International Journal of u- and e- Service, Science and Technology
Vol.8, No.3 (2015), pp.29-42
http://dx.doi.org/10.14257/ijunesst.2015.8.3.03

Feature Selection of Nonperforming Loans in Chinese Commercial Banks

Zhang Yu[1, 2], Yu Gang[1], Guan Yongsheng[1, 3] and Yang Donghui[1]

[1]*School of management, Harbin Institute of Technology, Harbin, Heilongjiang, China*
[2]*School of Economics, Harbin University of Science and Technology, Harbin, Heilongjiang, China*
[3]*LongJiang Bank, Province Heilongjiang, Harbin, Heilongjiang, China*
hester0524@163.com; yug@hit.edu.cn; 572387616@qq.com; ydh95130@163.com

Abstract

In recent years, huge amounts of nonperforming loans (NPLs) of commercial banks have become one of the biggest obstacles constraining reform and development in Chinese commercial banks. Finding a way to control the banks' NPLs is a core issue that it continues to be explored and researched in the finance. In this paper, PCA and relief algorithm in data mining methods were adopted to extract and analyze NPLs characteristics in commercial banks through contrasting the performing and nonperforming loans records, based on the predecessors' literatures. In this paper, a bank's loans data with 96 features and 10415 samples is collected. At last, we construct nonperforming loans of commercial banks classification model. Our research is very important for capturing warning signal timely, detection of NPLs and sound operation of commercial banks.

Keywords: *commercial banks; nonperforming loans; feature selection; PCA-relief algorithm*

1. Introduction

In recent years, the huge nonperforming loans (NPLs) of commercial banks have become one of the biggest obstacles to restrict commercial banks in China's reform and development. In the reform process, China's banking sector accumulated a huge amount of nonperforming loans [1]. In 1999-2000, these assets management companies -- China Cinda, China Huarong, China Orient and China Great Wall -- stripped of four state-owned commercial banks 1.3939 trillion yuan of bad loans and NPL ratio dropped by nearly 10 percent. While according to the five-grade judgment system (FGJS) of loan assets, state-owned commercial banks' bad loans ratio was still as high as 34.18%. In recent years, reduction of NPLs in Chinese state-owned commercial banks has been effective. By the end of 2001, Chinese state-owned commercial banks NPLs come to 25.37% of the loan balance. According to the latest survey, the top ten listed banks' NPLs was 492.9 billion yuan at the end of 2012, an increase of 64.7 billion yuan; NPLs ratio was 0.95%, down 0.01 percentage points over the same period of last year. With the deep-going of banking and related departments 'reform, we see the progress made in the control of NPLs in commercial banks. NPL ratio showed China's commercial banks has been in decline since 2003, while but falling range vary in different years and intervals. The biggest of bad loans drop represented a decline from the original 17.8% to 10% between 2003 to 2006 years, down of the international warning line. Since 2008, NPLs ratio decreased slowly, especially in the last two years. Hovering around 1%, the NPLs ratio in some quarters

ISSN: 2005-4246 IJUNESST

appears even rising. The explanation could be that the commercial bank's NPL ratio is now close to the limit at this stage of economic development. Only the healthy and sustainable development of economic can likely decrease non-performing loan ratio further.

Banks' high NPL is a difficult problem facing many countries. According to the experience of Japan, NPLs have a significant hazard. The biggest hazard of high rate of non-performing loans is that it can affect banks' supporting capacity to the economic. Banks of China lent extremely carefully recently, because too many bad loans have exercised a great influence on the banks' lending capacity. If you rely on the issue of base money to solve the problem of bad loans, it can easily lead to inflation. If you're careless with bad loans, a high increase in non-performing loans will lead to social and moral risk. If you increase efforts to deal with nonperforming loans, corporate chains may cause bankruptcies, increasing financial risks and social crisis. The presence of huge nonperforming loans of the banking industry will inevitably jeopardize the safe operation, reduce the bank's ability to resist risks and seriously lead to bank failures and even financial crisis. All of the countries of the world have a great concern on the NPLs of commercial banks. We want to find the features of NPLs, monitor and properly dispose of non-performing loans of commercial banks.

