Research on Credit Risk Assessment in Commercial Bank Based on Information Integration

GUO Yingjian, WU Chong
School of Management, Harbin Institute of Technology, P.R.China, 150001
guoyj@bankcomm.com

Abstract: Credit risk is the main risk faced by commercial bank during operation, how to manage credit risk is a focus for financial specialists. Therefore effective assessment of credit risk is very important. The risks faced by banks during operation can be regarded as a kind of research of uncertain problems' study. The essence of multi-sensor data integration is as follows: to have a best decision according to much information from different sensors. A kind of credit risk assessment model based on information integration is developed depending on the idea of multi-sensor information integration. This model includes three algorithms, as follows: BP neural network, SVM and DS evidence theory. This model has not only the classifying capacity of BP neural network and SVM, but also the decision ability of DS evidence theory. The essence of the model has two key steps, first we can obtain elementary decision, then we depend on DS evidence theory has a integration of processed outputs data by BP neural network and SVM. Depending on some data of certain bank, we have a simulation aiming at three models: BP model, SVM model and new developed model. The results show the developed new model can obtain better assessment compared with two models at same conditions.

Keywords: Credit risk, BP neural network, SVM, DS evidence theory

1 Introduction

Due to its special operating targets, extensive social connections and powerful influence, as a special enterprise engaging financial assets, commercial bank has gradually become the focus of risk. Commercial bank has to face all kinds of credit risks, particularly the risks because of the defaults by the transaction parties. Therefore, it is necessary to effectively establish a credit risk assessment system to promote the bank development and to avoid unnecessary losses. No matter in terms of actual or theoretical background, credit risk assessment has become a research concern and focus of economists and banks. Establishing a standardized and practical assessment system for commercial banks is of positive and practical significance to comprehensively improve the bank's management level and to effectively reduce and prevent credit risks: it is helpful to strengthen the credit risk management level of China's commercial banks; it is helpful to adapt to the international banks and to participate in the international competitions; it is helpful to develop new financial derivative products and to improve China's capital markets; it is helpful to learn from international risk management experience and to improve the supervision level of banking.

2 Research status of credit risk and its assessment model in commercial bank

The so-called credit risk means the possibility of loss to commercial banks caused by the defaults of the transaction parties (debtor and creditor). Credit risk is a two-way risk, which can be expressed by a function as $C = F (A, B, E)$; $C$ represents credit risk; $A$ represents the loan party; $B$ represents the bank; and $E$ represents the environment. Credit risk is a traditional type of risk faced by the commercial bank, which is one of the three major risks (credit risk, market risk and operational risk) of banks. In order to fully and accurately reflect the status of the bank's credit risks, to timely and effective response to potential credit risks, no matter the bank or the regulatory agencies have proposed high demands for the measurement of credit risks. Therefore, domestic and foreign experts and scholars have made a more
comprehensive research on the credit risk assessment from different aspects.

2.1 Traditional credit risk assessment model
Traditional credit risks basically include three categories: expert method, rating method and credit scoring method. Expert method is the bank credit expert's judgment based on subjective experience. The conspicuous advantage is the flexibility and the advantage in dealing with qualitative indicators. The basic idea of the rating method is to divide the bank loans into several levels and to endow the different levels with different loss reserve rate; then calculate the loss reserve and the sum; at last the necessary capital of the bank to prevent risks is obtained. Banks and regulators make use of this method to assess whether the lender's loss reserve is adequate.

