International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.23-34
http://dx.doi.org/10.14257/ijdta.2016.9.12.03

# Predicting Non Performing Loan of Business Bank with Data Mining Techniques

Wan Jie[1,2], Yue Zeng-lei[3], Yang Dong-hui[4], Zhang Yu[5], Liu Jiao[1], Liu Zhi[2] and Liu Jinfu[1]

[1]*School of Energy Science and Engineering, Harbin Institute of Technology, P. R. China, 150001*
[2]*Nangjing Qiuya Power Horizon Information Technology Company Limited, P. R. China, 210012*
[3]*Heilongjiang science and Technology Information Research Institute, P. R. China, 150001*
[4]*School of Economics and Management, Southeast University, P. R. China, 210096*
*Corresponding author: YANG Dong-hui, dhyang@seu.edu.cn*
[5] *School of Management, Harbin Institute of Technology, P. R. China, 150001*

***Abstract***

*The non-performing loans (NPL) prediction plays an important role in business bank. However, there is still a large gap between the requirement of prediction performance and current techniques. In this paper data mining approaches is used to predict the NPL. Both macroeconomic and bank-specific variables are collected to form the feature set firstly. Based on selected features, the study firstly applies single basic classifiers such as decision tree, k nearest neighbors and support vector machine (SVM) to model the problem of NPL. Bagging and AdaBoost are described in this paper as two different methods of multiple classifier fusion, to build prediction models. In this experiment, non-performing loans data with 96 features and 10415 instances of a business bank is collected. F-mean and The Area under the ROC Curve (AUC) are considered as metrics of classification performances. The results illustrate that multiple classifier fusion algorithms outperform single basic classifier. The model built by multiple classifiers fusion can produce better prediction results. Furthermore, the AdaBoost method performs much better than bagging method in processing NPL.*

***Keyword****: classification, class imbalance, data mining, non-performing loan, prediction*

## 1. Introduction

The problem of non-performing loan (NPL) is now extracting more attention of both developed and developing country governments [1]. Although NPLs have declined in Unite states, as of December 2012, lenders still held US$164 billion of distressed loans [2]. Europe's non-performing loans now have total been more than €1.2 trillion by 29 Oct 2013 [3]. Non-performing loan ratios of several countries can be found in Table 1. Those governments and finance sectors need to supervise bank loans more carefully. Many countries in East Asia are also suffering this problem, such as Japan, Korean and China. For example, in China, according to the data of China Banking Regulatory Commission, the balance of non-performing loans of Chinese commercial bank have been up to ￥592.1 billion by the end of December 2013, with an increase of ￥99.3 billion over the year of 2012. The NPL ratio of Chinese commercial bank was 1.0% in the year of 2013. However, the overall amount remained at a high level.

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2016 SERSC

**Table 1. Non-Performing Loan Ratios of Ten Countries from 2004 to 2012**

| Country | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
|---|---|---|---|---|---|---|---|---|---|
| US | 0.8 | 0.7 | 0.8 | 1.4 | 3 | 5.4 | 4.9 | 4.7 | 3.9 |
| Japan | 2.9 | 1.8 | 1.5 | 1.4 | 1.7 | 1.9 | 1.8 | -- | -- |
| France | 4.2 | 3.5 | 3 | 2.7 | 2.8 | 3.6 | 4.2 | -- | -- |
| Germany | 4.9 | 4 | 3.4 | 2.7 | 2.9 | 3.3 | -- | -- | -- |
| UK | 1.9 | 1 | 0.9 | 0.9 | 1.6 | 3.5 | 4 | -- | -- |
| Italy | 6.6 | 5.3 | 4.9 | 4.6 | 4.9 | 7 | 7.8 | -- | -- |
| Spain | 0.8 | 0.8 | 0.7 | 0.9 | 2.8 | 4.1 | 4.6 | 7.4 | 11.2 |
| Greece | 7 | 6.3 | 5.4 | 4.5 | 5 | 7.7 | 10.4 | 11.5 | 17.2 |
| Korea | 1.9 | 1.2 | 0.8 | 0.7 | 1.1 | 1.2 | 1.9 | 1.36 | 1.5 |
| China | 13.2 | 8.6 | 7.1 | 6.2 | 2.4 | 1.6 | 1.1 | 1.1 | 0.9 |