Based upon the research foundation of existing NPLs of commercial banks, we extracted the characteristic feature of NPLs and compared a large number of NPLs and performing loans of commercial banks recorded with data mining method. This can help China's commercial banks determine the nature of loans earlier to prevent the occurrence of NPLs as a point of referenced. Our research is very important for capturing warning signal timely, detection of non-performing loans and sound operation of commercial banks. The rest of the paper is organized as follows. In Section 2, we define the nonperforming loans, the scope of this study, and review of existing literature. In Section 3, we describe the research methods used in this paper, PCA-Relief algorithm. In Section 4, we experiment PCA-Relief algorithm to select the features of NPLs in a commercial bank loans data. Finally, Section 5 concludes the paper and sketches some future research directions.

2. The Definition and Literature Review

2.1 The Definition of Non-performing Loans of Banks

Formally, a NPL or a bad loan is defined to be a debt instrument (loan) whose contractual interest and principal payments are difficult to collect [2]. NPLs include that there are signs that it is impossible for the borrower to repay the original loan agreement business principal and interest of loans banks formed.

Definition of NPLs is built on the basis of loan classification. In China, NPLs were ever defined as doubtful loans, sluggish loans or overdue loans, according to People Bank of China issued 'Credit Rules' in July 17, 1995. Overdue loans are overdue (including extension after expiration) that the loans are not repaid (excluding bad loans and bad loans). Sluggish loans are overdue (including extension after expiration) beyond 2 years (including 2 years) and are still not repaid. Sluggish loans also include these ones that although not due or overdue for less than 2 years but have stopped production operations, lending project cessation (excluding doubtful loans). Doubtful loans are loans borrowers and guarantors have gone bankrupt and fail to pay off the loans.

At present, bank loans classification approach are represented by Australia two-grade judgment system (TGJS) and America five-grade judgment system (FGJS). The TGJS divided loans into normal and abnormal loans. In TGJS, abnormal loans are defined as bad loans. Under FGJS, a bank's loan portfolio can be classified into five major categories--pass, special mention, substandard, doubtful and loss, according to the degree

of risk of loans [3]. Many countries and regions carry out FGJS, including America, Southeast Asia countries and transition countries of Eastern Europe. They formally incorporated FGJS into the credit management system. China executed FGJS since 2002.

2.2 Related Works

Since the focus of our study is non-performing loans' characteristics in China's commercial banks, we briefly review traditional NPLs' research in Sect 2.2.

2.2.1 Causes and Countermeasures of NPL: Majority of Chinese previous works focus on the causes and countermeasures of NPLs. In general, the NPLs of state-owned commercial banks are due to the government, enterprises, banks' behavior mechanism, and also has the profound historical reasons and institutional reasons. Zhou Xiaochuan, the president of the central bank, pointed out that the causes of the non-performing loans of state-owned commercial banks mainly come from five aspects: the government's direct intervention, supporting to the state-owned enterprises, local administrative environment and legal environment, the leading industry structure adjustment and state-owned commercial banks' business problems.

Many Chinese scholars thought institutional reasons caused China's non-performing loans, because of the transmission mechanism in economic foundation leading to credit risk. Parnes (2012) divided loans classification error into type I and type II. He demonstrated the operational risk associated with type II errors in typical lending decisions made by banks. His paper also explored several forms of operational risk associated with the corresponding type I errors. Parnes' model is built on the ordinary problems, analyze their expected failure rates, compare their functionality, and further propose additional complexities within these models for general use by banks and other lending institutions [4].

2.2.2 Recognizing and predicting NPL: According to the existing literature, the main method to predict and recognize the bad loans of commercial banks is regression analysis, including simple linear regression, multiple regression, fuzzy regression, Logistic model and probit model. Some other scholars used the method of clustering analysis and decision tree [5-7]. The details are shown in Table 1.

Table 1. Research on the Recognitions in NPL

Methods	Time	Author	Conclusions
linear regression	2005	Chase *et al*	A linear regression framework was used to examine three determinants of NPLs [8].
Regression analysis and cluster analysis	2010	Yu Mingye *et al*	Customer risk characteristics that caused bad loans were summarized.
Financial-restraint model	2010	Greenidge *et al*	Financial-restraint model were used to control deposit and lending rates to create rent incentives in several countries [9].
Decision tree	2010	Chi Qingyun *et al*	Personal customers formed risks were analyzed to find out the characteristics of risk customers.
Fuzzy regression analysis	2011	Yang Jianhui *et al*	The method for reducing the balance of bad loans of banks was stated.
Logistic model	2011	Chen Muzi *et*	There is a relationship between the

		al	extreme zero recovery rate and GDP.
Fourier Flexible Functional form	2011	B. Maggi *et al*	The Fourier Flexible Functional form (FFF) also was adopted to estimate components of NPL [10].
Probit model	2012	Liu Ying	Probit model was established between the non-performing loans and three explanatory variables--customer credit rating scale, customer and customer economy type.
Dynamic panel data estimator	2012	Louzis	Specification to model NPL was adopted with real GDP growth rate, unemployment rate and real lending rates which had a strong effect on the level of NPL[11].
Regression analysis and discriminant analysis	2013	ZengYan	The relationship between macro-economic development indexes and China's non-performing loan rate was analyzed qualitatively with financial situation of enterprises those had a loan.

