TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
ISSN 2029-4913 / eISSN 2029-4921

2014 Volume 20(2): 292–311
doi:10.3846/20294913.2014.913275

MCDM APPROACH TO EVALUATING BANK LOAN DEFAULT MODELS

Gang KOU[a], Yi PENG[b], Chen LU[b]

[a]*School of Business Administration, Southwestern University of Finance and Economics, Chengdu, China*
[b]*School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, 610054, P. R. China*

Received 25 March 2013; accepted 10 November 2013

Abstract. Banks and financial institutions rely on loan default prediction models in credit risk management. An important yet challenging task in developing and applying default classification models is model evaluation and selection. This study proposes an evaluation approach for bank loan default classification models based on multiple criteria decision making (MCDM) methods. A large real-life Chinese bank loan dataset is used to validate the proposed approach. Specifically, a set of performance metrics is utilized to measure a selection of statistical and machine-learning default models. The technique for order preference by similarity to ideal solution (TOPSIS), a MCDM method, takes the performances of default classification models on multiple performance metrics as inputs to generate a ranking of default risk models. In addition, feature selection and sampling techniques are applied to the data pre-processing step to handle high dimensionality and class unbalancedness of bank loan default data. The results show that *K*-Nearest Neighbor algorithm has a good potential in bank loan default prediction.

Keywords: default risk model, bank loan, machine learning, multiple criteria decision making (MCDM), model selection.

Reference to this paper should be made as follows: Kou, G.; Peng, Y.; Lu, C. 2014. MCDM approach to evaluating bank loan default models, *Technological and Economic Development of Economy* 20(2): 292–311.

JEL Classification: D81, H81, G32.

Corresponding author Yi Peng
E-mail: *pengyi@uestc.edu.cn*

Introduction

Loans are often the major type of credit risk for banks. Over the years, many quantitative default risk models have been developed to assess the default risk of individual and corporate obligors and help banks to better manage credit risks and allocate economic capital (Bastos 2010).

According to the underlying techniques, default risk models and algorithms can be broadly classified into traditional statistical category and intelligence or machine-learning category (Kumar, Ravi 2007). Examples of traditional statistical techniques include simple univariate analysis (Beaver 1966), multiple discriminant analysis (MDA) (Altman 1968), multiple regression (Meyer, Pifer 1970), linear discriminant analysis (Altman 1973), logit model (Martin 1977; Ohlson 1980), goal programming (Srinivasan, Kim 1987), logistic regression (Wiginton 1980; Gilbert *et al.* 1990), quadratic discriminant analysis (QDA) (Banks, Prakash 1994), Bayesian analysis (Li 1999), Bayesian network models (Shumway 2001), and quadratic programming (Tseng, Lin 2005). Machine-learning category includes techniques like recursive partitioning algorithm (Marais *et al.* 1984; Frydman *et al.* 1985), neural network models (NN) (Tam, Kiang 1992), case-based reasoning (Buta 1994; Bryant 1997), decision trees (Henley 1995), *k*-nearest-neighbour (KNN) (Henley, Hand 1996), genetic programming (Varetto 1998; McKee, Lensberg 2002), rough sets (McKee 1998), neuro-fuzzy techniques (Gorzalczany, Piasta 1999), data envelopment analysis (Cielen *et al.* 2004), ensembles (Doumpos, Zopounidis 2007), soft computing (Ravi, Pramodh 2008), multi-criteria convex quadratic programming (Peng *et al.* 2008a), classification and regression trees (CART) (Khandani *et al.* 2010). Reviews and comparative studies on credit risk models can be found at Altman (1984), Rosenberg and Gleit (1994), Hand and Henley (1997), Galindo and Tamayo (2000), Kumar and Ravi (2007), Crook *et al.* (2007), and Verikas *et al.* (2010).

Although comparative analysis of default classification models has been conducted intensively, the findings of these studies do not always agree with each other. A default classification model that outperforms another using one set of dataset and performance measures may be disadvantaged under different circumstances. For example, Fernandez and Olmeda (1995) reported that the neural network model provided generally better results than CART-based decision tree algorithms in default prediction, whereas Galindo and Tamayo (2000) concluded that CART decision tree model was superior to neural networks in mortgage loan default prediction.

While important advances have been made in developing default classification models, less progress has been made in default model evaluation and selection. Lopez and Saidenberg (2000) proposed evaluation methods for credit risk models based on cross-sectional simulation. Sobehart *et al.* (2001) developed several performance metrics and a validation framework for quantitative default risk models. Stein (2002) discussed validation strategies for default prediction models. Medema *et al.* (2009) implemented a practical methodology to validate credit risk models. In addition to the inconclusiveness associated with default classification models, the limited availability of bank loan data presents another major challenge in assessing and selecting default classification models.

The objective of this study is to propose an approach to evaluate default classification models and generate a ranking of them using a combination of multiple performance metrics (Peng *et al.* 2011b). Though the correct-classification rate is a prominent measurement in evaluating default risk models, it is not a sufficient measure of a model's predictive power due to the highly unbalanced nature of default data (Khandani *et al.* 2010). Thus it is a common practice in default classification model evaluation to use several performance measures. Based on this observation, this study treats the model evaluation and selection problem as a multiple criteria decision making (MCDM) problem (Peng *et al.* 2011a; Zavadskas, Turskis 2011; Brauers, Zavadskas 2011; Antucheviciene *et al.* 2011; Fernando *et al.* 2012). Each performance measure is considered as a criterion and the quality of default classification model is determined by combining multiple performance metrics (e.g. accuracy, type-I error, and type-II error). TOPSIS, a multiple criteria decision making (MCDM) method, is used to rank the default prediction models.

The bank loan default prediction process includes data pre-processing, modelling, and model assessment and selection (Verikas *et al.* 2010). Two challenges in data pre-processing are high dimensionality and class unbalancedness. A large number of features may be collected in bank loan data and some of them are redundant or irrelevant. Discarding such features can improve the performance of prediction models (Fukunaga 1972). This paper examines the performances of two feature selection techniques (i.e. PCA and ICA) in the context of the proposed procedure and a real-life large corporate bank loan data. Bank loan datasets often have highly unequal frequencies of the normal and default records. Default models trained on random samples from unbalanced datasets may perform poorly on future bank loan default data (Zhang, Zhou 2004; Kou *et al.* 2005; Shi *et al.* 2005; Kou *et al.* 2014). The Synthetic Minority Oversampling Technique (SMOT) (Chawla *et al.* 2002) is utilized in this study to handle the unbalancedness by creating synthetic default records. The proposed approach is applied to a proprietary Chinese commercial bank loan dataset, which contains over 10,000 records of corporate loans with 120 variables.

The paper is structured as follows. Section 1 explains the proposed approach. Section 2 presents the design and the results of the experimental study. The final section concludes the paper with summaries and future research directions.

1. Research methodology

This section describes the proposed approach that this paper used to pre-processing, classifying, and evaluating default classification models. The major components of this approach are explained in sequence.

1.1. Proposed approach for bank loan default prediction

Figure 1 describes the proposed approach for bank loan default prediction that includes feature selection, sampling, classification, and model evaluation.