To a certain extent, high NPLs affect growth rate of Gross Domestic Product (GDP) in many cases. In the view of Hou (2007), NP Ls are a major cause of economic stagnation, thus hindering the economic growth and impairing economic efficiency [4]. Many empirical works show a negative relationship between NPLs and GDP growth [5]. In other point, increases in NPLs' rate are the main reason of reduction in earnings of banks. It makes financial situation even worse all over the world. As a reminder, the 1997 Asian Financial crisis triggered a wave of corporate bankruptcies and accumulation of NPLs in many East Asia countries.

Many researchers investigate the relationship of diverse economic elements and NPLs. Their aims are to identify the most significant determinants. In recent literature, binary logistic regression and various dynamic panel data techniques, such as panel vector autoregressive (panel VAR) model and autoregressive integrated moving average (ARIMA) models, have been popularly followed [6-8]. Those methods are proposed to test whether a specific economic or banking feature impact NPLs ratio. Louzis *et al*. argue that the type of determinants can be classified into macroeconomic (systematic) and bank-specific (idiosyncratic) variables [9].

In a new way, we aim to identify the potential NPLs using machine learning method before they lose control. Both macroeconomic and bank-specific features are collected in our experiment to reflect the credit status. To get more accurate identification, this paper adopts multiple classifier fusion technique which can integrate the advantages of base classifiers. The learning model can be used to predict NPLs for commercial bank managers.

This paper is organized as follows. Section 2 presents the related works of modeling NPL and NPL recognition. Section 3 introduces the methodology of multiple classifier fusion on NPL. Section 4 presents the experiments used to evaluate the effectiveness of the fusion strategies. Section 5 makes our concluding remarks.

## 2. Related Works

### 2.1. Modeling Non-Performing Loan

Majority of previous works focus on how to build a good model to analysis NPL. A linear regression framework was used to examine three determinants (nominal interest rates, inflation and GDP growth) of NPL by Chase *et al* (2005) [10]. Greenidge *et al*. (2010) added loan growth and the relative size of banks into those three elements to study the relationship of 5 banks [11]. They argued that bank sector control those numbers to reduce loans. Financial-restraint model were used to control deposit and lending rates to create rent incentives in several countries, such as China and Japan. Also, a two-period

 Copyright ⓒ 2016 SERSC

overlapping generations (OLG) model with a bank sector was presented by Barseghyan [12]. In this paper, the author analyzed the effect when a delay of government control and suggested bank sector should subtract resources from investment financing.
Another branch of modeling NPL is dynamic approach. The Markovian structure was suggested to forecast losses on a liquidating long term loan portfolio by Smith and Lawrence in 1995 [13]. But it was inappropriate in many common cases because their restrictive assumptions. Multivariate forecast model was more practical approach. Two variants of a dynamic panel framework were utilized to forecast the NPL ratio in Guy *et al*'s work [14]. Using panel data techniques, researchers wanted to account for the time persistence in the NPL structure, the pooled OLS, fixed effects and random effects models were the most popular techniques.

### 2.2. Features of Non-Performing Loans

Totally, features of non-performing loans can be classified as macroeconomic and bank-specific determinants. Macroeconomic variables are systematic factors that reflect economic situation of the whole country. Inflation, GDP growth and the money stock are used as predictors of macroeconomic environment [15]. More commonly, researchers use interest rates or the lending rates rather than money stock to evaluate the banking currency. In some works, the unemployment rate, which has a negative relation with inflation, is used as a systematic variable [9].