3. Methodology

In machine learning and statistics, feature selection, also known as attribute selection, is the process of selecting a subset of relevant features for use in model construction [12]. Feature selection is one of the key problems in pattern recognition. Feature selection results directly affect the accuracy of classifiers and generalized performance. According to [13], the process of feature selection is shown in Figure 1.

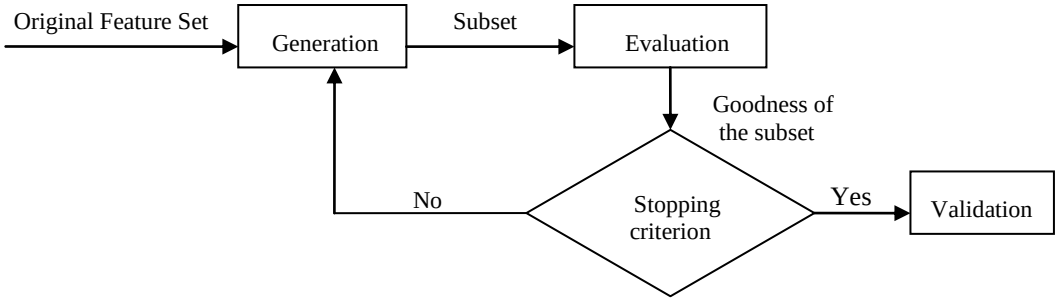


Figure 1. The Process of Feature Selection

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the two main categories of feature selection algorithms: wrappers and filters [14].

Wrapper methods use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally

intensive, but usually provide the best performing feature set for that particular type of model.

Filter methods use a proxy measure instead of the error rate to score a feature subset. This measure is chosen to be fast to compute, whilst still capturing the usefulness of the feature set. Common measures include the point wise [15]. Pearson product-moment correlation coefficient, inter/intra class distance or the scores of significance tests for each class/feature combinations [16, 17]. Filters are usually less computationally intensive than wrappers, but they produce a feature set which is not tuned to a specific type of predictive model.

In this section, a two-phase algorithm is presented. The proposed algorithm is presented in Figure 2. In the first phase, this algorithm proposes to collect data and calculate PCA indicators. It is noted that NPL is considered as one of the PCA indicators. Then, PCA is run to find a weight for each of PCA variables to be used for integrating all PCA indicators and calculate an overall efficiency score for each bank-year. The efficiency calculation mechanism of applied PCA is presented in Section 3.1. The second phase involves the use of Relief arithmetic to find the characteristic of NPLs. We use Weka software to execute the machine selection. The principle of Relief is presented in Section 3.2.

