Journal of Financial Risk Management, 2025, 14(1), 37-46
https://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

# Research on Credit Risk Rating of Commercial Banks Based on Support Vector Machine

## —Data from China's Listed Commercial Banks

**Shunquan Zhu**

School of Digital Finance, Guangzhou Huashang College, Guangzhou, China
Email: hdxgzsq@163.com

**How to cite this paper:** Zhu, S. Q. (2025). Research on Credit Risk Rating of Commercial Banks Based on Support Vector Machine. *Journal of Financial Risk Management, 14,* 37-46.
https://doi.org/10.4236/jfrm.2025.141003

**Received:** January 15, 2025
**Accepted:** February 25, 2025
**Published:** February 28, 2025



## Abstract

In the era of big data and artificial intelligence, machine learning is one of the hot issues in the field of credit rating. On the basis of combing the literature on credit rating methods at home and abroad, this paper uses the support vector machine (SVM) method of machine learning to set up a three-classification credit rating model for Chinese listed commercial banks, which is compared with the existing methods of credit rating, the data show that the model classification accuracy on the test set is 80%. The support vector machine (SVM) method can identify the credit rating of Chinese-listed commercial banks, and it has wide applicability and good promotion value. The paper enriches the traditional credit rating method and has important significance in standardizing the healthy development of the financial market.

## Keywords

Chinese Listed Commercial Banks, Support Vector Machine, Credit Risk Rating, Data Supporting

## 1. Research Background and Literature Review

In the era of big data and artificial intelligence, the credit rating of artificial intelligence machine learning has been widely concerned by academia, financial industry and government departments. Traditional credit rating methods, such as statistical and back-propagation (BP) neural network, are difficult to solve the problems of small sample, local minimum, high dimension, poor ability of function approximation and classification, low learning speed and need to guide learning. This seriously affects the quality of credit ratings. Therefore, it is necessary to explore a new credit rating method-credit rating for listed companies based on

least squares support vector machine (LSSVM)—to solve the above problems and improve the quality of credit classification. Therefore, this paper can not only enrich the research. The traditional credit rating method is of great significance to the healthy development of the financial market.

The research on credit rating has experienced the following types: qualitative rating method, statistical rating method, neural network rating method and KMV rating method. (1) Qualitative analysis rating method, mainly 5C element analysis method and LAPP principle, 5C element analysis method mainly analyzes the credit status from character, ability, capital, collateral guarantee and condition environment; LAPP principle mainly analyzes credit status from four aspects: liquidity, activity, profitability and potential. In addition, there are DuPont financial analysis system and Wall proportion analysis method. The disadvantages of these methods are strong subjectivity and great subjective influence by people. In order to overcome the problems of poor qualitative analysis ability, lack of overall summary and insufficient quantitative analysis, statistical rating method has been widely adopted abroad since the 1960s. In order to overcome the poor comprehensive analytical ability of the qualitative analysis method, they lack the overall generalization. Since the 1960s, statistical rating method has been widely used in foreign countries. (2) Statistical rating method, Beaver (1966) introduced the forecasting function of financial variables into the empirical field. The univariate model of credit classification is established. Altman (1968) first applied multivariate statistical method to credit classification. However, statistical law is strict regarding data. For example, the data must be obeyed from the multivariate normal distribution, there are no multiple collinearities between the variables, the covariance matrix of the paired samples should be the same, etc., but the data, in reality, cannot meet this requirement. Therefore, the follow-up scholars use Probit or Logistic to establish the model. Ohlson (1980) suggests that the logistic regression method should be used to establish the credit rating classification model. Compared with previous studies, Collins and Green (1982) showed that the effect ratio of Logistic model was higher than that of Logistic model. The effect of multivariate discriminant analysis model is good. Lo (1986) found that if the sample data is under normal distribution, the effect of the multivariate credit model is better than that of the Logistic regression model, and if the sample data does not obey the normal distribution, the effect of the multivariate credit model is better than that of the Logistic regression model. The effect of Logistic model is better than that of multivariate discriminant model. Therefore, Lo A. W. suggested that before establishing credit classification model, Anthony et al. (2001) used survival analysis to predict the financial distress of deposit and loan institutions, and used logarithm to test the distribution characteristics of data samples. Logistic distribution is the probability distribution of survival time. The results show that the prediction accuracy of Logistic distribution is higher than that of Probit method, and it is robust to the definition of dilemma and the cost of misjudgment. (3) The neural network rating method in the late 1980s and early 1990s. With the devel-

opment of information technology, the neural network method introduced credit rating. It can solve the problem of non-normal distribution and nonlinear credit classification, but it is difficult to solve the problem of small sample data, local minimum, and high dimension. The ability of function approximation and classification is weak and the learning speed is slow. (4) KMV rating method appeared in the late 1990s. There are many new credit classification models, the most representative of which are the CreditMetrics model established by JP Morgen Bank on the basis of VaR model in 1997 and the KMV model developed by KMV Company. However, there are many parameters to be determined when establishing CreditMetrics model, such as rank transfer matrix, correlation coefficient between assets, forward return rate and so on. These parameters come from the statistics of long-term data. At present, there are few similar statistics in China.