Real world data may be incorrect, duplicate, incomplete, missing, and dispersed. Thus they need to be cleaned, integrated, and transformed. A comprehensive description of methods and techniques for data preparation can be found at Han and Kamber (2006). Data preparation steps applied to the bank loan default data will be discussed in Section 2.1.

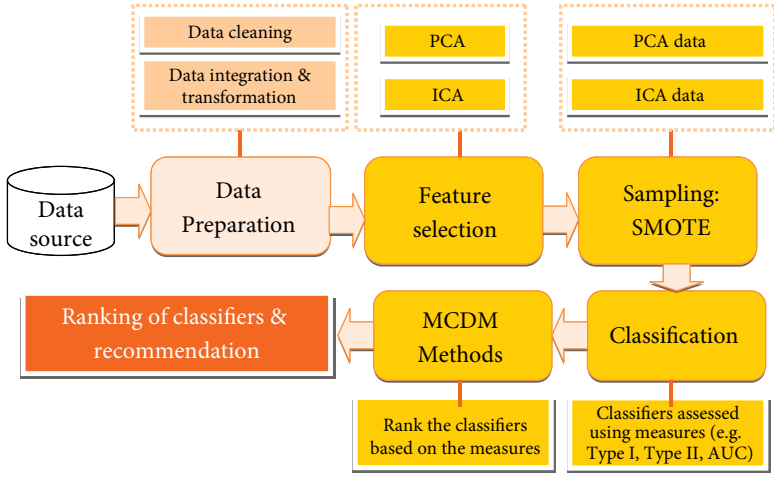

Fig. 1. Proposed approach for bank loan default classification

The following subsections present the four major components of the procedure: feature selection, sampling, classification algorithms, and the MCDM methods for model evaluation.

1.2. Feature selection techniques

Feature selection is an essential step in bank loan default prediction. It can not only reduce the computation complexity, but also improve the performance of prediction models by removing irrelevant variables. Many studies have been conducted in the area of corporate bankruptcy prediction to compare feature selection techniques (e.g. Ryu, Yue 2005; Tsai 2009), combine feature selection methods with classifiers (e.g. Back *et al.* 1996), or design systems integrating feature selection and prediction models (e.g. Lin, McClean 2001).

Principal component analysis (PCA) (Pearson 1901) is one of the most popular feature selection techniques used in bankruptcy prediction. Independent component analysis (ICA) (Jutten, Hérault 1991; Comon 1994), a relatively new technique, has also been adopted in feature selection with increasing frequency and has shown promise in some circumstances (e.g. Cao, Chong 2002; Kao *et al.* 2011). While PCA is suitable for datasets with Gaussian distributions, ICA can decompose non-Gaussian datasets into independent components. The effects of PCA and ICA on pattern recognition problems have been examined in several studies and the observations were mixed: Fortuna and Capson (2004) have shown that ICA outperformed PCA in its generalization ability. Baek *et al.* (2002), on the other hand, claimed that PCA outperformed ICA for face recognition and the difference in performance between the two techniques is statistically significant. Meanwhile, Moghaddam (2002), as well as Ekenel and Sankur (2004), observed that there are no significant performance differences between ICA and PCA.

In the bank loan default prediction field, few works have been conducted to compare the prediction performances of features selected by PCA and ICA. In this paper, both techniques are applied to the original bank loan data to compute principal and independent components. The two new datasets are then sampled and classified using the same sampling technique and classifiers.

1.2.1. Principal component analysis (PCA)

Principal component analysis (PCA) is a multivariate statistical technique that uses an orthogonal transformation to convert a set of often inter-correlated variables into a new set of orthogonal variables (Abdi, Williams 2010). PCA has a lot of applications and its goal in feature extraction is to find a mapping that simplifies the description of the dataset.

Let $X = (x_1, x_2, ..., x_m)^T$ be the original dataset of size $m \times n, m \le n,$ consisting of observed variable x_i of size $1 \times n, i = 1, 2, ..., m.$ Let **Y** be another $m \times n$ matrix related by a linear transformation **P**:

$$\mathbf{PX} = \mathbf{Y}, \tag{1}$$

where the rows of **P**, $\{p_1, ..., p_m\}$, are a set of new basis vectors, which are also called principal components, for representing the columns of **X**. Principal components can be obtained by a three-step algorithm (Shlens 2005): the first step selects a normalized direction in m-dimensional space along which the variance in **X** is maximized and save it as p_1; the second step finds another direction along which variance is maximized under the constraint that it is orthogonal to the preceding vector and save it as p_i; the third step repeats this procedure until *m* vectors are selected. The output is a set of ordered principal components $\boldsymbol{p}$. The dimensionality of the original data can be reduced by choosing the strongest principal components.

1.2.2. Independent component analysis (ICA)

ICA searches the linear transformation that minimizes the statistical dependence between its components and may be considered as an extension of PCA (Comon 1994). The ICA model is defined as (Hyvärinen, Oja 2000):

$$\mathbf{x} = \mathbf{As} = \sum_{i=1}^{n} a_i s_i, \tag{2}$$

where: **x** is the observed random vector with elements $x_1, ..., x_n$; **A** is the mixing matrix with elements a_{ij}; and **s** is the random vector with elements $s_1, ..., s_n$. Both **s** and **A** are assumed to be unknown and must be estimated using the observed vector **x** under several assumptions. First, the components s_i are assumed to be statistically independent. Second, the independent component (IC) is assumed to have non-Gaussian distributions. Third, the unknown mixing matrix is assumed to be square. Then the ICs can be obtained by compute **W**, the inverse of **A**:

$$\mathbf{s} = \mathbf{Wx}. \tag{3}$$

Non-Gaussianity is the key to estimating the ICA model and can be measured by kurtosis, negentropy, and approximations of negentropy (Hyvärinen, Oja 2000). Kurtosis

is the classical measure of non-Gaussianity and is adopted in this paper. The kurtosis of y is defined by:

$$\text{kurt}(y) = E\left\{y^4\right\} - 3(E\left\{y^2\right\})^2. \quad (4)$$

If y is assumed to zero mean and unit variance, the right-hand side becomes $\text{E}\left\{y^4\right\} - 3$. Thus, kurtosis is zero for a Gaussian random variable and nonzero for most non-Gaussian random variables.

Several algorithms have been developed to solve the ICA problem (Bell, Sejnowski 1995; Amari *et al.* 1996; Karhunen, Oja 1997; Hyvärinen 1999). This paper adopts the FastICA algorithm proposed by Hyvärinen and Oja (2000).

1.3. Sampling approach: SMOTE

Credit-default data are normally highly unbalanced because the proportion of the number of default loans to the population is normally below 10% (Paleologo *et al.* 2010). Unbalanced datasets cause at least two types of problems in default prediction (Zhang, Zhou 2004). The first problem is that the performance of classification models for the default class may be poor due to the overwhelming number of normal loan records in the samples. The second problem is related to classifier evaluation. Predictive accuracy, the most widely used measure for classifiers, can be misleading with unbalanced data. For example, suppose a dataset has 95% non-default firms and 5% default firms. An accuracy rate of 95% may be inadequate since the classifier could be correctly classified only the non-default firms. Possible ways to address this problem are sampling techniques and multiple performance measures. The performance measures will be discussed in Section 1.5.