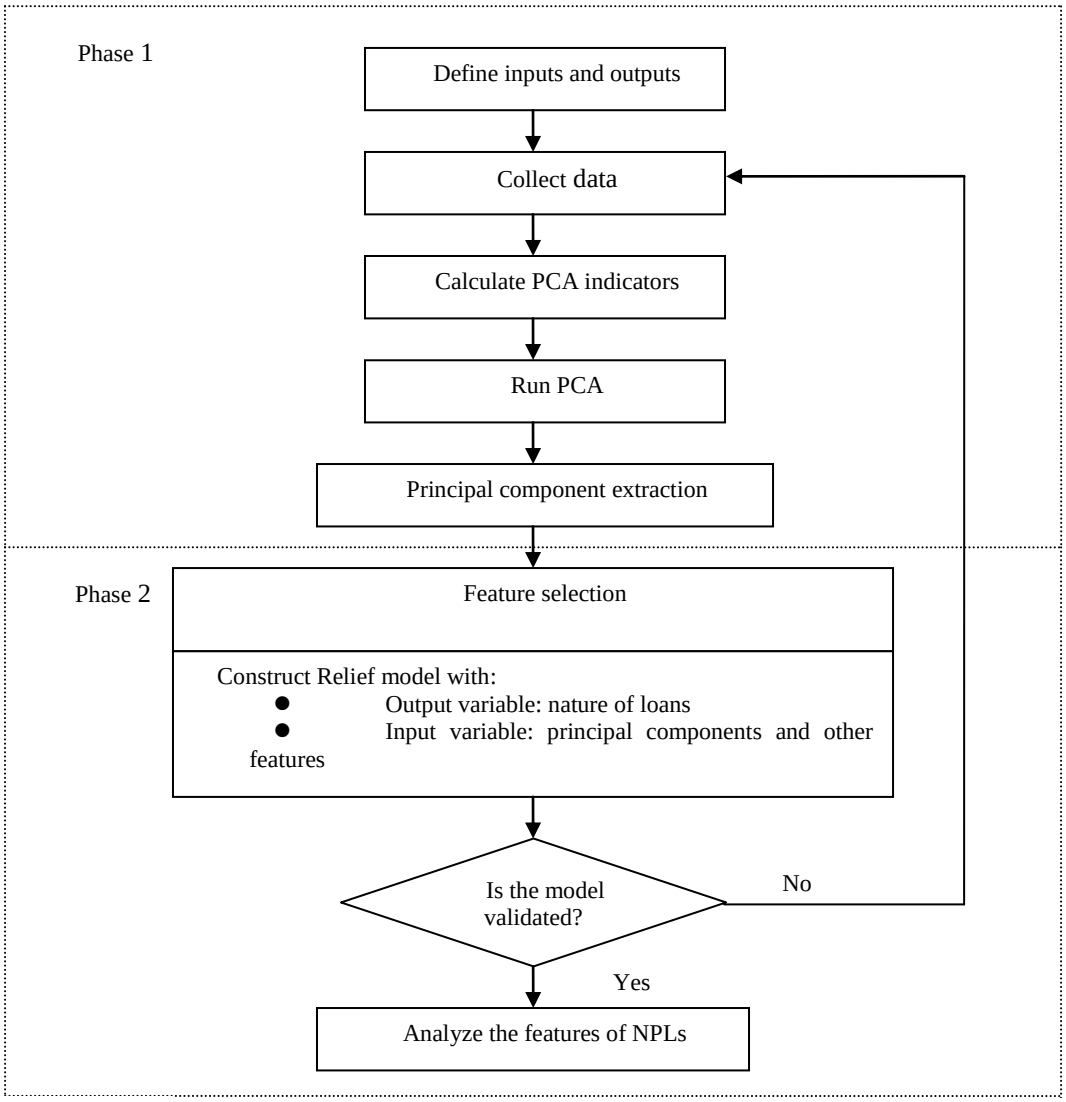


Figure 2. The PCA-Relief Algorithm

3.1. PCA

Principal component analysis (PCA), also called Karhunen-Loève transform, was proposed by Turk M and Pentland in 1990[18, 19]. PCA is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. From a mathematical point of view, this is a dimensionality reduction technique. PCA is widely used in multivariate statistics such as factor analysis[2].The basic idea of PCA is to find out a set of fewer unrelated and comprehensive index Fm to replace the original index X1,X2,…,XP (for example, p indexes). Then how to extract the comprehensive index set to extremely reflect the original variable while the new index separated from each other (unrelated to keep information do not overlap)? The specific steps of principal component analysis are as follows:

Step 1: Calculate the sample mean vector and covariance matrix.

Calculate the covariance matrix of sample data:

$$\sum = (s_{ij})\, p \times p \quad (1)$$

In which $s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_j)(x_{kj} - \overline{x}_j)$ i,j=1,2,…,p

Step 2: Calculate characteristic value λ_i and corresponding orthogonal unit vectors a_i.

The first m larger characteristic values, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m > 0$, are the variance corresponding to principle component variables. a_i Corresponding to λ_i is the coefficients of the original variables. And the principle component Fi is:

$$F_i = a_i' X \quad (2)$$

The variance (information) contribution rate of principle components can reflect the size of information:

$$a_i = \lambda_i / \sum_{i=1}^{m} \lambda_i \quad (3)$$

Step 3: Select the principal components by defining

The accumulative variance (information) contribution rate of G (m) can be calculated to determine the selected principle components.

$$G(m) = \sum_{i=1}^{m} \lambda_i / \sum_{k=1}^{p} \lambda_k \quad (4)$$

The first m principal components may be selected by satisfying, e.g., G (m) > 90%. *i.e.*, the first m principal components account for 90% contribution to the total sample variance [20].