The study of credit rating by Chinese scholars began in the mid-1990s. Wang et al. (1999), Chen (1999), Chen & Chen (2000), Wu (2001), Zhang & Cheng (2005), Shi et al. (2005), Zhu (2008), Wu et al. (2009), Yu et al. (2009), Zhu (2009) and other scholars have published some research results. Wu et al. (2009) established a support vector machine ensemble method based on five-level classification and applied Libsvm to a commercial silver. The empirical analysis of 176 groups of bank credit sample data shows that the proposed method has higher classification accuracy than other classification methods, which proves the feasibility and effectiveness of this method. The results of Yu et al. (2009) show that the least square fuzzy support vector machine model with variable penalty factor based on kernel principal component analysis has good classification results. Zhu (2009) applied option pricing method to the credit classification of listed companies.

## 2. Method and Principle of Support Vector Machine

Support Vector Machine (SVM) is a classifier based on support vector operations. "Machine" means a machine, which can be understood as a classifier, and a classifier is a classification function.

### 2.1. Linear classification

In training data, each data has $n$ attributes and a class of class flags. We can think that these data are in an n-dimensional space. Our goal is to find a $n-1$ dimensional hyperplane, which can divide the data into two parts or parts, each of which belongs to the same category.

In fact, there are many hyperplanes like this, and we have to find one of the best. Therefore, a constraint is added: the distance from the hyperplane to the nearest data point on each side is maximum. It also becomes the maximum interval hyperplane. This classifier also becomes the maximum interval classifier. A support vector machine (SVM) is a two-class or multiple-classifier.

For training data set T, the data can be divided into two categories: C1 and C2. For functions: $f(x)=wx^T+b$ (Linear classifier). Data for C1 classes $wx^T+b\geq 1$.

At least one of them. $x_i, f(x_i)=1$ This point is called the nearest point. Data for C2 classes $xw^T+b\le -1$. At least one of them $x_i, f(x_i)=-1$. This point is also the nearest point. The above two constraints can be combined as: $y_i f(x_i)=y_i\left(wx^T+b\right)\ge 1$. $y_i$ is the corresponding classification value of $x_i$ (−1 or 1). Seek $w$ and $b$, then the hyperplane function is $wx^T+b=0$. In order to obtain the optimal $f(x)$, in training data, we expected that the distance from each point to the hyperplane would be the largest.

### 2.2. Nonlinear Classification

One advantage of the SVM is that it supports non-linear classification. It is combined with Lagrange multiplier method and KKT condition, and the kernel function can generate nonlinear classifier.

$w=\sum_{i=1}^{n}\alpha_i y_i x_i$; $f(x)=\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b$, here $x_i$: training data $i$; $y_i$: label value of training data $i$; $\alpha_i$: the Lagrange multiplier of the training data $i$; $K(x_1,x_2)=\exp\left(-\frac{\|x_1-x_2\|^2}{2\sigma^{12}}\right)$: Kernel function; $\alpha,\sigma$ and $b$ is the value generated by the training data.

By adjusting $\sigma$ to match the size of the dimension, the larger $\sigma$, the lower the dimension.

### 2.3. Kernel Functions for Nonlinear Classification

According to the theory of machine learning, nonlinear problems can be transformed into linear problems by mapping to high dimensions. For example, a point in two dimensions. $\langle x_1,x_2\rangle$ can be mapped to a 5-dimensional space with five dimensions: $x_1,x_2,x_1x_2,x_1^2,x_2^2$. Mapping to high dimensions has two questions: how does one map? Another problem is that the calculation becomes more complicated. Fortunately, we can use the kernel function to solve this problem.

Kernel function, also known as kernel technique (kernel trick). The kernel function provides a method to calculate the inner product of the high dimensional space by the vector value of the original space, regardless of the way of mapping.

We can replace $K(x_1,x_2)$ with a kernel function.