Many idiosyncratic variables are also very important factors to affect NPL ratio. Because each bank has different policies to improve its performance of risk management, bank-specific variables cannot be ignored. Salas and Saurina (2002) found that the bank size was a valid variable to control bank's loan [16]. Greenidge and Grosvenor (2010) followed this view and augmented the relationship between Loan growth and size of five banks [11]. A final set of bank-specific variables is consist of bank's efficiency, bank size, moral hazard and credit expansion in the work of Belgrave *et al* [5]. As considering the lenders, the firm size, CEO's moral identity, ownership type of their enterprise are also relative with loan lose [1].

### 2.3. Data Mining of NPL

Most of previous works use traditional statistics to build prediction models. Recently, some researchers apply the method of data mining to feature the NPL variables and build a classifier. Many works have shown that data mining techniques are effective and efficient in finance fields. In the techniques of data mining, they considered all determinants as equal features. Jiménez *et al* (2007) featured the equivalent hazard rate, the growth in real gross domestic product, general economic, market and technological developments to analyze credit risk of individual bank loans [17]. Ravi *et al* (2007) collected 54 input variables in their dataset to predict the performance of 1000 community banks [18]. The previous mentioned macroeconomic and bank-specific determinants all can be used as features to build a classifier in this paper.

The prediction of NPL using data mining techniques is supervised method. We should know the labels of each instance. Five labels are more popular used in the non-performing loan that can be classified into passed, special Mention, substandard, doubtful, virtual loss and Loss (Unrecoverable) [9]. After selecting good features that give more contributions, they used the classifier to predict new NPL. Evolutionary nearest neighbor classifier (ENN), decision tree, case-based reasoning (CBR), support vector machine (SVM), artificial neural network (ANN) and back propagation neural network (BPN) have been used to build a model for analyzing credit risk and other finance problems. More literatures of classifiers used in previous research are summarized in [19]. In this paper, we apply several single classifiers to model the problem of non-performing loan. More importantly, multiple classifier fusion methods are used to find a higher classification results.

Copyright ⓒ 2016 SERSC

# 3. Methodology

## 3.1. Basic Classifiers

To predict the performance of business bank, some classifiers are good tools, such as K-nearest neighbors (KNN) [20], decision tree and support vector machine (SVM) [18]. In this part, we provide an overview of those three classifiers.

KNN is among the simplest of all machine learning methods. It is instance-based learning used for classification or regression. For classification, KNN algorithm weights the contribution of the neighbors. The more near neighbors contribute more to the average result than the more distant cases. K is a constant defined at first step. Specifically, for a new point, it needs to find k nearest samples in the training dataset and calculate the distances between the new sample point and all samples in the training dataset. Majority voting will give the new point a label.

Decision tree can produce easy-to-interpret rules which also be represented as sets of if-then rules. It produce learned tree by splitting and pruning criterion. The classification starts at the root node of the tree which has two descendants: the left sibling and the right sibling according to splitting criterion. The top-down fashion goes to deep branches to grow the tree. In each splitting step, algorithms choose the point that can move to the child branches until terminal node. Totally, previous target functions are two categories: entropy criterion and gini criterion. ID3 and C4.5 use an information entropy evaluation function as the selection criteria, while CART prefers the gini impurity criterion.

Support vector machine (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

$$< w \cdot x_i > +b \geq +1 - \xi_i, \text{ for } y_i = +1$$

$$< w \cdot x_i > +b \geq -1 - \xi_i, \text{ for } y_i = -1$$

Where,

$w$ is the weight of vector or normal vector of hyperplane;

$b$ is the constant term.

The minimization function can be expressed as:

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

Subject to:

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

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.

## 3.2. Strategy of Multiple Classifier Fusion

Using single classifier to model the NPL of business bank makes a new way to predict the probability of bad loans. The data mining techniques can save people out of manually

 Copyright ⓒ 2016 SERSC

checking bad loans.

Since single classifier cannot always give good enough accuracy, combining the decisions of different models, which also known as multiple classifier fusion, is logically put forward to tackle the diversity of training data. Bagging and Boosting are known as two different strategies that how to decide the fusion of a given set of complementary classifier. Next part, we describe those two methods in detail.