3.2 Relief Algorithm

Relief is a feature selection algorithm used in binary classification proposed by Kira and Rendell in 1992[17]. Its strengths are that it is not dependent on heuristics, requires only linear time in the number of given features and training instances, and is noise-tolerant and robust to feature interactions, as well as being applicable for binary or continuous data; however, it does not discriminate between redundant features, and low numbers of training instances fool the algorithm. Kononenko *et al*. proposed some updates to the algorithm (RELIEF) in order to improve the reliability of the probability approximation, make it robust to incomplete data, and generalizing it to multi-class problems [21].

Relief examines the features' differentiate ability by comparing features' differences between the same-class samples and different-class samples. If the difference among

same-class samples is small and large among different-class samples, the feature have strong ability to distinguish.

Take a data set $S=\{s_1,s_2,\ldots,s_m\}$ with m instances of p features, $s_i=\{s_{i1},s_{i2},\ldots,s_{ip}\}$, $1\le i\le m$, belonging to two known classes $C=\{C_l,C_2\}$. Within the data set, each feature should be scaled to the interval [0, 1] (binary data should remain as 0 and 1). The difference in two samples of s_i and s_j ($1\le i\ne j\le m$) in feature t ($1\le t\le p$) is defined as:

If feature t is a scalar type,

$$diff\ (t, s_i, s_j) = \begin{cases} 0 & s_{it} = s_{jt} \\ 1 & s_{it} \neq s_{jt} \end{cases} \tag{5}$$

If feature t is a numeric type,

$$diff\ (t, s_i, s_j) = \left| \frac{s_{it} - s_{jt}}{\max_t - \min_t} \right| \tag{6}$$

In which, $\max_t$ and $\min_t$ respectively means the maximum and minimum of feature t in the sample.

Firstly, we pick up a random sample s_i ($1\le i\le m$) from sample set and the samples closest to s_i (by Euclidean distance) from each class. The closest same-class sample is called Hit, and the closest different-class sample is called Miss. Then, update the weight vector wt. using Hit and Miss according to the formula (7)

$$w_t = w_{t-1} - diff\ (t, s_i, Hit)/r + diff\ (t, s_i, Miss)/r \tag{7}$$

Thus the weight of any given feature decreases if it differs from that feature in nearby instances of the same class more than nearby instances of the other class, and increases in the reverse case.

After r iterations, divide each element of the weight vector by m. This becomes the relevance vector. Features are selected if their relevance is greater than a threshold τ.

Kira and Rendell's experiments [17] showed a clear contrast between relevant and irrelevant features, allowing τ to be determined by inspection. However, it can also be determined by Cebysev's inequality for a given confidence level (α) that a τ of $1/\sqrt{\alpha \times m}$ is good enough to make the probability of a Type I error less than α, although it is stated that τ can be much smaller than that.

$W=\{w_1, w2\ldots WP\}$ is the feature weight vector finally. We can rank the features in descending order according to w_t ($1\le t\le p$), showing the ability to distinguish the categories from strong to weak sequence. The specific algorithm is as follows:

Algorithm: Relief algorithm

1. All feature weights is 0 and T is null;
2. for i=1 to m do
   1) select a sample R randomly;
   2) Find nearest neighbor samples H from the same sample set and find the nearest neighbor samples M from different samples set;
   3) for A=1 to N do
      $W(A)=W(A)-diff(A,R,H)/m + diff(A,R,M)/m$
3. for A=1 to N do
   If $W(A)\ge\delta$
      Added feature A to T
   End

4. Experiment

4.1 Dataset

We collect loan information of a commercial bank data in Harbin from January 2004 to March 2013. This bank lends money not only to individual but also to company, farm, and hospital and 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 we known, 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, contract amount, start date, due date, exercise rate, floating interest rate, and 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, and 4} to indicate pass, special mention, substandard, doubtful and loss respectively. Database of bank loans to the organization is shown in Figure 3.