There are many kinds of kernel functions, generally using Gauss kernel, the following formula:

$$K(x_1,x_2)=\exp\left(-\frac{\|x_1-x_2\|^2}{2\sigma^{12}}\right)$$

## 3. Variable Selection and Modeling Sample

We choose the credit rating of listed commercial banks to study the application of SVM. According to the evaluation reports of the authoritative credit rating agencies such as United, Zhongcongxin and Dagong International, the credit rating of commercial banks generally takes into account the following indicators: total as-

sets, total loans, total deposits, owner's equity, operating income, Net profit, rate of return on total assets, rate of return on net assets, non-performing loan ratio, provision coverage, inventory ratio, liquidity ratio, capital adequacy ratio, tier one capital adequacy ratio, core tier one capital adequacy ratio.

Considering the perfection and accuracy of data acquisition, only domestic commercial banks with credit ratings of AAA, AA+, AA are adopted. Banks with a credit rating of AA– have less or more incomplete data. In this paper, two variables are selected as characteristic values, liquidity ratio and profit ratio, for the following reasons: first, the size of assets varies among different banks in order to eliminate the impact of data absolute value differences on the study. It is appropriate to use financial ratio. The second is to make the drawing more intuitive, so only two eigenvalue vectors are taken for training; the third is that the liquidity ratio reflects the bank's solvency, and the profit ratio reflects the bank's profit capacity, which is an important indicator of credit rating.

### 3.1. Collect to Data

Because there are too many incomplete values of bank financial data obtained by tushare data interface, and the number of banks is too small. Therefore, the use of API on Uqer site to obtain.

Data operating income and net profit: income and revenue.

The profit ratio = NIncome/revenue.

Access to data on current assets and current liabilities:

Current assets = tradingFA + loanToOthBankFi + intReceiv + purResaleFa + CReserCB + derivAssets + deposInOthBfi + preciMetals + investAsReceiv;

current liabilities = CBBorr + depos + loanFrOthBankFi + tradingFL + soldForRepura + payrollPayable + taxesPayable + intPayable + bondPayable + deposFrOthBfi + derivLiab;

Current ratio = current assets/current liabilities.

The exact meaning of these variables can be found in the data menu in the mining platform.

In addition, we collect data by looking at bank annual reports and rating reports issued by credit rating agencies.

### 3.2. Read in Data

Import Python's numpy, pandas, sklearn, matplotlib module and train_test_split, pyplot, svm function.

```
import numpy as np
import pandas as pd
import sklearn
from sklearn.model_selection import train_test_split
from sklearn import svm
import matplotlib as mpl
import matplotlib.pyplot as plt
```

```
data=np.loadtxt('F:/2glkx/data/zonghe1.txt',delimiter=',')
# using loadtxt to read data can directly convert text data into arrays for easy operation
# to facilitate subsequent training, set the AAA to 3 AA to 2.5 AA to 2
Data = np.loadtxt('F:/2glkx/data/zonghe1.txt',delimiter=',')
# convert floating-point data to a string for later operations
data=data.astype(Str)
```

## 4. Training and Testing Results of Credit Rating of Commercial Banks Based on Support Vector Machine

### 4.1. Separation of Data into Training and Test Sets

```
x, y = np.split (data, (2,), axis = 1)
x_train, x_test, y_train, y_test = sklearn.model_selection.train_test_split (x, y, random_state = 1, train_size = 0.6)
```

1) split (data, segmentation position, axis 1 (horizontal partition) or 0 (vertical segmentation);

2) sklearn. model_selection. train_test_split randomly divided into training set and test set.

```
train_test_split (train_data, train_target, test_size = number, random_state = 0).
```

Parameter interpretation:

train_data: the sample feature set to be partitioned.

train_target: sample results to be partitioned.

test_size: the percentage of test_size: samples, if an integer, is the number of samples.

random_state: is the seed of random numbers.

Random number seed: in fact, the number of random numbers, in case of repeated trials, ensure that the same set of random numbers. For example, if you fill in 1 at a time, you will get the same random array if the other parameters are the same. But fill in 0 or not, each time will be different. The generation of random numbers depends on the seeds. The relationship between random numbers and seeds follows the following two rules: different seeds produce different random numbers and the same seeds produce the same random numbers even if the instances are different.

### 4.2. Training Classifier

```
# clf = svm.SVC (C = 0.1, kernel = 'linear', decision_function_shape = 'ovr')
clf = svm.SVC (C = 0.8, kernel = 'rbf', gamma = 100, decision_function_shape = 'ovr')
clf.fit(x_train, y_train.ravel())
```

When kernel = 'linear',The larger C is, the better the classification effect is, but it is possible to over-fit (defaul C = 1).