Many sampling techniques have been developed to deal with unbalanced datasets. The fundamental idea of these techniques is to oversampling the minority class and/or under-sampling the majority class (Chawla *et al.* 2002). Chawla *et al.* (2002) proposed the Synthetic Minority Oversampling TEchnique (SMOTE), an over-sampling approach in which the minority class is over-sampled by creating synthetic minority class records. The synthetic minority samples are generated along the line segments joining any or all of the k minority class nearest neighbours. SMOTE achieves better performance than over-sampling with replacement techniques because it creates larger decision regions, which leads to more coverage of the minority class. The experiment implements SMOTE using the SMOTE class (weka.filters.supervised.instance.SMOTE) of WEKA 3.7 (Witten, Frank 2005).

1.4. Classification algorithms and performance measures

Default risk prediction can be considered as a binary classification problem, which means bank loans are assigned to one of the two classes (default or non-default) after data analysis. In the classification framework, a model or classifier attempts to accurately predict class labels (default or non-default) of future individual loans based on two steps. In the first step, a classifier learns a function that maps independent variables or input variables to a dependent or output variable (class label). The second step tests the predictive accuracy of the classifier learned in the first step (Han, Kamber 2006).

1.4.1. Classification algorithms

Data mining and knowledge discovery (DMKD) has made great progress during the last twenty years (Peng *et al.* 2008b). As one of the major tasks of data mining and an important problem in research and practice, classification has wide business and scientific applications, such as credit risk management, default prediction, marketing promotion, financial loan evaluation, insurance premium calculation, and medical clinic decision. Five classification algorithms: Naïve Bayes, linear logistic regression, *k*-nearest-neighbour (*k*-NN), C4.5, and Classification and Regression Tree (CART), are chosen for the experimental study.

Naïve Bayes (Domingos, Pazzani 1997) models probabilistic relationships between input variables and the output variable. It estimates the class-conditional probability based on Bayes theorem and can represent simple distributions. Linear logistic regression (Cessie, Houwelingen 1992) models the probability of occurrence of an event as a linear function of a set of input variables and generalizes well to multiclass problems. The *k*-nearest-neighbour (Cover, Hart 1967) is an instance-based learning method. It classifies a new instance to the same class as the closest existing instance, and the closeness is measured by a distance metric. In the experimental study, *k*-NN is implemented by the IBK algorithm in WEKA (Witten, Frank 2005). C4.5 is a decision tree algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner (Quinlan 1993) and is one of the most extensively studied machine learning algorithms. Due to its outstanding performance and easy-to-interpret decision rules, C4.5 became a benchmark of classification algorithms. Classification and Regression Tree (CART) (Breiman *et al.* 1984) is a greedy algorithm for learning multivariate decision trees. It can be used to predict both continuous and categorical variables.

1.4.2. Performance measures

Widely used performance measures of default prediction systems are accuracy, Type-I error, Type-II error, and AUC (Verikas *et al.* 2010). Bank loan default prediction is normally a two-class problem. Suppose default and non-default loans are labelled as negative and positive, respectively. These measures can be defined as follows.

– Overall accuracy: Accuracy is the percentage of correctly classified instances. It is the most widely used classification performance metrics.

$$\text{Overall Accuracy} = \frac{TN + TP}{TP + FP + FN + TN}, \tag{5}$$

where TN, TP, FN, and FP stand for true negative, true positive, false negative, and false positive, respectively.

– True Positive (TP): TP is the number of correctly classified non-default instances. TP rate is also called sensitivity measure.

$$\text{True Positive rate/Sensitivity} = \frac{TP}{TP + FN}. \tag{6}$$

– False Positive (FP): FP is the number of default firms that is misclassified as non-default class. FP rate is called Type-I error. In default prediction, Type-I error is more important than Type-II error because it represents potential credit loss and Type-II

error implies potential loss of business. Actually, Altman *et al.* (1977) indicated that the costs of Type-I error are 35 times higher for banks than Type-II error costs.

$$\text{False Positive rate/Type-I error} = \frac{FP}{FP + TN}. \tag{7}$$

– True Negative (TN): TN is the number of correctly classified default instances. TN rate is also called specificity measure.

$$\text{True Negative rate/Specificity} = \frac{TN}{TN + FP}. \tag{8}$$

– False Negative (FN): FN is the number of non-default instances that is misclassified as default class. FN rate is also called Type-II error.

$$\text{False Negative rate/Type-II error} = \frac{FN}{FN + TP}. \tag{9}$$

– Area under the curve (AUC): ROC stands for Receiver Operating Characteristic, which shows the tradeoff between TP rate and FP rate. AUC represents the accuracy of a classifier. The larger the area, the better the classifier.

1.5. MCDM method: TOPSIS

There are many MCDM methods that can be used to rank alternatives. This study chose TOPSIS due to its good performance in previous studies and ease of use (Agrawal *et al.* 1991; Zanakis *et al.* 1998). Hwang and Yoon (1981) proposed the Technique for order preference by similarity to ideal solution (TOPSIS) method to rank alternatives over multiple criteria. It finds the best alternatives by minimizing the distance to the idea solution and maximizing the distance to the nadir or negative-ideal solution (Olson 2004).

A number of extensions and variations of TOPSIS have been developed over the years. The following TOPSIS procedure adopted from Opricovic and Tzeng (2004: 448–449) and Olson (2004: 722) is used in the experimental study.

Step 1: calculate the normalized decision matrix. The normalized value r_{ij} is calculated as:

$$r_{ij} = x_{ij} / \sqrt{\sum_{j=1}^{J} x_{ij}^2}, \; j = 1,...,J; \; i = 1,...,n, \tag{10}$$

where: J and n denote the number of alternatives and the number of criteria, respectively. For alternative A_j, the performance measure of the ith criterion C_i is represented by x_{ij}.

Step 2: develop a set of weights w_i for each criterion and calculate the weighted normalized decision matrix. The weighted normalized value v_{ij} is calculated as:

$$v_{ij} = w_i r_{ij}, \; j = 1,...,J; \; i = 1,...,n, \tag{11}$$

where w_i is the weight of the i^{th} criterion, and $\sum_{i=1}^{n} w_i = 1$.

Step 3: find the ideal alternative solution S^+, which is calculated as:

$$S^+ = \left\{v_1^+,...,v_n^+\right\} = \left\{\left(\max_j v_{ij} | i \in I'\right), \left(\min_j v_{ij} | i \in I''\right)\right\}, \tag{12}$$

where I' is associated with benefit criteria and I'' is associated with cost criteria. For the evaluation of default prediction models, accuracy and AUC are benefit criteria and have to be maximized, and the Type-I error and Type-II error are cost criteria and have to be minimized.

