International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.11, (2016), pp.211-220
http://dx.doi.org/10.14257/ijsip.2016.9.11.19

# Recognizing and Predicting the Non-Performing Loans of Commercial Banks

Zhang Yu[1,2], Guan Yongsheng[1,3], Yu Gang[1] and Lu Haixia[1]

[1]*School of Management, Harbin Institute of Technology, P.R. China*
[2]*School of Economics, Harbin University of Science and Technology, P.R. China*
[3]*LongJiang Bank, Province Heilongjiang, P.R. China*
*Hester0524@163.coml*

***Abstract***

*As the reform and opening up going into depth over the past three decades and more, the market economic system has been gradually established. The banking industry grows steadily in the process of the reform. It supports economic development, reduces and defends many financial risks in the processof the reform. However, there are many kinds of risks inside of banks, one of which is that the non-performing loans (NPLs) ratio is too high. Therefore, people should focus on how to accurately classify the banking loans into performing and non-performing ones and how to control and prevent the resulting crisis. This paper deeply analyses China's NPLs problem for the current period, recognizes and classifies loans types by adopting decision trees, Naïve Bayes and support vector machine (SVM) methods. The experiment result found that the decision trees method can well identify the performing loans and non-performing loans; its accuracy is as high as 94%.*

***Keywords:*** *commercial bank, non-performing loan, decision tree, Naïve Bayes, support vector machine*

## 1. Introduction

Over the past decade, the credit quality of loan portfolios across most countries in the world remained relatively stable until the financial crises hit the global economy in 2007-2008. Since then, average bank asset quality deteriorated sharply due to the global economic recession. The non-performing loans (NPLs) problem has received great attention in both developed and developing countries [1-2]. With the deepening of reform and opening-up, Chinese market economy system has been gradually built. In this process, a bank is facing various risks from the market and itself. In these risks, credit risk is an important issue every bank faces to solve. Credit risk refers to the risk that a borrower will default on any type of debt by failing to make required payments [3]. For example, borrowers are unable to repay their debts cause deterioration of assets quality, depositors' mass withdrawals form run. If there are a large number of NPLs in a bank's asset business, that means borrowers can not repay when the loans mature, the bank will suffer a great loss. According to supervision statistical data issued by CBRC (China Banking Regulatory Commission), by the end of the first quarter in 2014, non-performing loans balance of China's commercial banks is 646.1 billion, an increase of 54.1billion compared with the beginning of the year. NPLs rate is 1.04%, 0.04 percent points higher than the beginning of the year. The NPLs rate has been rising in past 9 consecutive quarters and topped 1% for the first time. Bankings' bad loans situation can not be optimistic. In order to meet the needs of financial reform and development, reflecting loans risk conditions timely, accurately and completely, China initiated the five-grade judgment system (FGJS) in four domestic commercial banks and three policy banks since 1998. FGJS adopted by China's commercial banks began to replace the 'one overdue--two slacks' system which had been used for many years. In 2004, China required all

ISSN: 2005-4254 IJSIP

banks to implement FGJS. According to FGJS, Chinese commercial banks loans can be divided into normal, attention, substandard, doubtful and loss. This kind of loans classification method evaluate the loans quality in the light of the loans risk degree, which aims to control forehand, reduce credit risk loss, follow the conservatism principle. Therefore, it is critical for a commercial bank to classify the customers and decide which loan is good and which one is bad before loans form.

Academics and policy makers have been focused on the risk status of loans such as credit decision-making and credit risk management [2, 4, 5].Many studies built models to analysis NPLs. Marais (1984) examined several issues in the experimental design and empirical testing of classification models focused on the classification of commercial bank loans as an illustration [6]. Khemraj (2009) attempted to ascertain the determinants of non-performing loans in the Guyanese banking sector using a panel dataset and a panel regression model [7-8]. Greenidge built various models to forecast NPLs in the banking sector of Barbados [9]. Beck, Jakubik and Piloiu (2013) studied the macroeconomic determinants of NPLs ratios across 75 countries. According their research, real GDP growth, share prices, the exchange rate, and the lending interest rate are found to significantly affect NPL [10]. Cifter (2015) examines the effect of bank concentration on the NPLs for ten Central and Eastern European (CEE) countries [11].