Credit scoring method is to take the financial ratio of evaluation object as explanatory variable, to establish regression model by the mathematical statistics method, to compare the credit score or default probability of the model output with the reference value and to measure the risk of the evaluation object. The application of this model is the most effective and widespread, which has been regarded as the mainstream research methods for credit risk by the international financial industry and the academic circle, including multiple discriminant analysis model, linear probability model, Logit model, Probit model, neural network model, support vector machine model and etc. In terms of multiple discriminant analysis model, the foreign scholar Altman, E.I, P. Narayanan[1], Altman, E.I, Saunders A [2] has successfully made use of the multiple discriminant analysis model for the first time to establish the famous five-variable Z-score method and the improved Zeta discriminant analysis method base on it; Shi Xiquan and Zou Xinyue[3], Zhang Aimin, Zhu Chunshan and Xu Jian[4], Wang Jiancheng[5] and Liang Qi[6] have made credit scoring research by using multiple discriminant analysis model and have conducted empirical studies. Bu Huimin[7] has established the commercial bank's credit scoring model based on the AHP analysis framework on the basis of the reality of China's commercial banks, which has provided a meaningful reference for the quantization and measurement of the commercial bank's credit risk. In terms of Logit & Probit linear regression model method, Press, S.J. and Wilson[8], Sjur W and Wijst,N[9] have established the Logit credit scoring model learning from the Logistic function. West, R.C[10] has obtained the default probability of every institution by using the Logit Model to analyze the credit risk situation of financial institutions. Smith and Lawerence[11] have obtained the optimal predicated loan default variable by using the Logit model. In terms of neural network model method, Odom, M.D. and Sharda[12], Desali VS, Crook J N and Overstreet GA[13] have established the neural network model method, which has been obtained results in the empirical study in a number of countries; WEST D[14] has established five different types of neural network model to study the accuracy of the commercial bank's credit evaluation; Malhotra R and Malhotra D,K[13] has identified the loan enterprise by using the neuro-fuzzy system. Zhang Zhongzhi, Fu Lin and Tang Huanwen[16] have made research on the artificial neural network model for building the commercial bank's credit risk; the empirical results show that the artificial neural network model has very high prediction accuracy. Wu Desheng and Liang Mu[17] have compared the pros and cons of different models used for enterprise credit evaluation; in view of the characteristics of the credit scoring, they have made empirical study and comparison of diagnosis behavior of the two networks by using Elman regression neural network and BP network modeling. In terms of support vector machine model method, Chang CC[18] has applied the SVM to the credit risk assessment process and has established the credit risk assessment model; domestic scholars Liu Min & Lin Chengde[19], Liu Yundao & Wu Chong[20] have established a credit risk assessment model by using the SVM technology and have conducted empirical studies, which has provided a new reference for the quantization and measurement of the commercial bank's credit risk.

2.2 Modern credit risk assessment model
In the 1990s, a number of new commercial credit risk measurement model were introduced, including the famous CreditMetrics credit measurement model of well-known bank JP Morgan, the EDF model of KMV company, the Credit Risk + model of CSFB and the Credit Portfolio View model of McKinsey & Company. These methods have become the commercial bank's credit risk measurement model and the standard management method. The reason for these models to become outstanding representatives of
modern risk measurement models are as follows: they have explained the risk measurement from the aspect of a new theory; they have resolved some important problems that cannot be resolved by the past traditional credit risk measurement models; and they have created a new era of risk measurement. Both traditional and current models are related with some internal factors of the financial system, therefore many credit risk assessment models cannot be used directly. In short, the commercial bank's credit risk assessment is a multi-level multi-factor comprehensive evaluation system. Therefore, the research core is to seek a better comprehensive evaluation system.

3 Credit risk assessment model in commercial bank based on information integration

3.1 The idea of information integration
Information Integration is also known as data integration; information integration is essentially the multi-sensor data integration. Information integration is information processing technology developed in the recent few decades, which is to make comprehensive analysis on the information obtained by various means and at any times as a whole and to lay a foundation for the decision-making and control. Take into account that the commercial bank's credit risk assessment is comprehensively evaluated based on a variety of factors; therefore, the information integration method should be improved appropriately. The credit risk assessment applied to the commercial banks not only deepens the commercial bank's evaluation system, but also contributes to practical application. Based on this point, the model has not only made use of the mature BP neural network and support vector machine, but also has introduced the DS evidence theory to establish an information integration model based on the three for the commercial bank's credit risk assessment.

3.2 The establishment of index system
Before determining the credit risk assessment model, a suitable risk assessment system must be chose. In view that the sample data capacity used in this paper us not great, the value range of all indexes is broader and the data is gliding property, the twice or three times standard deviation test is adopted to omit the abnormal data and ultimately obtain 189 sample data; the statistical software SPSS is adopted to make factor analysis on the obtained data. Main purpose to obtain the explanation factor is to determine the number of minimum factor for interpreting the correlation between the observed variables. According to the eigenvalue criteria, the number of interpretation factor under the condition of the current sample data is 4; meanwhile, it can seen from the factor load matrix that, the economic implications of the factors are obvious. The sample data of the 14 financial indicators can be divided into four interpretation factor as loan enterprise's solvency factor, profitability factor, asset management factor and enterprise development capacity factor. The above four factors are determined by the comprehensive index; then the output of neural network is adopted to judge the loan's risk type. At present, in accordance with the requirements of the People's Bank, the loan's risk can be divided into the following five types: normal level, special mentioned level, substandard level, doubtful level and loss level.