Step 4: find the negative-ideal alternative solution S^-, which is calculated as:

$$S^- = \left\{v_1^-,...,v_n^-\right\} = \left\{\left(\min_j v_{ij}|i\in I'\right),\left(\max_j v_{ij}|i\in I''\right)\right\}. \tag{13}$$

Step 5: Calculate the separation measures, using the n-dimensional Euclidean distance. The separation of each alternative from the ideal solution is calculated as:

$$D_j^+ = \sqrt{\sum_{i=1}^{n}(v_{ij}-v_i^+)^2},\ j=1,...,J. \tag{14}$$

The separation of each alternative from the negative-ideal solution is calculated as:

$$D_j^- = \sqrt{\sum_{i=1}^{n}(v_{ij}-v_i^-)^2},\ j=1,...,J. \tag{15}$$

Step 6: Calculate a ratio R_j^+ that measures the relative closeness to the ideal solution and is calculated as:

$$R_j^+ = D_j^- / (D_j^+ + D_j^-),\ j=1,...,J. \tag{16}$$

Step 7: Rank alternatives by maximizing the ratio R_j^+.

2. Experimental study

The experiment is designed to test the proposed approach using a large real-life bank loan default dataset provided by a Chinese commercial bank. The following subsections describe the data, the experimental design, and the results.

2.1. Bank loan data

The dataset which was provided by a Chinese commercial bank contains 10,868 short-term, medium-term, and long-term loans granted to various sizes of companies over the period January 2002 – June 2004. Companies included in the data cover a wide range of industries, such as manufacturing, information technology, pharmacy, telecommunication, foreign trading, energy, agriculture, public utility, and real estate.

The raw dataset was pre-processed to remove irrelevant, sensitive, and incorrect data. Originally, there are 120 variables and some of them are sensitive and irrelevant to the default prediction, such as company identification codes, company names, and company addresses. Second, instances with more than 50% missing values were considered as incomplete or incorrect and deleted from the dataset. Pre-processing step removed 224 instances and 16 variables.

Different definitions of 'defaulted' or 'bad' loans are available in the literature (e.g. Marais *et al.* 1984; Hurt, Felsovalyi 1998; Basel Committee… 2004; Dermine, Carvalho 2006). Because the dataset used in this study was provided by a Chinese commercial bank,

the following five-category loan classification system used by China's state-owned commercial banks (SCB) and joint shareholding commercial banks (JSCB) is adopted. The five-category criteria classify bank loans as pass, special-mention, substandard, doubtful, and loss. The last three categories are treated as defaulted loans by SCB and JSCB. A loan is classified as 'substandard' when the obligor's ability to pay its credit obligations is in question depending on its normal business revenues. A loan is classified as 'doubtful' when the obligor cannot pay interest and/or principal. A loan is classified as 'loss' when principal and interest cannot be recovered or only a small portion can be recovered after taking all possible measure or legal procedure is started (Bank loan classification regulation 2002; Banks to adopt five-category loan classification system 2003).

According to this definition, the sample data has 9,928 normal loans and 716 defaulted loans. The 104 variables describe five important financial aspects of companies as pointed out by Altman (1968), including liquidity, profitability, leverage, solvency, and activity. This is a typical bank loan data with two characteristics: a large number of variables and highly unbalanced (only 6.7% default loans). Table 1 compares the normal and defaulted loans in the dataset from four aspects (McKee, Lensberg 2002). RMB is the abbreviation of "renminbi", the official currency of China.

Table 1. Comparison of subsample characteristics

Variable or ratio	Normal loans	Defaulted loans
Total assets		
maximum	RMB.26,854,906,969	RMB.1,792,017,755.50
Mean	RMB.559,394,670.58	RMB.87,387,234.37
minimum	RMB.640.1	RMB.1827.6
S.D.	RMB.2,447,286,504.76	RMB.183,269,496.1
Total revenues		
maximum	RMB.16,993,642,023	RMB.1,620,347,617
Mean	RMB.333,414,922.10	RMB.44,486,523.13
minimum	RMB.-1,631,589.70	RMB.-335,266.87
S.D.	RMB.1,104,160,852	RMB.159,020,842.9
Net income/Total assets		
Mean	2.915901569	–2.104
S.D.	43.47841237	47.665
Cash/Current liabilities		
Mean	0.354	0.102
S.D.	4.117	0.184

The two subsamples have some different characteristics. For example, the normal loans have larger mean total assets and mean total revenues than the defaulted loans. The Net income/Total assets ratio and the Cash/Current liabilities ratio of the normal loan category are also higher than the defaulted loan category. On the other hand, the normal and defaulted loans exhibit some similar features. For instance, both normal and defaulted loans have

wide range values for total assets and revenues. The defaulted subsample has a negative Net income/Total assets ratio and has 45% companies with nonpositive net incomes. Although the normal group has a positive Net income/Total assets ratio, 3,425 companies (or 34%) in the subsample have nonpositive net incomes. This feature is similar to the one reported by McKee and Lensberg (2002). The mixed subsample characteristics reflect the real life problem in loan default prediction.

2.2. Experimental design

The experiment was carried out according to the following process:

Input: The bank loan default dataset

Output: Rankings of classifiers

Step 1. Prepare the dataset: remove irrelevant variables; handle missing values by filling in with the variable average; eliminate incorrect data.

Step 2. Select relevant variables using PCA and ICA to get two sets of data. PCA is implemented using the PrincipalComponents class of WEKA 3.7 (Witten, Frank 2005; Hall *et al*. 2009) and ICA is implemented using the FastICA operator of RapidMiner 5.0 (Mierswa *et al*. 2006).

Step 3. Applying SMOTE approach to the PCA and ICA datasets to generate synthetic default instances, which is implemented using the SMOTE class of WEKA 3.7 (Witten, Frank 2005; Hall *et al*. 2009).

Step 4. Train and test the selected classification algorithms using WEKA 3.7 on randomly sampled partitions (i.e. 10-fold cross-validation) of the sampled PCA and ICA datasets, respectively.

Step 5. Evaluate classification algorithms using TOPSIS, which is implemented using MATLAB 7.0 (Matlab 2005).

Step 6. Generate the ranking of classification algorithms provided by TOPSIS for the PCA and ICA datasets, respectively.

END

After removing irrelevant variables, there are 104 variables in the dataset. Non-Gaussianity is the key to estimating the ICA model and can be verified by kurtosis. Figure 2 illustrates that all variables satisfy non-Gaussianity since none of them has zero kurtosis.

PCA and ICA methods are applied to the pre-processed data to reduce the dimensionality. A threshold of 85% is set for both PCA and ICA methods. The resulting datasets after applying PCA and ICA have 31 and 32 variables, respectively.

The SMOTE approach is used to create synthetic default instances for the PCA and ICA datasets. Figure 3 compares the different class sizes of default before and after applying the SMOTE. The default data increase from 716 records to 3365 records.