Previous domestic studies always use financial ratios in the quantitative research of bank loans analysis because of their quantifiability and obtainability [12-14]. However, traditional financial analysis has some limitations. What's more, the credit enterprises are not in a closed system, they would be inevitably influenced and confined by the macroeconomic and market environments. Dai Xiaomin (2005) extended the Bayesian discriminate model and neural network model of credit ratings by considering non-financial ratios and macroeconomic factors. Empirical results indicated that by introducing industry-relative ratios and non-financial ratios (with lag of macroeconomic factors) into traditional models based on only financial ratios, the total classified accuracy and predictive power would be significantly improved, and the neural network approach has more classified accuracy than Bayesian discriminate model [11]. Khemraj (2009) found that both bank specific and macroeconomic factors impacts on the loan portfolios of commercial bank in Guyana [7]. Kester Guy (2011) used a series of bank idiosyncratic variables and macroeconomic factors to explain non-performing loans [15]. Abedola (2011) employed ARDL of Pesaran and Shin [16-17] to examine the effects of some macroeconomic variables which include industrial production index, interest rate and producer price index [18].

The remainder of the study is organized as follows. Section 2 describes our panel dataset and research methodology. In this section special attention is given to the motivation for selecting each variable included in our data mining model. Section 3 presents the empirical analysis, and Section 4 concludes with several policy implications.

## 2. Data and Methodology

### 2.1. Data and Data Preprocessing

In order to avoid oversampling and potential self selection bias, resulting in overestimation of model predictive ability, this study did not adopt same proprotion among FGJS samples. The number of normal loan samples is higher than that of non-performing loans in the adoped data set, which is correspond to the actual situation. This study uses the NPLs problem in China in 2013 as its empirical context. The data come from a commercial bank database in Harbin from January 2004 to March 2013. There are 96 features and 10415 instances in the data set. As aforementioned, both bank-specific variables and macroeconomic variables are determinants to influence the situation of bank loan. Therefore, we add annual inflation rate, real effective exchange rate (REER), and

 

annual growth in real GDP into original dataset. According to the bank manager's mark on each instance, the data is labeled with five categories. We follow the bank rule and label different types of loan as {0, 1, 2, 3, 4} to indicate normal, attention, substandard, doubtful and loss respectively.

Concentrated in the raw data, there are many feature variables which have nothing to do with the loan quality classification, such as bank loans customers' descriptions and bank loans related records. These feature variables should not be removed from the raw dataset. Some other data have an empty data, so before identification and classification, we first preprocess the bank loans data. Data preprocessing is shown in Figure 1.

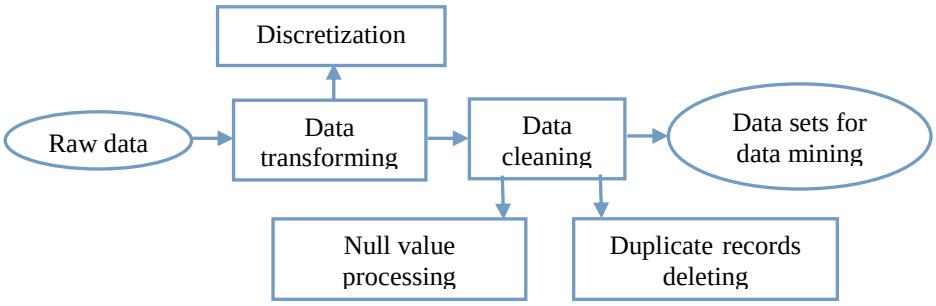


**Figure 1. Raw Data Preprocessing**

Four methods of data preprocessing methods are widely used: data cleaning, redundant data processing, data transforming and data reduction. Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, deduplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers believed of data. Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data. Textual data spellcheckers can be used to lessen the amount of mistyped words, but it is harder to tell if the words themselves are correct [19].

In this paper, we cleaned data with the following steps:

1. Deleting some features with missing data from the raw database. We set that we will remove a feature being without 5 samples.

2. Using the mean value properties of filling the missing "*custormers' ages*".

3. The use of global constants to replace the missing value. For examples, we use "null" to replace the missing value.

4. Washing the dirty data. In general, the heterogeneous data in the database data are not correct, often inevitably have incomplete, inconsistent, inaccurate and repeated data, called 'dirty data'. We unify heterogeneous data.