In kernel = 'rbf', the smaller the gamma value is, the more continuous the clas-

sification interface is, and the more "scattered" is, the better the classification effect is, but it may be over-fitted.

When decision_function_shape = 'ovr' is defined as one v rest, that is, one category is divided from other categories, and when decision_function_shape = 'ovo' is defined as the one v one, is divided into two categories, the result of multi-classification is simulated by two-classification method.

The following results were obtained:

SVC (C = 0.8, cache_size = 200, class_weight = None, coef0 = 0.0,
decision_function_shape = 'ovr', degree = 3, gamma = 100, kernel = 'rbf',
max_iter = −1, probability = False, random_state = None, shrinking = True,
tol = 0.001, verbose = False)

After several attempts, gamma = 100 can achieve the best effect.

print (clf.score (x_train, y_train))

print (clf.score (x_test, y_test))

The following results were obtained:

0.7962962962962963

0.5277777777777778

It can be seen from above that the accuracy of the SVC classifier is 0.80 for the training set and 0.53 for the test set.

## 5. Visual Graphics for Credit rating of Commercial Banks

# determining the range of coordinate axes x,y represents two characteristics respectively

x = x.astype(float)

y = y.astype(float) #Convert the x,y data type back to a later operation

x1_min, x1_max = x [:, 0].min(), x [:, 0].max () # Range of column 0

x2_min, x2_max = x [:, 1].min(), x [:, 1].max () # Range of column 1

x1, x2 = np.mgrid[x1_min:x1_max:200j, x2_min:x2_max:200j]

#Generating grid sampling points

grid_test = np.stack ((x1.flat, x2.flat), axis = 1) # Test point

grid_hat = clf.predict (grid_test)

grid_hat = grid_hat.reshape (x1.shape)

# Predictive classification value grid_hat = grid_hat.reshape (x1.shape)

#Make it the same shape as the input

# Draw a Visualized Drawing

cm_light = mpl.colors.ListedColormap (['#A0FFA0', '#FFA0A0', '#A0A0FF'])

cm_dark = mpl.colors.ListedColormap (['g', 'r', 'b'])

x1 = x1.astype (float)

x2 = x2.astype (float)

grid_hat = grid_hat.astype (float)

plt.pcolormesh (x1, x2, grid_hat, cmap = cm_light)

plt.scatter (x [:, 0], x [:, 1], c = 'b', edgecolors = 'k', s = 25, cmap = cm_dark) #sample

```
plt.xlabel (u'Current Ratio', fontsize = 13)
plt.ylabel (u'Profit Ratio', fontsize = 13)
plt.xlim (x1_min, x1_max)
plt.ylim (x2_min, x2_max)
plt.show ()
```

Get the figure as shown in Figure 1.

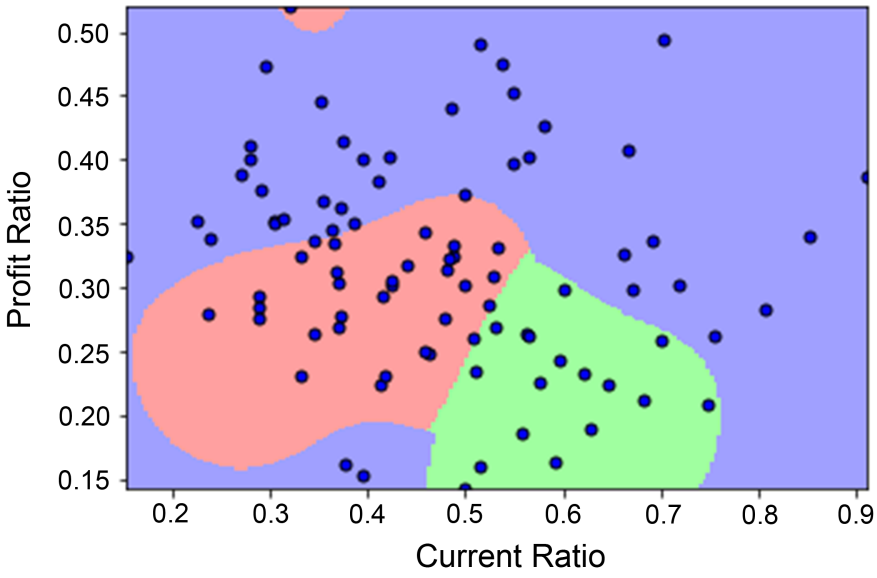


Figure 1. Three categories of credit ratings of commercial banks.