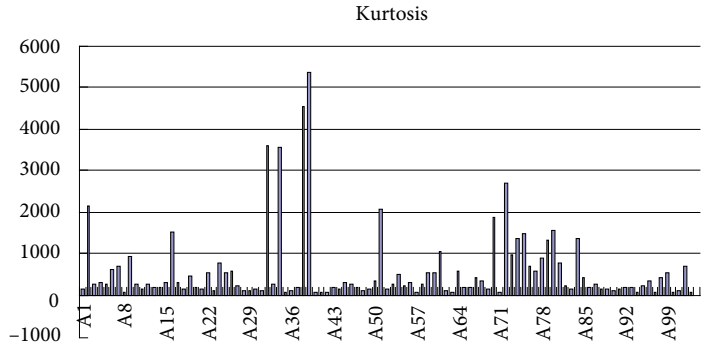

Fig. 2. Kurtosis for variables

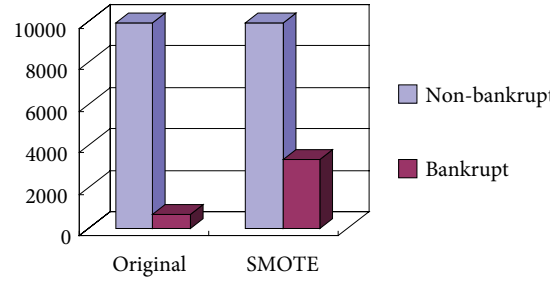

Fig. 3. Comparison of class sizes before and after applying SMOTE

2.3. Results and discussions

2.3.1. Classification results

Table 2 summarizes classification results on the PCA test dataset. The best accuracy, Type-I error rate, Type-II error rate, and AUC are highlighted in boldface and italic. *K*-NN has the highest accuracy of 96.67%, followed by C4.5 and CART. Naïve Bayes and logistic have the lowest Type-I and Type-II error rate, respectively. However, the low Type-I or Type-II error rate is achieved by sacrificing the other type of error rate. The 6.4% Type-I error rate of naïve Bayes is accompanied with a 69.7% Type-II error rate and the 1.2% Type-II error rate of logistic is accompanied with a 48.18% Type-I error rate. For the measure of AUC, the performances of *K*-NN, C4.5 and CART are close and the average rate of the five classifiers reaches 90.2%.

Table 2. Classification results on the PCA test dataset

	Accuracy	Type-I errors	Type-II errors	AUC
Naive Bayes	0.4674	**0.0642**	0.6977	0.75
Logistic	0.8655	0.4818	**0.012**	0.9
K-NN	**0.9667**	0.0771	0.0178	**0.96**
C4.5	0.9564	0.1017	0.0231	0.95
CART	0.9522	0.1176	0.0231	0.95

Table 3 compares the five classifiers using the same set of measures on the ICA test dataset. *K*-NN is still the best classifier in terms of the accuracy and AUC. Similarly, Naïve Bayes and logistic have the lowest Type-I and Type-II error rate, respectively. At the same time, these two classifiers also have the highest Type-II and Type-I error rates.

Table 3. Classification results on the ICA test dataset

	Accuracy	Type-I errors	Type-II errors	AUC
Naive Bayes	0.3897	**0.0536**	0.8066	0.64
Logistic	0.7633	0.8791	**0.0101**	0.79
K-NN	**0.9665**	0.0902	0.0135	**0.97**
C4.5	0.9342	0.164	0.0311	0.92
CART	0.9266	0.1922	0.0315	0.91

2.3.2. PCA vs. ICA

To compare the effects of PCA and ICA on the classification results, the experiment applied the selected classification algorithms to the PCA and ICA datasets. The averaged results of accuracy, AUC, Type-I error rate, and Type-II error rate for the PCA and ICA data are summarized in Table 4. It can be seen that PCA outperforms ICA on the four measurements.

Table 4. Comparison of classification results using the PCA and ICA datasets

	Accuracy	Type-I errors	Type-II errors	AUC
PCA	0.836933	0.256317	0.130133	0.86
ICA	0.786617	0.396417	0.148817	0.788333

2.3.3. TOPSIS results

The rankings of classifiers generated by TOPSIS for the PCA data and the ICA data are summarized in Table 5. The rankings of the five classifiers using the PCA and ICA datasets are identical and *K*-NN is ranked the best classifier.

Table 5. Ranking of classifiers by TOPSIS

	TOPSIS(PCA)	TOPSIS(ICA)
Naive Bayes	0.4169	0.3992
Logistic	0.681	0.6204
K-NN	**0.9831**	**0.9752**
C4.5	0.9609	0.9264
CART	0.9478	0.9111

In this specific dataset, it chanced that one classifier outperformed other classifiers in most performance measures. In reality, it is normal that classifiers with the best scores on some measures may perform poorly on other measures. In those circumstances, the MCDM methods are helpful in providing a final ranking of classification algorithms. It is worth to mention that the performance measures used by this study are mainly accuracy-related. In

real application in commercial banks, there are other issues need to be considered when selecting classifiers. For instance, probability of default (PD) is a commonly used indicator in bankruptcy prediction and thus a classifier which can provide PD information may be preferred over a classifier which cannot provide PD information by commercial banks.

Conclusions

This paper concentrated on feature selection, unbalanced data, and model assessment in bank loan default prediction. It developed a procedure to address the three problems. Firstly, ICA and PCA were used to select relevant features. Secondly, the SMOT approach was utilized to deal with the unbalanced data by creating synthetic default examples. Thirdly, TOPSIS, a multiple criteria decision making method, was utilized to rank a selection of default prediction models.

An experimental study using a large real-life bank loan default dataset provided by a Chinese commercial bank was conducted to validate the proposed process. PCA and ICA generated two separate datasets and tested in the sampling, classification, and model evaluation steps. After feature selection, PCA and ICA reduced the dimensionality from 104 to 31 and 32, respectively. The reduced datasets were then balanced using the SMOTE approach and classified with five selected classification algorithms. The performances of classifiers were measured using accuracy, Type-I error rate, Type-II error rate, and AUC. Finally, TOPSIS was applied to the PCA and ICA datasets to evaluate the classifiers based on the values of performance metrics. The experimental results showed that *K*-NN has good potential in default prediction. In addition, the outcome indicated that there is no significant difference between PCA and ICA on default prediction of the specific dataset.

Future research issues include introducing more feature selection techniques, sampling approaches, classification algorithms, and MCDM methods to the process. Ranking results generated by MCDM methods may be very different. Another research direction is to resolve this disagreement and help decision-makers pick the most suitable classifier(s). Experimental study with more default datasets may also be conducted to test the process.

Acknowledgements

This research has been partially supported by grants from the National Natural Science Foundation of China (#71325001 for Yi Peng, #71222108 for Gang Kou), the Fundamental Research Funds for the Central Universities, the Research Fund for the Doctoral Program of Higher Education (#20120185110031) and Program for New Century Excellent Talents in University (NCET-12-0086 and NCET-10-0293).