### 3.2.1. Bagging

Bagging (Abbreviation from Bootstrap aggregating) is a randomly resampling-based technique and can improve performance for unstable learning algorithms [21]. Supposed that m classes in label set, bagging equally samples instances from original training dataset $D_n = \{(x_i, y_j) | 1 \le i \le n, 1 \le j \le m\}$. And individual models are built separately. The bagging algorithm can be expressed as follows and the process is drawn in Figure1.

Step 1: In each iteration t, t=1,…T, sample n instances with replacement from training dataset.

Step 2: Apply the learning algorithm to the sample and get k models.

Step 3: Each decision $d_{ki}$ of k models receives equal weight ($\omega_1 = \omega_2 = \ldots = \omega_k$) and belongs to $\{0,1\}$.

Step 4: Store the resulting model final decision H based on majority voting, which is calculated by $y_j = \sum_{k=1}^{K} d_{kj}, \; j = 1, \ldots, m$.

Step 5: Predict class of new instance using model.

### 3.2.2. AdaBoosting

Different from bagging method, boosting technique samples instances according to sampling error rate. The training cases that are difficult to classify is believed to be more important. Boosting weights a model's contribution by its confidence. Multiple models developed in sequence by assigning higher weights for those instances. Therefore, boosting algorithm can produce classifiers that are more accurate on fresh data than ones generated by bagging algorithm. The final prediction function is formulated by weighted voting for the classification problem. In this way, a weak classifier finally produces better and better results by iteratively learning.

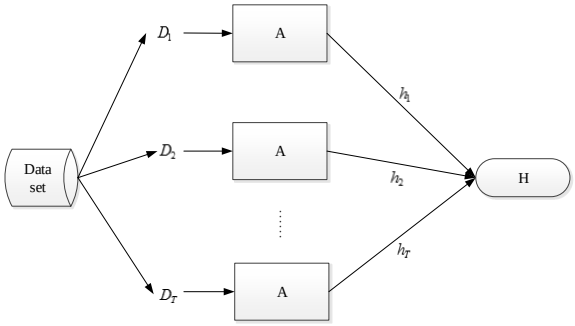



**Figure 1. The Process of Bagging Method**

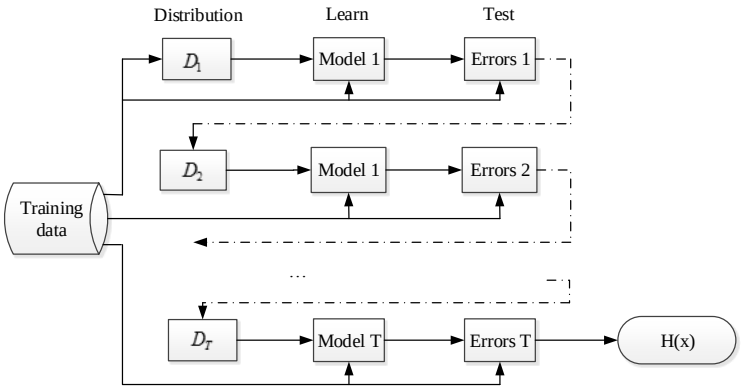



**Figure 2. The Process of AdaBoost Method**

There are many variants on the idea of boosting. AdaBoost (Adaptive boosting) is one of the best performance boosting algorithm and widely used in real application, such as e-mail filtering and text classification [22]. A training set of N examples is used to generate the first model. Then repeat it again to generate the next model in which step the misclassified cases get higher weights. At last, one combines predictions from individual models weighted of accuracy of the models. The process of AdaBoost can be seen in Fig 2.

## 4. Experiment

### 4.1. Dataset

The present study collects loan information of a business bank data in Harbin from January 2004 to March 2013. This bank lends money not only to individual but also to company, farm, hospital, media group. Since the number of individual loan far less than organization, we focus the latter one whose prediction is more meaningful. In the data set, there are 96 features and 10415 instances totally. As aforementioned, both bank-specific variables and macroeconomic variables are determinants to influence the situation of bank loan. Therefore, we add inflation, GDP growth and the money stock (M0 and M1) into original dataset. The features contain accounts number, accounting agencies, types and classes of loan, payments, contract amount, start date, due date, exercise rate, floating interest rate, GDP growth and so on. 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 passed, special Mention, substandard, doubtful, virtual loss and Loss respectively.