Notice that the data set has many more instances of normal class than other four types. There are no instances of doubtful class. In the view of label distribution, we find a problem that class imbalance is much significant. The quantity of four classes in raw data set is show in Table 1. In this situation, most of the classifier are biased and have poor performance of classification on minor classes.

**Table 1. Quantity of Four Classes in Raw Data Set**

| Classes | Quantity of raw dataset instances |
|---|---|
| 0 | 8768 |
| 1 | 323 |
| 2 | 577 |
| 4 | 225 |
| Total | 9893 |

We got training set of 1490 samples and testing set of 624 test samples by random selection from raw dataset. In the training set, there are 742 normal samples, 213 attention samples, 385 substandard samples and 150 loss samples. While in the testing set, there are 247 normal samples, 110 attention samples, 192 substandard samples and 75 loss samples. The quantity of four classes in training set and testing set is show in Table 2.

**Table 2. Quantity of Four Classes in Training Set and Testing Set**

| Classes | Quantity of training set instances | Quantity of testing set instances |
|---|---|---|
| 0 | 742 | 247 |
| 1 | 213 | 110 |
| 2 | 385 | 192 |
| 4 | 150 | 75 |
| Total | 1490 | 624 |

Then, we process feature selection. We use CPA-relief algorithm1to select features, seen [20]. Finally there are 31 features including 30 predictive variables and 1 objective variable in the data set. The results of feature selection are shown in Table 3. Sij signifies the predictive variables and C signifies the objective variable.

**Table 3. Predictive and Objective Variables for NPLs Feature Selection**

| | | | |
|---|---|---|---|
| $S_{i1}$ | *Customer number* | $S_{i16}$ | *Executed rate* |
| $S_{i2}$ | *Loan type* | $S_{i17}$ | *Overdue interest rate type* |
| $S_{i3}$ | *Payment* | $S_{i18}$ | *Extension mark* |
| $S_{i4}$ | *Contract amount* | $S_{i19}$ | *Received interest mark* |
| $S_{i5}$ | *Loan date* | $S_{i20}$ | *Table accumulative interest* |
| $S_{i6}$ | *Maturity period* | $S_{i21}$ | *Off-table accumulative interest* |
| $S_{i7}$ | *Due date* | $S_{i22}$ | *A class of overdue account* |
| $S_{i8}$ | *Last value date* | $S_{i23}$ | *Dull account* |
| $S_{i9}$ | *Interest bearing cycle* | $S_{i24}$ | *Doubtful account* |
| $S_{i10}$ | *Floating manner* | $S_{i25}$ | *Table normal interest account* |
| $S_{i11}$ | *Overdue interest* | $S_{i26}$ | *Off-table compound interest* |
| $S_{i12}$ | *Floating rate* | $S_{i27}$ | *Off-table non-accrual account* |
| $S_{i13}$ | *Principal account cancelled* | $S_{i28}$ | *Off-table interest-owned account* |
| $S_{i14}$ | *Interest account cancelled* | $S_{i29}$ | *Rate type* |
| $S_{i15}$ | *Executed compound interest rate* | $S_{i30}$ | *Charge amount* |
| *C* | *Loan nature* | | |

Data mining proposes several classification methods derived from the field of statistics and artificial intelligence. Three methods, which enjoy a good reputation for their classification capabilities, are employed in this research. These methods are decision tree, Naïve Bayes and SVM.

### 2.2. Decision Tree

Each decision or event may lead to two or more events that lead to different consequences. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.

 

A decision tree consists of 3 types of nodes:Decision nodes - commonly represented by squares;Chance nodes - represented by circles and End nodes - represented by triangles.

In the machine learning, a decision tree is a prediction model. It represents a mapping relationship between feature attributes and feature values. Every node in the tree represents a certain feature attribute, and each branch path represents a possible feature values. The leaf node represents the path's feature value from a decision node to the end node. A decision tree only has a single possible output. If you would want another output, independent decision trees can be setup to handle different outputs. A decision tree is often be used for analyzing data and doing predictions in data mining.

Among classification tools, decision trees have several advantages.

1. A decision tree is simple to understand and interpret. People are able to understand decision tree models after a brief explanation. In the process of learning, users don't need to have a lot of background knowledge. A decision tree can directly reflects the characteristic of data.