References

Abdi, H.; Williams, L. J. 2010. Principal component analysis, *Wiley Interdisciplinary Reviews: Computational Statistics* 2(4): 433–459. http://dx.doi.org/10.1002/wics.101

Agrawal, V. P.; Kohli, V.; Gupta, S. 1991. Computer aided robot selection: the multiple attribute decision making approach, *International Journal of Production Research* 29(8): 1629–1644. http://dx.doi.org/10.1080/00207549108948036

Altman, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* 23(1): 589–609. http://dx.doi.org/10.2307/2325319

Altman, E. I. 1973. Predicting railroad bankruptcies in America, *Bell Journal of Economics* 4(1): 184–211. http://dx.doi.org/10.2307/3003144

Altman, E. I.; Haldeman, R. G.; Narayanan, P. 1977. ZETA analysis, *Journal of Banking and Finance* 1(1): 29–54. http://dx.doi.org/10.1016/0378-4266(77)90017-6

Altman, E. I. 1984. The success of business failure prediction models, *Journal of Banking and Finance* 8(2): 171–198. http://dx.doi.org/10.1016/0378-4266(84)90003-7

Amari, S.; Chichocki, A.; Yang, H. 1996. A new learning algorithm for blind signal separation, *Advances in Neural Information Processing Systems* 8: 757–763.

Antucheviciene, J.; Zakarevicius, A.; Zavadskas, E. K. 2011. Measuring congruence of ranking results applying particular MCDM methods, *Informatica* 22(3): 319–338.

Back, B.; Laitinen, T.; Sere, K. 1996. Neural network and genetic algorithm for bankruptcy prediction, *Expert Systems with Applications* 11(4): 407–413. http://dx.doi.org/10.1016/S0957-4174(96)00055-3

Baek, K.; Draper, B. A.; Beveridge, J. R.; She, K. 2002. PCA vs. ICA: a comparison on the FERET data set, *In JCIS*, 824–827.

Bank loan classification regulation. 2002. The People's Bank of China.

Banks to adopt five-category loan classification system [online], [cited 08 October 2011]. 2003. *People's Daily*. Available from Internet: http://english.peopledaily.com.cn/200309/06/eng20030906_123830.shtml

Banks, W. J.; Prakash, L. A. 1994. On the performance of linear programming heuristics applied on a quadratic transformation in the classification problem, *European Journal of Operational Research* 74(23): 23–28. http://dx.doi.org/10.1016/0377-2217(94)90325-5

Basel Committee on Banking Supervision, 2004. International Convergence of Capital Measurement and Capital Standards. Basel.

Bastos, J. A. 2010. Forecasting bank loans loss-given-default, *Journal of Banking and Finance* 34(10): 2510–2517. http://dx.doi.org/10.1016/j.jbankfin.2010.04.011

Beaver, W. H. 1966. Financial ratios as predictors of failure, *Journal of Accounting Research* 4: 71–111. http://dx.doi.org/10.2307/2490171

Bell, A. J.; Sejnowski, T. J. 1995. An information maximization approach to blind separation and blind deconvolution, *Neural Computation* 7(6): 1129–1159. http://dx.doi.org/10.1162/neco.1995.7.6.1129

Brauers, W. K. M.; Zavadskas, E. K. 2011. Multimoora optimization used to decide on a bank loan to buy property, *Technological and Economic Development of Economy* 17(1): 174–188. http://dx.doi.org/10.3846/13928619.2011.560632

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. 1984. *Classification and regression trees*. Wadsworth International Group, Belmont, California.

Bryant, S. M. 1997. A case-based reasoning approach to bankruptcy prediction modeling, *International Journal of Intelligent Systems in Accounting, Finance and Management* 6(3): 195–214. http://dx.doi.org/10.1002/(SICI)1099-1174(199709)6:3<195::AID-ISAF132>3.3.CO;2-6

Buta, P. 1994. Mining for financial knowledge with CBR, *AI Expert* 9(2): 34–41.

Cao, L. J.; Chong, W. K. 2002. Feature extraction in support vector machine: a comparison of PCA, XPCA and ICA, in *Proceedings of the 9th International Conference on Neural Information*, 18–22 November, 2002, Singapore, 1001–1005.

Cessie, S. le; Houwelingen, J. C. 1992. Ridge estimators in logistic regression, *Applied Statistics* 41(1): 191–201. http://dx.doi.org/10.2307/2347628

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Oversampling Technique, *Journal of Artificial Intelligence Research* 16: 341–378.

Cielen, A.; Peeters, L.; Vanhoof, K. 2004. Bankruptcy prediction using a data envelopment analysis, *European Journal of Operational Research* 154(2): 526–532. http://dx.doi.org/10.1016/S0377-2217(03)00186-3

Comon, P. 1994. Independent component analysis – a new concept?, *Signal Processing* 36(3): 287–314. http://dx.doi.org/10.1016/0165-1684(94)90029-9

Cover, T. M.; Hart, P. E. 1967. Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* 13(1): 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964

Crook, J. N.; Edelman, D. B.; Thomas, L. C. 2007. Recent developments in consumer credit risk assessment, *European Journal of Operational Research* 183(3): 1447–1465. http://dx.doi.org/10.1016/j.ejor.2006.09.100

Dermine, J.; Neto De Carvalho, C. 2006. Bank loan losses-given-default: a case study, *Journal of Banking and Finance* 30(4): 1219–1243. http://dx.doi.org/10.1016/j.jbankfin.2005.05.005

Domingos, P.; Pazzani, M. 1997. On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning* 29(203): 103–130. http://dx.doi.org/10.1023/A:1007413511361

Doumpos, M.; Zopounidis, C. 2007. Model combination for credit risk assessment: a stacked generalization approach, *Annals of Operations Research* 151(1): 289–306. http://dx.doi.org/10.1007/s10479-006-0120-x

Ekenel, H. K.; Sankur, B. 2004. Feature selection in the independent component subspace for face recognition, *Pattern Recognition Letters* 25(12): 1377–1388. http://dx.doi.org/10.1016/j.patrec.2004.05.013

Fernandez, E.; Olmeda, I. 1995. Bankruptcy prediction with artificial neural networks, *Lecture Notes in Computer Science* 930: 1142–1146. http://dx.doi.org/10.1007/3-540-59497-3_296

Fernando, A. F.; Spahr, R.; Santos, P.; Rodrigues, P. M. M. 2012. A multiple criteria framework to evaluate bank branch potential attractiveness, *International Journal of Strategic Property Management* 16(3): 254–276. http://dx.doi.org/10.3846/1648715X.2012.707629

Fortuna, J.; Capson, D. 2004. Improved support vector classification using PCA and ICA feature space modification, *Pattern Recognition* 37(6): 1117–1129. http://dx.doi.org/10.1016/j.patcog.2003.11.009

Frydman, H.; Altman, E. I.; Kao, D. 1985. Introducing recursive partitioning for financial classification: the case of financial distress, *Journal of Finance* 40(1): 269–291. http://dx.doi.org/10.1111/j.1540-6261.1985.tb04949.x

Fukunaga, K. 1972. *Introduction to statistical pattern recognition*. New York: Academic Press. 369 p.

Galindo, J.; Tamayo, P. 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications, *Computational Economics* 15(1–2): 107–143. http://dx.doi.org/10.1023/A:1008699112516

Gilbert, L. R.; Menon, K.; Schwartz, K. B. 1990. Predicting bankruptcy for firms in financial distress, *Journal of Business Finance and Accounting* 17(1): 161–171. http://dx.doi.org/10.1111/j.1468-5957.1990.tb00555.x

Gorzalczany, M. B.; Piasta, Z. 1999. Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support, *Information Sciences* 120(1–4): 45–68. http://dx.doi.org/10.1016/S0020-0255(99)00070-5

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. 2009. The WEKA data mining software: an update, *SIGKDD Explorations* 11(1): 10–18. http://dx.doi.org/10.1145/1656274.1656278

Han, J.; Kamber, M. 2006. *Data mining: concepts and techniques*. 2nd edition. San Francisco: Morgan Kaufmann Publishers. 770 p.