### 4.2. Data Pre-processing

In the original data set, many features are organization descriptions and bank memories that have no use to classify. Those features should not be included. After data pre-processing, a neat feature set contains 21 features. Based on this feature set, there are lots of empty data should be deleted firstly. Finally, the data set has 9893 instances. In the view of label distribution, we find a problem that class imbalance is much significant. The data set has many more instances of passed class than other four types. Rushi Longadge (2013) argued that classifier predicts everything as major class and ignores the minor class [23]. In this situation, most of the classifier are biased and have poor performance of classification on minor classes. In order to eliminate this problem, we should pick the data of major class which equals to the number of other classes. We randomly sample a number of the major class equaled to sum of other classes. We do our experiment of data pre-processing using MATLAB 7.0. The randomly sampling command in MATLAB is "randperm(n)". The number is equal to the average of sum of other three classes that is 375.

 Copyright ⓒ 2016 SERSC

Totally, the data set contains 1500 instances to build classification model

### 4.3. Results and Analysis

In this paper, the classification of non-performing loan of business bank is handled with the help of WEKA. The first step is how to select features that make the classification performance of NPL much better. In this work, information gain method, which performs better on feature selection than other methods [24], is adopted to operate feature extraction. We follow rule that information gain should be larger than 0.0025. We find that the 14th and 16th features (Types of normal lending rates and signs of pre-charge interest) have no contribution to the results of information gain. Therefore, the feature set includes 19 features at last.

Secondly, 10-fold cross validation is chosen to randomly segment the dataset into 9 training set and 1 test set. Decision tree (J48), k-nearest neighbor (IBK) and SVM are applied as basic classifiers. Bagging and AdaBoost are also carried out as "meta" classifiers.

To measure the results of classifiers, precision and recall are two common metrics. While, F-mean is a integrate value that coordinate those two metrics. Its formula is expressed as followed.

$$F-mean=\frac{2*precsion*recall}{precision+recall} \tag{1}$$

The Receiver Operating Characteristic (ROC) is another useful metric used to measure true positive rate (TPR) and false positive rate (FPR). A ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between them. The closer to the left upper corner of the unit square of the ROC curve, the better the classification results. The Area under the ROC Curve, known as the AUC, is the summary of ROC Curve that gives a better understanding of real values and predicted values. AUC is the most often used method to compare the metric of different classification models.

We choose the F-mean and AUC to capture the performance of classification models. The results are collected in Table 2. As shown in Figure 3, a snapshot clearly displays the curves of those two metrics. The observation from the indicators of classification results are following:

For basic classifiers, the decision tree has the best F-mean immediately followed by 1-NN. As considering AUC, 3-NN makes the true predicted values much more. But SVM which known as a strong classifier has not the better performance on the data of NPL of business bank.

K-NNs have relative better results than SVM in the metric of F-mean. Only compare three different classifiers of nearest neighbors, 1NN is the best of three KNNs in metric of F-mean but the worst in the metric of AUC. From Figure3, we can see that the curves of F-mean and AUC among three classifiers of nearest neighbors have negative correlation.

In the case of Bagging and AdaBoost, their results are much better than basic classifiers both in the metric of F-mean and AUC. They have 5.1% and 7.8% enhancements than the worst of basic classifiers in the metric of F-mean, while 8.2% and 8.5% enhancements than the worst of basic classifiers in the metric of AUC. Therefore, we can say that multiple classifier fusion methods build better prediction models.