A	B	C	D	E	F	G	H	I	J	K
客户号	贷款类	发放金额	合同金额	起息日	存	到期日	上次计息	计	浮动方	逾期计息
0000020186	5050	2500000.0	2500000.0	20040927	001	20050623	20041221	1	0	4
0000031965	5540	10000000.0	10000000.0	20041203	003	20061129	20060921	1	0	4
0000028145	5090	3000000.0	3000000.0	20040930	001	20050928	20050928	1	0	4
0000022288	5050	150000000.0	150000000.0	20040826	001	20050825	20050621	1	0	4
0000004247	5250	560000.0	560000.0	20030924	003	20050324	20071221	1	0	4
0000000528	5010	20000000.0	20000000.0	20040828	001	20050828	20041221	1	0	4
0000020688	5010	7000000.0	7000000.0	20040930	003	20060728	20060621	1	0	4
0000020918	5050	10000000.0	10000000.0	20040927	001	20050525	20050321	1	0	4
0000028137	5010	20000000.0	20000000.0	20041223	001	20051221	20051221	1	0	4
0000015374	5250	3600000.0	3600000.0	20040324	001	20050318	20050321	1	0	4
0000014871	5130	20000000.0	20000000.0	20040920	001	20050909	20050621	1	0	4
0000022288	5050	30000000.0	30000000.0	20040827	001	20050825	20050621	1	0	4

Figure 3. A Bank's Loan Data Table (Partial)

4.2 Data Pre-processing

Data pre-processing is an important step in the data mining process. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time.

In the original data set, many features are organization descriptions and bank memories that have no use to classify. Some features' data is obviously error and some others representation methods are not unified. For example, loan officer of some sample is data code, while some others are represented as text name. Those features should not be included. We remove them.

After data pre-processing, a neat feature set contains 31 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 original data set, loan nature is respectively defined as 5 grades: *pass, special mention, substandard, doubtful* and *loss*. Because the purpose of this study is to identify the characteristics of non-performing loans, we classify loan nature into two grades as *normal* and *non-performing* loans. Therefore, the data is reclassified. We define the *pass, special mention, substandard* loans as *normal* loans, while *doubtful* and *loss* as *non-performing loans* data. We transform a multi-classification problem into two-class classification problem.

4.3 Results and Analysis

Our experimental task is to seek and analysis the characteristics of NPLs, based the type of bank loans that already artificially set. Therefore, we set the objective variable (C) as *the nature of loan*. There are 31 prediction variables (S_i) shown in Table 2.

Table2. Prediction Variables and Objective Variables for 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*		

Table 3 presents some descriptive statistics of the collected data.

Table 3. Descriptive Statistics of the Collected Data

Variable	N	Min	Max	mean	SD
S_{i1}	10415	1017840829	2000860508	2.00E9	1.344E7
S_{i2}	10415	5000	256590	136737.75	99418.813
S_{i3}	10415	0	549426724	8946317.38	1.890E7
S_{i4}	10415	0	549426724	10036689.74	2.174E7
S_{i5}	10415	19950308	20130228	20105627.02	21228.873
S_{i6}	9893	1	106	13.22	33.257
S_{i7}	10176	19960307	20221227	20115926.83	20952.438
S_{i8}	10174	20040921	20130228	20113038.46	19306.783
S_{i9}	10414	0	6	.83	1.624
S_{i10}	10012	0	1	.08	.275
S_{i11}	10174	0	6	3.34	.785
S_{i12}	10415	-44	161	26.31	23.417
S_{i13}	6	2.E16	2.E16	2.36E16	6.269E14
S_{i14}	6	2.E16	2.E16	2.36E16	6.269E14
S_{i15}	10415	0	22	1.63	4.263
S_{i16}	10415	0	22	8.06	1.808
S_{i17}	10415	210	220	210.08	.900
S_{i18}	10415	0	2	.97	.380

S_{i19}	10415	0	1	1.00	.035
S_{i20}	10415	0	358995	209.88	6299.384
S_{i21}	10415	0	4345619	4683.71	95866.919
S_{i22}	1079	1.E15	9.E16	1.92E16	1.393E16
S_{i23}	437	1.E15	9.E16	1.83E16	1.670E16
S_{i24}	210	1.E15	9.E16	2.08E16	2.215E16
S_{i25}	2811	1.E15	9.E16	2.12E16	1.476E16
S_{i26}	489	1.E15	9.E16	1.90E16	1.353E16
S_{i27}	217	1.E15	2.E16	1.58E16	8.642E15
S_{i28}	234	1.E15	2.E16	1.91E16	7.749E15
S_{i29}	10415	0	100	1.56	4.675
S_{i30}	10411	0	780000	1094.46	17652.180

In Phase 1, we use software SPSS17.0 to execute PCA. After standardized transformation, the correlation matrix can be obtained according to the specific eigenvalues. The PCA results are presented in Table 4.