Hand, D. J.; Henley, W. E. 1997. Statistical classification methods in consumer credit scoring: a review, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 160(3): 523–541. http://dx.doi.org/10.1111/j.1467-985X.1997.00078.x

Henley, W. E. 1995. *Statistical aspects of credit scoring*: PhD Dissertation. Department of Statistics, The Open University, Milton Keynes.

Henley, W. E.; Hand, D. J. 1996. A k-nearest-neighbour classifier for assessing consumer credit risk, *Journal of the Royal Statistical Society. Series D (The Statistician)* 45(1): 77–95. http://dx.doi.org/10.2307/2348414

Hurt, L.; Felsovalyi, A. 1998. Measuring loss on Latin American defaulted bank loans, a 27-year study of 27 countries, *Journal of Lending and Credit Risk Management* 80: 41–46.

Hwang, C. L.; Yoon, K. 1981. *Multiple attribute decision making methods and applications*. Berlin Heidelberg: Springer. 259 p. http://dx.doi.org/10.1007/978-3-642-48318-9

Hyvärinen, A. 1999. Survey on independent component analysis, *Neural Computing Surveys* 2: 94–128.

Hyvärinen, A.; Oja, E. 2000. Independent component analysis: algorithms and applications, *Neural Networks* 13(4–5): 411–430. http://dx.doi.org/10.1016/S0893-6080(00)00026-5

Jutten, C.; Hérault, J. 1991. Blind separation of sources, Part I: an adaptive algorithm based on neuromimetic architecture, *Signal Processing* 24(1): 1–10. http://dx.doi.org/10.1016/0165-1684(91)90079-X

Kao, L. J.; Lu, C. J.; Chiu, C. C. 2011. Efficiency measurement using independent component analysis and data envelopment analysis, *European Journal of Operational Research* 210(2): 310–317. http://dx.doi.org/10.1016/j.ejor.2010.09.016

Karhunen, J.; Oja, E. 1997. A class of neural networks for independent component analysis, *IEEE Transactions on Neural Networks* 8(3): 486–503. http://dx.doi.org/10.1109/72.572090

Khandani, A. E.; Kim, A. J.; Lo, A. W. 2010. Consumer credit-risk models via machine-learning algorithms, *Journal of Banking and Finance* 34(11): 2767–2787. http://dx.doi.org/10.1016/j.jbankfin.2010.06.001

Kou, G.; Peng, Y.; Shi, Y.; Wise M.; Xu, W. 2005. Discovering credit cardholders' behavior by multiple criteria linear programming, *Annals of Operations Research* 135(1): 261–274. http://dx.doi.org/10.1007/s10479-005-6245-5

Kou, G; Peng, Y.; Wang, G. 2014. Evaluation of clustering algorithms for financial risk analysis using MCDM methods, *Information Sciences* (in Press). http://dx.doi.org/10.1016/j.ins.2014.02.137

Kumar, P. R.; Ravi, V. 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review, *European Journal of Operational Research* 180(1): 1–28. http://dx.doi.org/10.1016/j.ejor.2006.08.043

Li, K. 1999. Bayesian analysis of duration models: an application to Chapter 11 bankruptcy, *Economics Letters* 63(3): 305–312. http://dx.doi.org/10.1016/S0165-1765(99)00052-X

Lin, F. Y.; McClean, S. 2001. A datamining approach to the prediction of corporate failure, *Knowledge Based Systems* 14(3–4): 189–195. http://dx.doi.org/10.1016/S0950-7051(01)00096-X

Lopez, J. A.; Saidenberg, M. R. 2000. Evluating credit risk models, *Journal of Banking and Finance* 24(1–2): 151–165. http://dx.doi.org/10.1016/S0378-4266(99)00055-2

Marais, M. L.; Patel, J.; Wolfson, M. 1984. The experimental design of classification models: an application of recursive partitioning and bootstrapping to commercial bank loan classifications, *Journal of Accounting Research* 22: 87–113. http://dx.doi.org/10.2307/2490861

Martin, D. 1977. Early warning of bank failure: a logit regression approach, *Journal of Banking and Finance* 1(3): 249–276. http://dx.doi.org/10.1016/0378-4266(77)90022-X

MATLAB 2005. The MathWorks, Inc., Natick, MA 01760 [online], [cited 12 December 2012]. Available from Internet: http://www.mathworks.com/products/matlab

McKee, T. E. 1998. A mathematically derived rough set model for bankruptcy prediction, in Brown, C. E. (Ed.). *Collected Papers of the Seventh Annual Research Workshop on Artificial Intelligence and Emerging Technologies in Accounting, Auditing and Tax*. Artificial Intelligence/Emerging Technologies Section of the American Accounting Association, Sarasota, Florida.

McKee, T. E.; Lensberg, T. 2002. Genetic programming and rough sets: a hybrid approach to bankruptcy classification, *European Journal of Operational Research* 138(2): 436–451. http://dx.doi.org/10.1016/S0377-2217(01)00130-8

Medema, L.; Koning, R. H.; Lensink, R. 2009. A practical approach to validating a PD model, *Journal of Banking and Finance* 33(4): 701–708. http://dx.doi.org/10.1016/j.jbankfin.2008.11.007

Meyer, P. A.; Pifer, H. 1970. Prediction of bank failures, *The Journal of Finance* 25(4): 853–868. http://dx.doi.org/10.1111/j.1540-6261.1970.tb00558.x

Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; Euler, T. 2006. YALE: rapid prototyping for complex data mining tasks, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)*, 20–23 August, 2006, Philadelphia, USA, 935–940.

Moghaddam, B. 2002. Principal manifolds and probabilistic subspaces for visual recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(6): 780–788. http://dx.doi.org/10.1109/TPAMI.2002.1008384

Ohlson, J. A. 1980. Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research* 18(1): 109–131. http://dx.doi.org/10.2307/2490395

Olson, D. L. 2004. Comparison of weights in TOPSIS models, *Mathematical and Computer Modelling* 40(7–8): 721–727. http://dx.doi.org/10.1016/j.mcm.2004.10.003

Opricovic, S.; Tzeng, G. H. 2004. Compromise solution by MCDM methods: a comparative analysis of VIKOR and TOPSIS, *European Journal of Operational Research* 156(2): 445–455. http://dx.doi.org/10.1016/S0377-2217(03)00020-1

Paleologo, G.; Elisseeff, A.; Antonini, G. 2010. Subagging for credit scoring models, *European Journal of Operational Research* 201(2): 490–499. http://dx.doi.org/10.1016/j.ejor.2009.03.008

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space, *Philosophical Magazine* 2(6): 559–572. http://dx.doi.org/10.1080/14786440109462720