If contrasting the last two classification methods, we also find that AdaBoost outperforms Bagging method. The F-mean of AdaBoost is higher than Bagging with 2.7% increase. The value of AUC of AdaBoost is a little bit larger than Bagging method. In detail, we want to know the exact classification results on each type of NPL. Therefore, we also collect results of diverse classes of non-performing loans which can be seen in Table3. The F-mean on each class of AdaBoost is better than that of Bagging. The AUC on Class 0, Class 1 and Class 2 of AdaBoost's are all better than Bagging method. The AUC value of Bagging and AdaBoost is very near on Class 4. As a whole, AdaBoost algorithm is much

Copyright ⓒ 2016 SERSC

better than bagging algorithm in processing the data of non-performing loans of business bank.

**Table 2. The Classification Results of Different Classifiers**

| Classifier | TP | FP | Precision | Recall | F-Mean | AUC |
|---|---|---|---|---|---|---|
| J48 | 0.833 | 0.069 | 0.832 | 0.833 | 0.833 | 0.908 |
| 1-NN | 0.831 | 0.071 | 0.831 | 0.831 | 0.831 | 0.88 |
| 2-NN | 0.815 | 0.076 | 0.821 | 0.815 | 0.811 | 0.907 |
| 3-NN | 0.8 | 0.084 | 0.8 | 0.8 | 0.798 | 0.919 |
| SVM | 0.805 | 0.072 | 0.814 | 0.805 | 0.809 | 0.894 |
| Bagging | **0.851** | **0.065** | **0.855** | **0.851** | **0.849** | **0.962** |
| AdaBoost | **0.877** | **0.056** | **0.88** | **0.877** | **0.876** | **0.965** |

**Table 3. Classification Results on Different Classes of Bagging and Adaboost**

| Method | Class | Number | TP | FP | Precision | Recall | F-Mean | AUC |
|---|---|---|---|---|---|---|---|---|
| Bagging | Class 0 | 375 | 0.963 | 0.054 | 0.855 | 0.963 | 0.906 | 0.992 |
| | Cass 1 | 322 | 0.867 | 0.051 | 0.824 | 0.867 | 0.845 | 0.964 |
| | Class 2 | 578 | 0.83 | 0.103 | 0.834 | 0.83 | 0.832 | 0.941 |
| | Class 4 | 225 | 0.693 | 0.006 | 0.951 | 0.693 | 0.802 | 0.961 |
| AdaBoost | Class 0 | 375 | 0.965 | 0.038 | 0.894 | 0.965 | 0.928 | 0.991 |
| | Cass 1 | 322 | 0.87 | 0.035 | 0.873 | 0.87 | 0.871 | 0.972 |
| | Class 2 | 578 | 0.879 | 0.099 | 0.848 | 0.879 | 0.863 | 0.951 |
| | Class 4 | 225 | 0.738 | 0.007 | 0.949 | 0.738 | 0.83 | 0.951 |

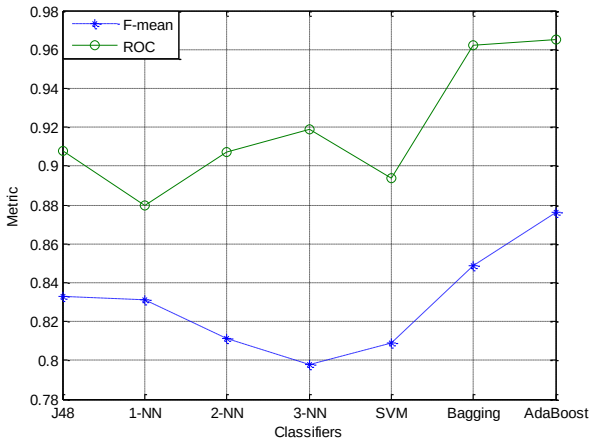



**Figure 3. Curves of F-mean and AUC among different Classifiers**

## 5. Conclusion

In this paper, both macroeconomic and bank-special variables are considered as determinants of non-performing loans. To predict the NPL, three single classifiers are applied to model the problem of NPL of business bank. Also, bagging and AdaBoost algorithms are two important multiple classifier fusion methods which also are applied to the NPL data. The classification results show that data mining techniques can good at