3.3 Commercial bank's credit risk assessment model based on the BP neural network
BP network is the abbreviation of Back-Propagation Networks. The basic idea of BP learning algorithm is to transfer the mapping problem of neural network input/output into non-linear optimization problem, to use the gradient descent algorithm in the optimization, to use iterative computation correct the network weight and to realize the minimum mean square error between the network output and the expected output. BP neural network is a kind of error back propagation recurrent neural network, which has good generalization ability, pattern recognition and classification advantages and is in line with characteristic of commercial bank's credit risk classification. The network structure of BP neural network is shown as Fig.1:
The credit risk assessment model based on BP neural network is established according to the network structure in Fig.1 and the research reality of the credit risk, as shown in Fig.2:

The selection of the network input & output and the hidden layers should comply with the following rules:
According to the commercial bank's credit risk index system, the input layer is set up as four-dimensional input and five-dimensional output structure; the four-dimensional vector of the input layer includes the loan enterprise's solvency factor, profitability factor, asset management factor and enterprise development capacity factor; the vector of five-dimensional output represents the customer's credit rating; the five credit ratings are normal level, special mentioned level, substandard...
level, doubtful level and loss level, represented by \((10000), (01000), (00100), (00010)\) and \((00001)\) respectively.

The choice of hidden layers is very important. The role of hidden node is to store mapping rules by using weight of each hidden node. If there are few hidden nodes, the network's ability to extract information from the samples is poor, which is not enough to reflect the detailed mapping; if there are too many nodes, the interference or noise in the samples will be remembered as rules, which may affect the accuracy of the network performance. Generally speaking, the complex nonlinear mapping requires that the network has more hidden nodes to enhance the mapping capacity. This paper adopts the trial-and-error method, which means that gradually adjust the number of hidden nodes in the same training sample, until the network reaches the minimum error; to set up the number of hidden layers as 1, choose "hidden layer neurons =2 * input neurons +1" as a reference formula and set up as 9; the author also has made other choices, whose classification results are all worse than that of 9.

### 3.4 Commercial bank's credit risk assessment model based on Support Vector Machine

Support Vector Machine (abbreviated as SVM) is a machine learning method put forward by Vapnik and other people in accordance with the statistical learning theory. SVM has advantages as the global optimization, simple structure, good promotion ability and etc., therefore, SVM has been more and more widely used in many research fields, including pattern recognition, regression estimation and density estimation and etc. At present, SVM has been applied to the commercial bank's credit risk assessment. Generally speaking, the bank's customer credit assessment issue can be attributed to intelligent system design, that is, through learning the existing credit sample data (the so-called training sample), simulate the rule of the assessment system and make use of the simulated rule to classify the data to be judged.

The basic idea of SVM: mapping the input vector to a high-dimensional feature space \(Z\) through the pre-selected nonlinear mapping input vector and construct an optimal separating hyperplane in this space. The classification function of SVM is similar to the neural network in the form; the output is a linear combination of intermediate nodes. Each intermediate node has a corresponding support vector machine, as shown in Fig.3.

![Fig.3 The structure scheme of support vector machine](image)

Commercial bank's credit risk assessment by using SVM has also become a focus. At present, the classifications based on SVM are basically centralized in the double-value classification; while the research in this paper mainly adopts the multiple-value classification. In the research in this paper, the following classification ideas are adopted:

Provided that the bank's credit rating is divided into five levels: normal level, special mentioned level, substandard level, doubtful level and loss level, all classifications can be completed by three steps. First of all, classify the normal level, special mentioned level, and substandard level into a big category and
classify the doubtful level and loss level into a big category, respectively represented as C11 and C12; C11 includes the normal level, special mentioned level and the C12 includes the doubtful level and loss level. Then further classify the C11 and C12 by using SVM; C11 is divided into C111 (normal level) and C112 (special mentioned level and substandard level); C12 is divided into C121 (doubtful level) and C122 (loss level). At last, further classify the C112 into C1121 (special mentioned level) and C1122 (substandard level) by using C112. Actually, it is to gradually make classification by using the binary tree and SVM, thus obtain the multi-category classification, as shown in the Fig.4:

![Diagram](example.png)

**Fig.4 Commercial bank's credit risk assessment model based on SVM**

### 3.5 Commercial banking credit risk assessment model based on information integration

The information integration model used by this paper is established on the basis of BP neural network, SVM and DS evidence theory (the model is recorded as IFCA). At first, make a introduction of the basic concepts and core algorithms of the DS evidence theory.