Table 4. Results of PCA

Component	Initial Eigenvalues			Extraction Sums of squared loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.480	23.579	23.579	4.480	23.579	23.579
2	2.361	12.429	36.008	2.361	12.429	36.008
3	1.788	9.410	45.418	1.788	9.410	45.418
4	1.570	8.265	53.683	1.570	8.265	53.683
5	1.247	6.564	60.246	1.247	6.564	60.246
6	1.132	5.959	66.205	1.132	5.959	66.205
7	1.013	5.332	71.537	1.013	5.332	71.537
8	1.008	5.305	76.842	1.008	5.305	76.842
9	.978	5.149	81.991			
10	.863	4.542	86.534			
11	.715	3.765	90.299			
12	.547	2.880	93.179			
13	.421	2.215	95.394			
14	.331	1.745	97.139			
15	.201	1.059	98.198			
16	.112	.588	98.786			
17	.107	.565	99.351			
18	.071	.374	99.725			
19	.052	.275	100.000			
Extraction Method: Principal Component Analysis.						

According to the principle of eigenvalues>1 PCA extract 8 principal components with the cumulative contribution of 76.842%.

In Phase 2, we use Weka 3.6 tools (Select attributes interface) to execute relief algorithm. We select Relief algorithm in attribute evaluator and Ranker in search method. Set the parameters as numNeighbours = 10, sample Size = -1, seed = 1, sigma = 2. In order to ensure the accuracy of the feature selection, we choose ten-fold cross certification to test. The results of relief algorithm are presented in Table 5.

Table 5. Average Merit of Features

feature	Average merit	feature	Average merit	feature	Average merit
S_{i1}	0	S_{i9}	0.05	S_{i17}	0.008
S_{i2}	0.013	S_{i10}	0.024	S_{i18}	0.009
S_{i3}	0.007	S_{i11}	0.055	S_{i19}	0.001
S_{i4}	0.009	S_{i12}	0.047	S_{i20}	0
S_{i5}	0.034	S_{i13}	0	S_{i21}	0
S_{i6}	0.031	S_{i14}	0	S_{i22}	0.003
S_{i7}	0.016	S_{i15}	0.008	S_{i23}	0.001
S_{i8}	0.023	S_{i16}	0.037		

The relationship of features' weight can be show as follows:

$S_{i11}< S_{i9}< S_{i12}< S_{i16}< S_{i5}< S_{i6}< S_{i10}< S_{i8}< S_{i7}< S_{i2}< S_{i18}< S_{i4}< S_{i15}< S_{i17}< S_{i3}< S_{i22}< S_{i19}< S_{i23}$

As we set the threshold of features' weight as 0.01, the selected features are S_{i11}、S_{i9}、S_{i12}、S_{i16}、S_{i5}、S_{i6}、S_{i10}、S_{i8}、S_{i7} and S_{i2}. The results in Table5show that S_{i11}'*Overdue interest'* is the most important influence. It is indicated that one can judge whether a loan is a NPL though *overdue interest* firstly, followed by *Interest bearing cycle* and *Floating rate*. Features S_{i1}, S_{i3}, S_{i14}, S_{i20} and S_{i21} have an average weight of 0, indicating that these features irrelevant to the loan nature. We should focus on the specific features such as '*Overdue interest*'. The following is the analysis of it.

S_{i11} named as *overdue interest*. Contract Law 207 regulate that 'when the borrower fails to repay the loan, he should pay overdue interest in accordance with the contract or relevant rules of the state'. Overdue loan interest calculation method mainly includes two kinds: one is according to the contract agreed; the other is calculated in accordance with the relevant regulations of the people's bank.

4.4 Expansion test

Based on selected ten features, this paper use decision tree C4.5 to classify the original data. The classification results are shown in Table 6.

Table 6. Classification Results of C4.5

	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	

There are 4 types of loans in the original data set namely {0, 1, 2, 4}. The experimental results show that based on feature selection, 98.7% of {0} is classified correctly and the overall average accuracy rate is 94.3%. It is indicated that the selected features can distinguish the NPLs from the loans data effectively.

5. Conclusion

The feature selection of NPLs of commercial bank is an important and widely studied topic. It can help the bank sector and manager to make right decision to avoid heavy losses. The major contribution of this paper is to apply PCA-relief algorithm to build NPLs feature selection models. To extract the feature of NPLs, PCA and relief algorithms are dimension reduction methods. The experimental results show that the feature of Overdue interest, Interest bearing cycle, floating rate, executed rate, Loan date, Maturity period, Floating manner, Last value date, Due date and Loan type can classify the loan nature rather well. Through the results of this research, commercial banks can determine the nature of a loan earlier to reduce bad loans rate and prevent the occurrence of bad loans to avoid credit risk.