2. Decision trees have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes.

3. The calculation speed is fast and it is easy to transform classification rules.

4. Decision trees can be combined with other decision techniques.

In the other hand, decision trees also have some disadvantages. For data including categorical variables with different number of levels, information gain in decision trees are biased in favor of those attributes with more levels. Calculations can get very complex particularly if many values are uncertain and/or if many outcomes are linked.

Decision trees contain a number of different algorithms, mainly divided into three categories:

(1) Algorithms based on statistics, represented as CART. In this kind of algorithms, there are two branches for the non-end nodes.

(2) Algorithms based on information theory, represented as ID3 and C4.5. In this kind of algorithms, the number of determined by the one of sample classes.

(3) The algorithms represented as AID and CHAIN. In this kind of algorithm, the number of non-end nodes branches is between two and the number of sample classes.

In the late 1970s and early 1980s, J.Ross Quinlan, a machine learning researcher, developed a decision tree algorithm [21-22], called the ID3 (Interactive Dichotomiser3). Decision tree (C4.5), a modified ID3 algorithm, was developed by Quinlan in 1993 [23]. This paper constructed a classification and recognition model for commercial bank nonperforming loans based on C4.5. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = S_1, S_2, \cdots$ of already classified samples. Each sample $S_i$ consists of a p-dimensional vector $(x_{1,i}, x_{2,i}, \ldots, x_{p,i})$, where the $x_j$ represent attributes or features of the sample, as well as the class in which $S_i$ falls.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists.

C4.5 uses GainRatio(A) to measure and select attributes and uses SplitInfoA (D) to normalize the information gain. SplitInfoA(D) is similar to Info(D):

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \quad (1)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (2)$$

This research selected the attribute with maximum gain ratio as the split attribute.

### 2.3. Naïve Bayes

Bayesian classification is a statistical classification method. In machine learning, Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong naïve independence assumptions between the features [24-25]. This hypothesis is called class-condition assumptions. The working process of Naïve Bayes classification is as follows:

(1) Let D be the training set, represented by a vector $X = \{x_1, x_2, ...x_n\}$ representing some n features A1, A2,... An, (dependent variables). It assigns to this instance probabilities $P(C_i|X)$ for each of i possible outcomes or classes.

(2) Suppose there are m classes, C1, C2,... , Cm. Given a sample of X, the classifier will predict X is belong to the class with the highest posterior probability. That is to say, Naïve Bayes classification predict X belongs to the Ci, if and only if

$$P(C_i|X) > P(C_j|X) \qquad 1 \le j \le m, j \ne i$$

Then, we maximize $P(C_i|X)$. Ci with max $P(C_i|X)$ is call maximum posterior probability hypothesis. According Bayes theorem,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (4)$$

(3)Because P(X) is a constant for all classes, we may maximize $P(X|C_i)P(C_i)$. If the prior probability of classes is unknown, we usually assume classes probability is equally: $P(C_1) = P(C_2) = ... = P(C_m)$ Then, maximize $P(X|C_i)$ based on the prior probability. Otherwise, maximize $P(X|C_i)P(C_i)$.

Theoretically speaking, Naïve Bayes has the smallest error rate compared with all other classification algorithms.

### 2.4. Support Vector Machine

Support vector machine(SVM), introduced by Vapnik in 1995 [26-27], is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.

SVM is primarily a two-class classifier. The optimization criterion here is the width of the margin between the classes, *i.e.*, the empty area around the decision boundary defined by the distance to the nearest training patterns [28-29]. These patterns, called support vectors, finally define the classification function. Their number is minimized by maximizing the margin.

SVM maps the original training data to a higher-dimensional space with nonlinear mapping, presumably making the separation easier in that space. In the new dimensional space, the optimal linear separating hyperplane is searched. Using an appropriate higher-dimensional nonlinear mapping, two classes of data can be always divided by the hyperplane. To keep the computational load reasonable, the mappings used by SVM

 

schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function $k(x, y)$ selected to suit the problem [30].

SVM outperforms other classifiers in many fields. It is now one of the most popular tools and powerful learning algorithms. SVM trains with a learning method from optimization theory to find the optimal hyper plane in a high-dimensional space. It maps input vectors non-linearly spread on optimal hyper plane into a high dimensional feature space using various kernels. The learned examples that are closest to the optimal hyper plane are support vectors. Then support vectors are used to linearly separate feature spaces. Given a training set of instance-label pairs $(x_i, y_i)$ , i=1,2,…m, where $x_i \in R^n$ , $b \in R$ , $y_i \in \{1, -1\}$, and with the non-negative slack variables $\xi_i \geq 0$, the data points can be correctly classified by

$$< \omega \cdot x_i > + b \geq +1 - \xi_i, \text{ for } y_i = +1 \quad (5)$$

$$< \omega \cdot x_i > + b \geq -1 - \xi_i, \text{for } y_i = -1 \quad (6)$$

Where, $\omega$ is the weight of vector or normal vector of hyperplane; $b$ is the constant term.