DS evidence theory is a method to process uncertain information developed on the basis of the Bayesian theory. DS theory has so far established a fairly perfect theoretical foundation, which not only can process the uncertainty caused by randomicity, but also can process the uncertainty caused by ambiguity. DS method can continuously narrow the assumption set through the accumulation of credit assignment function and distinguish the "unknown" and the "uncertainty". Another potential advantage of DS method is that it does not require the prior probability and the conditional probability density. In view of its advantages in this area, is can be completely used in the uncertainty assessment and also can be used in the credit risk assessment.

The principle of information integration assessment model established on the basis of BP network, SVM and DS evidence theory is as follows:

The detailed design algorithm of credit risk assessment model based on BP Network is as follows:

By using the above established BP neural network model, the commercial bank’s data can establish a four-input and five-output structure and a hidden layer (there are nine hidden layer nodes); the network output setup is a five-dimensional structure, including \((0\ 1\ 0\ 0\ 0), (0\ 1\ 0\ 0\ 0), (0\ 0\ 1\ 0\ 0), (0\ 0\ 0\ 1\ 0)\) and \((0\ 0\ 0\ 0\ 1)\); setting up the output layer as sigmoid function, the intervention of network output's every components can be realized. Network output may result in the situation that the sum of the five-dimensional vectors is not equal to 1; provided the vector of arbitrary output is
\( (y_1, y_2, y_3, y_4, y_5) \) and make \( \sigma = \sum_{i=1}^{5} y_i \) to enable the network output normalized as \( (y'_1, y'_2, y'_3, y'_4, y'_5) \), among which \( y'_i = \frac{y_i}{\sigma}, \quad i = 1, 2, 3, 4, 5 \) and \( y'_i \) is the basic trust degree of situation \( i \).

The detailed algorithm of credit risk assessment model based on SVM is as follows:

In order to effectively ease the complexity of the calculation, SVM adopts the LS-SVM. Kernel function adopts the radial basis function. The SVM contains two outputs at each time as +1 and -1, respectively representing the corresponding classification.

For the recognition rate of sample classification, the following rules are adopted:

Correct classification rate of samples: \( c_i = \frac{C_i}{N_i} \),

Average classification accuracy rate of samples: \( c = \frac{\sum_j C_j}{\sum_j N_j} \),

Sample miss rate: \( l_i = \frac{L_i}{N_i} \).

In the above Test index formula, \( N_i \) represents the sample number of the class \( i \), \( C_i \) represents the correctly recognized samples in class \( i \) and \( L_i \) represents the non-recognized samples in class \( i \) defect sample.

The algorithm schematic diagram combining the BP neural network, SVM and DS evidence theory is shown as Fig.5:
Start