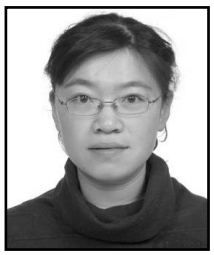
At the end of the experiment, we use decision tree C4.5 to classify the nature of loans. Further experiment will be carried out to find better classification method for predicting NPLs. In addition, in the expansion test, we find that the difference of accuracy rate is very large among the nature of loans. In the view of label distribution, we find a problem that class imbalance is much significant. The data set has many more instances of pass class than other four types. Future research will be launched to solve this problem.

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] F. Hajialiakbari, M. H. Gholami and J. Roshandel, "Assessment of the effect on technical efficiency of bad loans in banking industry: a principal component analysis and neuro-fuzzy system", Neural Computing and Applications, vol. 23, no. 1, **(2013)**, pp. 315-322.
[3] K. Guy, "Non-performing loans", Research and Economic Analysis, vol. 37, no. 1, **(2011)**, p. 10.
[4] D. Parnes, "Modeling operational risk for good and bad bank loans", Journal of Operational Risk Volume, vol. 7, no. 4, **(2012)**, p. 13.
[5] Z. Zhang and M. Cao, "Research of credit risk of commercial bank's personal loan based on CHAID decision tree", Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2nd International Conference, IEEE, **(2011)**.
[6] Z. Zenglian, "Research of Default Risk of Commercial Bank's Personal Loan Based on Rough Sets and Neural Network", Intelligent Systems and Applications (ISA), 3rd International Workshop, IEEE, **(2011)**, pp. 1-4.
[7] M. He, N. Liu and E. SIia, "Discrimination for non-performing loans recovery: a method of support vector machines based on wavelet transform", Information Science and Engineering (ISISE), International Symposium, IEEE, **(2010)**, pp. 88-92.
[8] 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.
[9] K. Greenidge and T. Grosvenor, "Forecasting Non perfoming loans in Barbados", Journal of Business, Finance & Economics in Emerging Economies, vol. 5, no. 1, **(2010)**, pp. 79-108.
[10] B. Maggi and M. Guida, "Modeling non-performing loans probability in the commercial banking system: efficiency and effectiveness related to credit risk in Italy, Empirical Economics, vol. 41, no. 2, **(2011)**, pp. 269-291.
[11] D. P. Louzis, A. T. Vouldis 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.
[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", The Journal of Machine Learning Research, vol. 3, **(2003)**, pp. 1157-1182.
[13] M. Dash and H. Liu, "Feature selection for classification", Intelligent data analysis, vol. 1, no. 3, **(1997)**, pp. 131-156.
[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", The Journal of Machine Learning Research, vol. 3, **(2003)**, pp. 1157-1182.

[15] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization", ICML, vol. 97, **(1997)**, pp. 412-420.
[16] G. Forman, "An extensive empirical study of feature selection metrics for text classification", The Journal of machine learning research, vol. 3, **(2003)**, pp. 1289-1305.
[17] K. Kira and L. A. Rendell, "A practical approach to feature selection", Proceedings of the ninth international workshop on Machine learning, Morgan Kaufmann Publishers Inc., **(1992)**.
[18] M. Turk and A. Pentland, "Face processing: Models for recognition", Advances in Intelligent Robotics Systems Conference, International Society for Optics and Photonics, **(1990)**.
[19] L. Jing, "Feature extraction of machine sound using wavelet and its application in fault diagnosis", NDT & E International, vol. 34, **(2001)**, pp. 25-30
[20] J. Zhu, "Data envelopment analysis vs. principal component analysis: An illustrative study of economic performance of Chinese cities", European Journal of Operational Research, vol. 111, no. 1, **(1998)**, pp. 50-61.
[21] I. Kononenko, E. Šimec and M. R. Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF", Applied Intelligence, vol. 7, no. 1, **(1997)**, pp. 39-55.

Authors



Zhang Yu. 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.

Yu Guang. 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 Journals, such as information Science journal, IEEE Transactions on Reliability, *et al.* Her current research interest includes different aspects of Artificial Intelligence and Machine learning.