 Copyright ⓒ 2016 SERSC

predicting NPL of business bank. In the other point, bagging and AdaBoost methods have better performance than basic classifiers such as decision tree, k nearest neighbor and SVM. Moreover, based on the metrics of F-mean and AUC on four classes of NPL, the AdaBoost algorithm performs much better than bagging algorithm. Further experiments will be carried out to precisely find more determinants as features for predicting NPL. Other multiple classifier fusion methods will be tried to check whether or not they give more accurate prediction results in the future.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No.71501040 and 2014 Nanjing "321 Introduction Plan" Candidate Project.

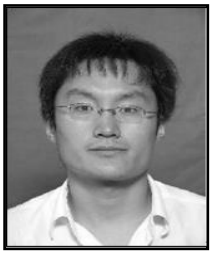
## References

[1] 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.
[2] "U.S. Banks' Bulk Loan Sales to Grow, Aid in NPA Reduction", Fitch Ratings, www.fitchratings.com/gws/en/fitchwire/fitchwirearticle/U.S.-Banks'-Bulk?pr_id=782581, **(2013)**.
[3] A. Cárdenas, "The Spanish Savings Bank Crisis: History, Causes and Responses", IN3 Working Paper Series, vol. 9, **(2013)**, pp. 4-41
[4] Y. Hou and D. Dickinson, "The non-performing loans: Some bank-level evidence", // the 4th International Conference on Applied Financial Economics, Samos Island, Greece, vol. 8, **(2007)**, pp. 105-137.
[5] A. Belgrave, K. Guy and M. Jackman, "Industry Specific Shocks and Non-Performing Loans in Barbados", Review of Finance & Banking, vol. 4, no. 2, **(2012)**, pp. 123-133.
[6] M. Xu, "Resolution of non-performing loans in China", Glucksman Fellowship, vol. 4, **(2005)**, pp. 1-62.
[7] K. Kauko, "External deficits and non-performing loans in the recent financial crisis", Economics Letters, vol. 115, no. 2, **(2012)**, pp. 196-199.
[8] Y. S. Lee and L. I. Tong, "Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming", Knowledge-Based Systems, vol. 24, no. 1, **(2011)**, pp. 66-72.
[9] D. P. Louzis, A. T. Vouldisn and V. L. Metaxas, "Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios", Journal of Banking & Finance, vol. 36, no. 4, **(2012)**, pp. 1012-1027.
[10] C. Karen, K. Greenidge, W. Moore and D. Worrell, "Quantitative Assessment of a Financial System: Barbados", International Monetary Fund, Monetary and Financial Systems Department, **(2005)**, pp. 5-76.
[11] K. Greenidge and T. Grosvenor, "Forecasting Non-performing loans in Barbados", Journal of Business, Finance & Economics in Emerging Economies, vol. 5, no. 1, **(2010)**, pp. 79-108.
[12] L. Barseghyan, "Non-performing loans, prospective bailouts, and Japan's slowdown", Journal of Monetary Economics, vol. 57, no. 7, **(2010)**, pp. 873-890.
[13] S. L. Douglas and E. C. Lawrence, "Forecasting losses on a liquidating long-term loan portfolio", Journal of Banking & Finance, vol. 19, no. 6, **(1995)**, pp. 959-985.
[14] K. Guy and S. Lowe, "Non-performing Loans and Bank Stability in Barbados", Economic Review, vol. 37, no 3, **(2011)**, pp. 77-99.
[15] K. A. Demirgüç and E. Detragiache, "Cross-country empirical studies of systemic bank distress: a survey", National Institute Economic Review, vol. 192, no. 1, **(2005)**, pp. 68-83.
[16] V. Salas and J. Saurina, "Credit risk in two institutional regimes: Spanish commercial and savings banks", Journal of Financial Services Research, vol. 22, no. 3, **(2002)**, pp. 203-224.
[17] S. Ongena, J. L. Peydro-Alcalde and J. Saurina, "Hazardous times for monetary policy: what do twenty-three million bank loans say about the effects of monetary policy on credit risk", Centre for Economic Policy Research, **(2007)**, pp. 6-25.
[18] V. Ravi, H. Kurniawan and P. N. K. Thai, "Soft computing system for bank performance prediction", Applied Soft Computing, vol. 8, no. 1, **(2008)**, pp. 305-315.
[19] S. W. Lin, Y. R. Shiue and S. C. Chen, "Applying enhanced data mining approaches in predicting bank performance: A case of Taiwanese commercial banks", Expert Systems with Applications, vol. 36, no. 9, **(2009)**, pp. 11543-11551.
[20] G. Yu, D. H. Yang and H. X. He, "An automatic recognition method of journal impact factor manipulation", Journal of Information Science, vol. 37, no. 3, **(2011)**, pp. 235-245.
[21] N. C. Oza, "Online bagging and boosting", Systems, man and cybernetics, 2005 IEEE international conference on. IEEE, vol. 3, **(2005)**, pp. 2340-2345.
[22] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, **(2005)**, pp. 358-372.
[23] R. Longadge and S. Dongre, "Class Imbalance Problem in Data Mining: Review", International Journal