Input the debt and credit record of enterprise

Data Feature Extraction

Determine whether the data is in line with the request of neural network

Establish the BP network and SVM model

BP network output and SVM output

BP output and SVM output normalized respectively

Integrate the normalized data of BP and SVM by using DS algorithm

Make decision by using the DS output

End

Fig.5 Commercial bank's credit risk assessment model based on IFCA

The BP network output is as above; for SVM, use fuzzy math idea to analyze the sample's dependence degree on each category of samples; through the one-by-one classification of SVM, obtain the five types of basic trust degree of certain result, provided as \((y_1^*, y_2^*, y_3^*, y_4^*, y_5^*)\), among which \(y_i^*\) is the trust degree of situation \(i\) through the SVM. According to \((y_1^*, y_2^*, y_3^*, y_4^*, y_5^*)\) and \((y_1^*, y_2^*, y_3^*, y_4^*, y_5^*)\) obtained under every situations, make use of DS evidence reasoning to integrate \((y_1^*, y_2^*, y_3^*, y_4^*, y_5^*)\) and \((y_1^*, y_2^*, y_3^*, y_4^*, y_5^*)\). Provided that the normal level, special mentioned level, substandard level, doubtful level and loss level can be represented by \(A_1, A_2, A_3, A_4\) and \(A_5\) and the BP network output results of certain data are \((y_1^*, y_2^*, y_3^*, y_4^*, y_5^*)\), make \(m_i(A_i) = y_i^*, \ i = 1, 2, 3, 4, 5\); in the same way, make \(m_2(A_i) = y_i^*, \ i = 1, 2, 3, 4, 5\); among which, \(m_i(A_i)\) represents the \(A_i\) trust degree of certain input obtained through BP algorithm; \(m_2(A_i)\) represents the \(A_i\) trust degree of certain input obtained through SVM (\(A_i\) and \(A_j\) are independent). Therefore, the composite formulas based on DS include:
\[ K = \sum_{i\neq j} m_i(A_i)m_j(A_j), \ 1 \leq i, j \leq 5 \]

\[ m(A_i) = \sum_{i\neq j} m_i(A_i)m_j(A_j)/(1-K), \ i = 1, 2, 3, 4, 5 \]

Make \( \hat{y}_i = m(A_i), i = 1, 2, 3, 4, 5 \), the \( \hat{y}_i \) is the \( A_i \) trust degree finally obtained by DS reasoning, thus the recognition rate of this information integration method can be obtained to make the evaluation decision.

4 Empirical Analysis

As the targeted customer of each bank is different, there is a big difference in all commercial banks credit risk assessment. The algorithm established in this paper is a extensively applied method for the commercial bank's risk assessment. But in specific empirical analysis, related tests only can be done to certain data to determine the feasibility of algorithm and to further make appropriate changes for the algorithm, in order to be used in commercial bank's other credit risk assessments.

The empirical data in this paper is a group of data collected from certain commercial bank; after the collection and sorting, there are totally 336 samples and finally 189 sample data; choose 100 samples as the training sample data and the rest 89 samples are predications. As the non-linear characteristics of the targeted system, the selection of the initial value and the convergence of the network has certain relationship; make normalization processing for the above data (the normalization algorithm adopted in this paper), \( data_{nn} = (data - \text{min})/\text{max} - \text{min} \), among which \( data_{nn} \) is the network input data-normalized data; \( data \) is the processed data obtained by certain module; \( \text{max} \) is the maximum value of processed data of corresponding module; \( \text{min} \) is the minimum value of processed data of corresponding module.

For BP network and the SVM training simulation, adopt the cross-validation algorithm; the 100 samples are divided into 10 groups; for each training, take 9 groups for training and take the rest group for testing; in the ten times of training, all groups have the opportunity to be used as testing sample.

For BP network, carry out training and testing in accordance with the above structure; for SVM, adopt the radial basis function method. For the simulation results of the above data, see Tab.1.

<table>
<thead>
<tr>
<th></th>
<th>Method</th>
<th>BP Network Model</th>
<th>SVM Model</th>
<th>IFCA Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>Accuracy (%)</td>
<td>78.3</td>
<td>80.6</td>
<td>84.5</td>
</tr>
<tr>
<td></td>
<td>Error rate (%)</td>
<td>21.7</td>
<td>19.4</td>
<td>15.5</td>
</tr>
</tbody>
</table>

It can be seen from the Tab.1, compared with the current widely used methods such as BP neural network and SVM, the commercial bank's credit risk assessment method based on information integration proposed by this paper is 4% higher in terms of assessment accuracy. The method has integrated the advantages of BP network SVM and DS evidence theory, thus has proved the feasibility and effectiveness of the model through the empirical analysis.

5 Conclusion
Commercial bank's credit risk is a major problem that the commercial bank must face, thus the commercial bank's credit risk assessment is a hot research issue of the bank management currently. The currently widely used methods are neural network and support vector machine. The paper has established a new information integration assessment model by combining the advantages of BP network and SVM classification and by using the characteristic that the DS evidence theory is good at dealing with the issue of uncertainty. The method has taken into account a few of the advantages and has proved the feasibility and effectiveness of the model through relevant empirical analysis, to lay an important theoretic foundation for the further practical application.

In the future further study, analyze a large number of empirical data, thus to improve the credit risk assessment model, making it has practical application value.

References