## 6. Analysis and Conclusion of the Results of Three Categories of Credit Rating of Commercial Banks

It can be seen from Figure 1 that the scattered sample points are divided into three categories according to the classifier trained by the sample set, that is, AAA, AA A, which is distributed in three color regions. The purple area above is AAA grade commercial banks. It can be seen that its profit margin is the highest. The liquidity ratio can be seen to be between 0.3 and 0.5 through scattered point distribution, but there are also some very high liquidity ratios, which exceed 0.5. The red area is AA commercial bank; its profit margin level is lower than AAA grade commercial banks, higher than AA grade, and its liquidity ratio is also between 0.3 and 0.5. Green area is AA grade bank. AA-rated banks have low-profit margins but a high liquidity ratio, mostly between 0.5 - 0.8.

Because the number of data samples is limited by the actual situation (the number of banks), and the data are not all from the same source (excellent mining, bank annual reports, Rating assessment reports, and the final use of training data is recomputed (profit margin = net profit/operating income, liquidity ratio = current assets/current liabilities), with significant errors, Only two characteristic variables (profit margin, liquidity ratio) are selected. Therefore, it cannot fully reflect the credit rating of commercial banks. However, we can see from Figure 1 that profit margin is an important indicator of bank credit, and it reflects the bank's

earnings. Profit ability: The higher the credit level of the bank, the stronger its profitability.

## Fund Program

This research is the phased achievements of the scientific research ability improvement project of Key construction disciplines in Guangdong Province (No: 2024ZDJS113), the Application Oriented Demonstration Major of Guangzhou Huashang College (No: HS2024SFZY08) and the construction project of the core course teaching and Research Section of Financial Technology major of Guangzhou Huashang College (No: HS2024ZLGC43), etc.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance, 23,* 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Anthony, H., Catanach, J., & Perry, S. E. (2001). An Evaluation of the Survival Model's Contribution to Thrift Institution Distress Prediction. *Journal of Managerial Issue, 8,* 401-417.

Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research, 4,* 71-111. https://doi.org/10.2307/2490171

Chen, J. (1999). Empirical Analysis of Financial Deterioration Forecast of Listed Companies. *Accounting Research, 4,* 31-38.

Chen, X., & Chen, Z. H. (2000). Theoretical Method and Application of Enterprise Financial Distress Research. *Investment Research, 6,* 23.

Collins, R. A., & Green, R. D. (1982). Statistical Methods for Bankruptcy Forecasting. *Journal of Economics and Business, 34,* 349-354. https://doi.org/10.1016/0148-6195(82)90040-6

Lo, A. W. (1986). Logit versus Discriminant Analysis: A Specification Test and Application to Corporate Bankruptcies. *Journal of Econometrics, 31,* 151-178. https://doi.org/10.1016/0304-4076(86)90046-1

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research, 18,* 109-131. https://doi.org/10.2307/2490395

Shi, X. J. et al. (2005). Research on Consistency between Options and Accounting Information Credit Model. *Theory and practice of Systems Engineering, 10,* 11-20.

Wang, C. F. et al. (1999). Credit Risk Assessment of Commercial Banks Based on Neural Network Technology. *Theory and practice of Systems Engineering, 9,* 24-32.

Wu, C. et al. (2009). Research on Credit Risk Assessment Model of Commercial Banks Based on Five-Level Classification Support Vector Machine Integration. *Forecast, 4,* 57-61.

Wu, S. N. (2001). Research on Forecasting Model of Financial Distress of Listed Companies in China. *Economic Research, 6,* 46-55.

Yu, L. et al. (2009). A Least Squares Fuzzy Support Vector Machine Model with Variable

Penalty Factor based on Kernel Principal Component Analysis and Its Application in Credit Classification. *Systems Science and Mathematics, 10,* 1311-1326.

Zhang, M., & Cheng, T. (2005). Dynamic Perspective of Empirical Research on Financial Early Warning of Listed Companies. *Financial Research, 31,* 62-70.

Zhu, S. Q. (2009). Research on Credit Classification Modeling and Application of Listed Companies Based on Option Pricing Theory. *Statistics and Information Forum, 7,* 23-38.

Zhu, Y. F. (2008). Feature Selection and Effect Evaluation of Microenterprise Credit Evaluation Based on Neural Network. *Statistics and Information Forum, 4,* 8-11.