4), 521-530. https://doi.org/10.1007/s40815-015-0063-y

Qian, L., Bing, Y., Yi, L., Naiyang, D., & Ling, J. (2012). Constructing support vector machine ensemble with segmentation for imbalanced datasets. *Neural Comput. Applic*.

Reining, A. (2001). *Monte Carlo Simulation in the Integrated Market and Credit Risk Portfolio Model*. Algorithmic Inc.

Saunders, A. (1999). *Credit Risk Measurement.* New York: John Wiley & Sons.

Shinong, W., & Shizhong, H. (2001). Research on prediction model of financial distress of listing Corporation in China (in Chinese). *Economics Research Journal, 6*, 46-55.

Shu-Ting, L., Bor-Wen, C., & Chun-Hung, H. (2009). Prediction model building with clustering-launched classification and support vector machines in credit scoring. *Expert Syst. Appl., 36*(4), 7562-7566. https://doi.org/10.1016/j.eswa.2008.09.028

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decis. Support Syst., 46*(1), 287-299. https://doi.org/10.1016/j.dss.2008.06.013

Terry, H. (2015). Credit scoring using the clustered support vector machine. *Expert Syst. Appl., 42*(2), 741-750. https://doi.org/10.1016/j.eswa.2014.08.029

Wenbing, X., Qi, F., & Hu, W. (2007). The credit evaluation model based on support vector machine and risk, (in Chinese). *Chinese Science Abstracts (Chinese Edition), 13*(22), 284.

Wun-Hwa, C., & Jen-Ying, S. (2006). A study of Taiwan's issuer credit rating systems using support vector machines. *Expert Syst. Appl., 30*, 427-435. https://doi.org/10.1016/j.eswa.2005.10.003

Yifeng, Z., & Chengde, L. (2005). The comparison on classification method of credit risk assessment in commercial bank, (in Chinese). *The 24th Chinese control conference*, 1734-1737.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Syst. Appl., 42*(3), 508-516. https://doi.org/10.1016/j.eswa.2014.12.006

Zhihui, L., & Meng, L. (2005). Empirical Research on the credit risk identification model in Chinese commercial banks (in Chinese). *Economic Science, 5*, 61-71.