The minimization function can be expressed as:

$$\underset{\omega, b, \xi}{Min} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{m} \xi_i$$

Subject to: $y_i (< w \cdot x_i > + b) + \xi_i - 1 \geq 0$ , $\xi_i \geq 0$

Several kernel functions, such as polynomial kernel, sigmoid kernel and radial basis kernel help the SVM in finding the optimal solution. The best kernel parameters, C and gamma, can be got with LIBSVM tools which will be adopted in this paper.

## 3. Results of Experiment

After confirmed the theme of data mining and preprocessed the experimental data, the next task is applying appropriate methods to do empirical analysis based on the mining aim according to the data, combined with the actual characteristics of experimental data preprocessing. In this paper, we constructed the commercial bank non-performing loan classification models by using software Weka 3.6. The specific steps are as follows.

(1)Establish data Resource Node and input a certain commercial bank loan database which has been preprocessed;

(2)Establish Filter node, filter parts of data according to the requirements of subsequent modeling nodes;

(3)Establish a Type Node, which means defined the variable type in the dataset, and set the variable “input” and “output” property. “Loan nature” is the output variable and others are input variables.

(4)Apply random sampling method of weka3.6 to set training dataset and testing dataset.

(5) Establish C4.5 decision tree node, Naïve Bayes node and SVM node and generate analysis node to analyse the related prediction results.

### 3.1. Results of Decision Tree Classification Model

This research constructed a decision tree to classify non-performing loans with training data produced by the sample nodes after segmentation. In this paper, C4.5 algorithm is used, the choice of the C4.5 node in the model area. Through the selection of experts, the global pattern of pruning, the misclassification loss, improve the decision tree model, in order to get the best results. Then, test by the test data set, the model is evaluated. The

results of decision tree classification model are shown in Table 4. Average classification accuracy rate is 94.3%. The classification accuracy rate of Class0 is 98.7%. The classification accuracy rate of Class1 is 76.8%. The classification accuracy rate of Class2 is 62.6%.The classification accuracy rate of Class4 is 67%.

**Table 4. Results of Decision Tree Classification Model**

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| | 0.987 | 0.257 | 0.961 | 0.987 | 0.974 | 0.927 | 0 |
| | 0.768 | 0.006 | 0.803 | 0.768 | 0.785 | 0.925 | 1 |
| | 0.626 | 0.016 | 0.782 | 0.626 | 0.695 | 0.903 | 2 |
| | 0.67 | 0.002 | 0.889 | 0.67 | 0.764 | 0.893 | 4 |
| Weighted Avg. | 0.943 | 0.224 | 0.94 | 0.943 | 0.94 | 0.924 | |

### 3.2. Results of Naïve Bayes Classification Model

The results of naïve bayes classification model are shown in Table 5. From Table 5, we can see that naïve bayes has good classification accuracy rate for the Class 0, but classifier cannot classify Class 2, Class 1 and Class 4 exactly.

**Table 5. Results of Naïve Bayes Classification Model**

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| | 0.858 | 0.405 | 0.943 | 0.858 | 0.898 | 0.794 | {0} |
| | 0.307 | 0.096 | 0.165 | 0.307 | 0.214 | 0.742 | {2} |
| | 0.591 | 0.043 | 0.316 | 0.591 | 0.412 | 0.896 | {1} |
| | 0.578 | 0.011 | 0.556 | 0.578 | 0.566 | 0.913 | {4} |
| Weighted Avg. | 0.811 | 0.367 | 0.868 | 0.811 | 0.835 | 0.797 | |

### 3.3. Results of SVM Classification Model

In this research, voting is performed in such a way:

Assume: Class0=Class1=Class2=Class4=0;

(Class0, Class1) -classifier, if Class0 win, Class0=class0+1; otherwise, Class1= Class1+1;

(Class0, Class2) -classifer, if Class0 win, Class0= Class0+1; otherwise, Class2= Class2+1;

...