Peng, Y.; Kou, G.; Shi, Y.; Chen, Z. 2008a. A multi-criteria convex quadratic programming model for credit data analysis, *Decision Support Systems* 44(4): 1016–1030. http://dx.doi.org/10.1016/j.dss.2007.12.001

Peng, Y.; Kou, G.; Shi, Y.; Chen, Z. 2008b. A descriptive framework for the field of data mining and knowledge discovery, *International Journal of Information Technology & Decision Making* 7(4): 639–682. http://dx.doi.org/10.1142/S0219622008003204

Peng, Y.; Kou, G.; Wang, G.; Shi, Y. 2011a. FAMCDM: a fusion approach of MCDM methods to rank multiclass classification algorithms, *Omega* 39(6): 677–689. http://dx.doi.org/10.1016/j.omega.2011.01.009

Peng, Y.; Kou, G.; Wang, G.; Wu, W.; Shi, Y. 2011b. Ensemble of software defect predictors: an AHP-based evaluation method, *International Journal of Information Technology & Decision Making* 10(1): 187–206. http://dx.doi.org/10.1142/S0219622011004282

Quinlan, J. R. 1993. *C4.5: programs for machine learning*. San Francisco: Morgan Kaufmann. 302 p.

Ravi, V.; Pramodh, C. 2008. Threshold accepting trained principal component neural network and feature subset selection: application to bankruptcy prediction in banks, *Applied Soft Computing* 8(4): 1539–1548. http://dx.doi.org/10.1016/j.asoc.2007.12.003

Rosenberg, E.; Gleit, A. 1994. Quantitative methods in credit management: a survey, *Operations Research* 42(4): 589–613. http://dx.doi.org/10.1287/opre.42.4.589

Ryu, Y. U.; Yue, W. T. 2005. Firm bankruptcy prediction: experimental comparison of isotonic separation and other classification approaches, *IEEE Transactions On Systems, Management and Cybernetics – Part A: Systems and Humans* 35(5): 727–737. http://dx.doi.org/10.1109/TSMCA.2005.843393

Shi, Y.; Peng, Y.; Kou, G.; Chen, Z. 2005 Classifying credit card accounts for business intelligence and decision making: a multiple-criteria quadratic programming approach, *International Journal of Information Technology & Decision Making* 4(4): 1–19. http://dx.doi.org/10.1142/S0219622005001775

Shlens, J. 2005. *A tutorial on principal component analysis*. Institute for Nonlinear Science. 13 p.

Shumway, T. 2001. Forecasting bankruptcy more accurately: a simple hazard model, *The Journal of Business* 74(1): 101–124. http://dx.doi.org/10.1086/209665

Sobehart, J.; Keenan, S.; Stein, R. 2001. Benchmarking quantitative default risk models: a validation methodology, *ALGO Research Quarterly* 4(1/2): 57–72.

Srinivasan, V.; Kim, Y. H. 1987. Credit granting: a comparative analysis of classification procedures, *The Journal of Finance* XLII(3): 665–681. http://dx.doi.org/10.1111/j.1540-6261.1987.tb04576.x

Stein, R. M. 2002. Benchmarking default prediction models: pitfalls and remedies in model validation, *Journal of Risk Model Validation* 1(1): 77–113.

Tam, K. Y.; Kiang, M. Y. 1992. Managerial applications of neural networks: the case of bank failure predictions, *Management Science* 38(7): 926–947. http://dx.doi.org/10.1287/mnsc.38.7.926

Tsai, C. 2009. Feature selection in bankruptcy prediction, *Knowledge-Based Systems* 22(2): 120–127. http://dx.doi.org/10.1016/j.knosys.2008.08.002

Tseng, F.-M.; Lin, L. 2005. A quadratic interval logit model for forecasting bankruptcy, *Omega* 33: 85–91.

Zavadskas, E. K.; Turskis, Z. 2011. Multiple criteria decision making (MCDM) methods in economics: an overview, *Technological and Economic Development of Economy* 17(2): 397–427. http://dx.doi.org/10.3846/20294913.2011.593291

Varetto, F. 1998. Genetic algorithm applications in the analysis of insolvency risk, *Journal of Banking and Finance* 22(10–11): 1421–1439.

Verikas, A.; Kalsyte, Z.; Bacauskiene, M.; Gelzinis, A. 2010. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey, *Soft Computing* 14(9): 995–1010. http://dx.doi.org/10.1007/s00500-009-0490-5

Wiginton, J. C. 1980. A note on the comparison of logit and discriminant models of consumer credit behavior, *Journal of Financial and Quantities Analysis* 15(3): 757–770. http://dx.doi.org/10.2307/2330408

Witten, I. H.; Frank, E. 2005. *Data mining: practical machine learning tools and techniques*. 2nd edition. San Francisco: Morgan Kaufmann. 560 p.

Zanakis, S. H.; Solomon, A.; Wishart, N.; Dublish, S. 1998. Multi-attribute decision making: a simulation comparison of select methods, *European Journal of Operational Research* 107(3): 507–529. http://dx.doi.org/10.1016/S0377-2217(97)00147-1

Zhang, D.; Zhou, L. 2004. Discovering golden nuggets: data mining in financial application, *IEEE Transactions on Systems, Man, and Cybernetics* 34(4): 513–522. http://dx.doi.org/10.1109/TSMCC.2004.829279

Gang KOU is a Professor and Executive Dean of School of Business Administration, Southwestern University of Finance and Economics. He is the Managing Editor of International Journal of Information Technology & Decision Making and series Editor of Quantitative Management (Springer). Previously, he was a Professor of School of Management and Economics, University of Electronic Science and Technology of China, and a Research Scientist in Thomson Co., R&D. He received his PhD in Information Technology from the College of Information Science & Technology, University of Nebraska at Omaha; got his Master's degree in Department of Computer Science, University of Nebraska at Omaha; and BS degree in Department of Physics, Tsinghua University, Beijing, China. He has participated in various data mining projects, including data mining for software engineering, network intrusion detection, health insurance fraud detection and credit card portfolio analysis. He has published more than eighty papers in various peer-reviewed journals and conferences. He has been Keynote speaker/workshop chair in several international conferences. He co-chaired Data Mining contest on The Seventh IEEE International Conference on Data Mining 2007 and he is the Program Committee Co-Chair of the 20th International Conference on Multiple Criteria Decision Making (2009) and NCM 2009: 5th International Joint Conference on INC, ICM and IDC. He is also co-editor of special issues of several journals, such as *Journal of Multi Criteria Decision Analysis*, *Decision Support Systems*, *Journal of Supercomputing and Information Sciences*.

Yi PENG is a Professor of School of Management and Economics, University of Electronic Science and Technology of China. She received her PhD in Information Technology from the College of Information Science & Technology, University of Nebraska at Omaha and got her Master's degree in Department of Info. Science & Quality Assurance, University of Nebraska at Omaha and BS degree in Department of Management Information Systems, Sichuan University, China. Her research interests cover knowledge discover in database and data mining, multi-criteria decision making, data mining methods and modelling, knowledge discovery in real-life applications.

Chen LU got his Master's degree in Management Science from School of Management and Economics, University of Electronic Science and Technology of China.