Copyright ⓒ 2016 SERSC 

of Computer Science and Network (IJCSN), vol. 2, no. 1, **(2013)**, pp. 1305.
[24] D. H. Yang and G. Yu, "A method of feature selection and sentiment similarity for Chinese micro-blogs", Journal of Information Science, vol. 39, no. 4, **(2013)**, pp. 429-441.

## Authors



**Jie Wan,** He received B. E. and M. E. from Harbin Institute of Technology, Harbin, China in 2007 and 2010, respectively. Then studying for his PHD in Harbin Institute of Technology. His research interests are focused on data mining algorithm and its application in Engineering. He has published more than 10 journal and conference papers in these domains.

**Zenglei Yue**, She received Bachelor degree from Beijing Forestry University in 2007, and Master degree from Chinese Academy of Sciences in 2012.After that she has been working in the Institute of Science and Technology of Heilongjiang Province. Now she is an associate researcher in the Institute. He research focus on science and technology policy and decision. She has published 3 papers in these domains.

**Donghui Yang**, He is an assistant professor at Department of management science and engineering, Southeast University. He received his Ph.D degree of Management in 2014 and Master degree of Management in 2009 from Harbin institute of Technology. His major is management science and engineering. He received his Bachelor degree of Management from Harbin Engineering University in 2006. His research interesting includes data mining, text mining, social network analysis and recommender system.

**Zhang Yu**, He received bachelor and master degrees from Harbin Institute of Technology, Harbin, China in 2002 and 2004 respectively. After that she started working in Harbin University of Science and Technology. Now she is an associate professor in the economic college. Her research interests are focused on machine learning and financial auditing. She has published more than 10 journal and conference papers in these domains.

**Liu Jiao**, She received B. E. and M. E. from Harbin Institute of Technology, Harbin, China in 2012 and 2014, respectively. Then studying for her PHD in Harbin Institute of Technology. Her research interests are focused on data mining algorithm and its application in Engineering.

 Copyright ⓒ 2016 SERSC

**Liu Zhi**, He graduated from Heilongjiang Academy of Social Sciences in 2002.After that he has been working in Zhengtong Electronic Technology Co., LTD and other S&T companies respectively. He joined Nanjing Power Horizon Science and Technology Co., LTD in 2014 as vice manager, in charging of Marketing Department. His research interests are focused on big data analysis.

**Jinfu Liu**, He received B. E., M. E. and PhD degrees from Harbin Institute of Technology, Harbin, China in 2000, 2002 and 2008, respectively. After that He started working with Harbin Institute of Technology as assistant professor. Now He is an associate professor with this institute. His research interests are focused on intelligent modeling and fault diagnosis. He has published more than 20 journal and conference papers in these domains.

Copyright ⓒ 2016 SERSC

 Copyright ⓒ 2016 SERSC