(Class2, Class4) –classifer, if Class2 win, Class2= Class2+1; otherwise, Class4= Class4+1;

The decision is the Max(Class0, Class1, Class2, Class4).

The first step of constructing a SVM model is to construct the sample set according to analysis. Secondly, choose the type of kernel function. This article intends to adopt RBF function as inner product kernel function to build the model. The form of RBF function is:

$$K(x, x_i) = exp\left\{-\frac{\|x-x_i\|^2}{2\sigma^2}\right\} \quad (7)$$

Via cross-checking, this research determine that $\sigma^2$ = 4 , C = 10, then used Matlab6.5 tool kit to analyze.

The summary of SVM results is shown in Table 6.

Through analysis, the model can identify accurately 10341 loan samples from raw dataset. The identification accuracy rate is 99.2895%.

 

**Table 6. Results of SVM Classification Model**

| | | |
|---|---|---|
| Correctly Classified Instances | 10341 | 99.2895 % |
| Incorrectly Classified Instances | 74 | 0.7105 % |
| Kappa statistic | 0.0261 | |
| Mean absolute error | 0.0002 | |
| Root mean squared error | 0.0139 | |
| Relative absolute error | 32.6286% | |
| Root relative squared error | 99.965% | |
| Coverage of cases (0.95 level) | 99.2895% | |
| Mean rel. region size (0.95 level) | 1.3514% | |
| Total Number of Instances | 10415 | |

## 4. Discussion and Conclusions

The high non-performing loans of commercial Banks will lead to huge financial risk. Whether for the government or a commercial bank itself, it is necessary to recognize and predict a non-performing loan. But there is a lot of work to do when we artificially audit bank loans, also influenced by subjective factors. Data mining techniques, which have advanced classification and prediction capabilities, can facilitate auditors in accomplishing the task of non-performing loans detection. This paper construct three classification predict models of decision tree, naïve bayes and support vector machine for bank loans risk classification. Among of these three classification methods, the classification accuracy rate of decision tree is the highest. Based on the actual production data of the commercial bank, by using decision tree classification model analysis, obtained 0.943% correct classification rate.

Some limitations of the present study bear mentioning. First and foremost, bank loans actual data is difficult to be acquired. Although the models constructed in this research are based on a certain commercial bank, the models need to be tested in other banks loan dataset. Another limitation is that this research only uses three classification methods to recognize and predict the non-performing loans; some other clustering methods do not in use. These will be the main tasks of the next research.

## References

[1] A. Campbell, "Bank insolvency and the problem of nonperforming loans", Journal of Banking Regulation, vol. 9, no. 1, **(2007)**, pp. 25-45.
[2] J. Li and C. K. Ng, "The Normalization of Deviant Organizational Practices: The Non-performing Loans Problem in China", Journal of business ethics, vol. 114, no. 4, **(2013)**, pp. 643-653.
[3] D. Duffie and K. J. Singleton, "Credit risk: pricing, measurement, and management", Princeton University Press, **(2012)**.
[4] A. Campbell, "Bank insolvency and the problem of nonperforming loans", Journal of Banking Regulation, vol. 9, no. 1, **(2007)**, pp. 25-45.
[5] S. T. Li, W. Shiue and M. H. Huang, "The evaluation of consumer loans using support vector machines", Expert Systems with Applications, vol. 30, no. 4, **(2006)**, pp. 772-782.
[6] M. L. Marais, J. M. Patell and M. A. Wolfson, "The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications", Journal of accounting Research, **(1984)**, pp. 87-114.
[7] T. Khemraj and S. Pasha, "The determinants of non-performing loans: an econometric case study of Guyana", **(2009)**.
[8] G. Jimenez, V. Salas and J. Saurina, "Determinants of collateral", Journal of financial economics, vol. 81, no. 2, **(2006)**, pp. 255-281.
[9] K. Greenidge and T. Grosvenor, "FORECASTING NON-PERFORMING LOANS IN BARBADOS", Journal of Business, Finance & Economics in Emerging Economies, vol. 5, no. 1. **(2010)**.
[10] R. Beck, P. Jakubik and A. Piloiu, "Non-performing loans: What matters in addition to the economic cycle?", **(2013)**.
[11] A. Çifter, "Bank concentration and non-performing loans in Central and Eastern European countries", Journal of Business Economics and Management, vol. 16, no. 1, **(2015)**, pp. 117-137.
[12] R. Duchin and D. Sosyura, "Safer ratios, riskier portfolios: Banks′ response to government aid", Journal of Financial Economics, vol. 113, no. 1, **(2014)**, pp. 1-28.

[13] G. B. D. Xiaomin, "An Research on the Five-level Classification Model of Domestic Commercial Banks".
[14] J. Y. Uppal and I. U. Mangla, "Islamic Banking and Finance Revisited after Forty Years: Some Global Challenges", Journal of Finance, **(2014)**.
[15] K. Guy and S. Lowe, "Non-performing Loans and Bank Stability in Barbados", Economic Review, vol. 37, no. 3, **(2011)**, pp. 77-99.
[16] M. H. Pesaran and Y. Shin, "Generalized impulse response analysis in linear multivariate models", Economics letters, vol. 58, no. 1, **(1998)**, pp. 17-29.
[17] M. H. Pesaran, Y. Shin and R. J. Smith, "Bounds testing approaches to the analysis of level relationships", Journal of applied econometrics, vol. 16, no. 3, **(2001)**, pp. 289-326.
[18] S. S. Abedola, W. S. W. Yusoff and J. Dalahan, "An ARDL approach to the determinants of nonperforming loans in Islamic banking system in Malaysia", Kuwait Chap Arabian J Bus Manag Rev, vol. 1, no. 1, **(2011)**, pp. 20-30.
[19] J. M. Hellerstein, "Quantitative data cleaning for large databases", United Nations Economic Commission for Europe (UNECE), **(2008)**.
[20] Z. Yu, Y. Gang, Y. Guan and D. Yang, "International Journal of u- and e- Service, Science and Technology, vol. 8, no. 3, **(2015)**, pp. 29-42.
[21] J. R. Quinlan, "Simplifying decision trees", International journal of man-machine studies, vol. 27, no. 3, **(1987)**, pp. 221-234.
[22] J. R. Quinlan, P. J. Compton and K. A. Horn, "Inductive knowledge acquisition: a case study", Proceedings of the Second Australian Conference on Applications of expert systems. Addison-Wesley Longman Publishing Co., Inc., **(1987)**, pp. 137-156.
[23] J. R. Quinlan, "C4. 5: Programming for machine learning", Morgan Kauffmann, **(1993)**.
[24] Y. Lin, J. Wang and R. Zou, "An Improved Naïve Bayes Classifier Method in Public Opinion Analysis", Proceedings of the 4th International Conference on Computer Engineering and Networks. Springer International Publishing, **(2015)**, pp. 219-225.
[25] B. F. Z. Al-Bayaty and S. Joshi, "Comparative Analysis between Naïve Bayes Algorithm and Decision Tree to Solve WSD Using Empirical Approach", Lecture Notes on Software Engineering, vol. 4, no. 1, **(2016)**.
[26] P. B. Schiilkop, C. Burgest and V. Vapnik, "Extracting support data for a given task", **(1995)**.
[27] V. Vapnik, S. E. Golowich and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing", Advances in neural information processing systems, **(1997)**, pp. 281-287.
[28] K. Chen, W. Lu and J. Yang, "Support Vector Machine", **(2014)**.
[29] P. Tsyurmasto, M. Zabarankin and S. Uryasev, "Value-at-risk support vector machine: stability to outliers", Journal of Combinatorial Optimization, vol. 28, no. 1, **(2014)**, pp. 218-232.
[30] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, ", Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8. **(2007)**.

## Authors

**ZhangYu**, she received the Master of Management in Management Department(2004) from Harbin Institute of Technology(HIT). Now She is a teacher of Harbin university of science and technology and majoring in PhD of Management in Management Department from HIT. Her current research interests include different aspects of financial data mining and Machine learning.

**YuGuang**, she got a Bachelor Degree (1985) and a Master Degree (1990) of Engineer in power engineering department of Harbin Institute of Technology(HIT). She got a PhD of Management science and Engineering in HIT in 2007. She has been a professor in College of Management, a graduate and doctoral tutor in Harbin Institute of Technology since 2008. She is a peer review of many SCI Journal, such as information Science journal, IEEE Transactions on Reliability,et al. Her current research interests includes different aspects of Artificial Intelligence and Machine learning.

 