Knowledge-Based Systems 41 (2013) 16–25

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods

Ligang Zhou

Faculty of Management and Administration, Macau University of Science and Technology, Taipa, Macau

ARTICLE INFO

Article history:
Received 9 July 2012
Received in revised form 10 December 2012
Accepted 20 December 2012
Available online 3 January 2013

Keywords:
Bankruptcy prediction
Imbalanced dataset
Undersampling
Oversampling
Classification

ABSTRACT

Corporate bankruptcy prediction is very important for creditors and investors. Most literature improves performance of prediction models by developing and optimizing the quantitative methods. This paper investigates the effect of sampling methods on the performance of quantitative bankruptcy prediction models on real highly imbalanced dataset. Seven sampling methods and five quantitative models are tested on two real highly imbalanced datasets. A comparison of model performance tested on random paired sample set and real imbalanced sample set is also conducted. The experimental results suggest that the proper sampling method in developing prediction models is mainly dependent on the number of bankruptcies in the training sample set.

1. Introduction

When a company applies for a loan from a creditor, the creditor needs to answer a question, "Is it possible that the borrower will go bankrupt and will not repay the loan?" Before an investor make investment in stock of a company, the investor always worries about the bankruptcy of the company which may cause a loss of all investment. Therefore, it is important for creditors and investors to be able to predict corporate bankruptcy.

From a statistical point of view, the corporate bankruptcy prediction problem is a typical classification problem, in which a company is classified into a non-bankrupt class or a bankrupt class in terms of the company's features. The quantitative methods are usually used to the catch the relationship between a company's bankruptcy and its financial information in the most recent fiscal year before its bankruptcy or other information in the same period reflecting the company's operating environment, such as industry position or macroeconomic environment. This relationship is often described as a corporate bankruptcy prediction model (CBPM) which is constructed with a part of historical observations and is evaluated with another part of historical observations. With the assumption that the relationship holds in the future, the model can be used to predict a company's bankruptcy in the future with the currently available information of the company.

The development of these corporate bankruptcy prediction models is a data-fitting based empirical research and the typical processes of models development are shown as Fig. 1. It shows that the performance of models is dependent on a series of processes, such as sampling, features selection, modeling and evaluation criteria [1].

For the features selection in the development of CBPMs, a lot of research has been conducted. Beaver [2] identified 30 different ratios considered to be important factors for forecasting corporate bankruptcy and tested them by a univariate discriminant analysis model on 79 pairs of bankrupt/non-bankrupt firms; the empirical results showed that "working capital funds flow/total assets" and "net income/total assets" were the two most efficient ratios that could correctly classify 90% and 88% of the firms, respectively. Altman [3] selected five ratios, employed a multivariate discriminant analysis model, and tested the model on 33 pairs of bankrupt/non-bankrupt firms. The model could correctly identify 90% of the firms one year prior to failure. The five selected ratios were: working capital/total assets, retained earnings/total assets, EBIT/total assets, market value equity/book value of total debt, and sales/total assets. Ravi Kumar and Ravi [4] reviewed 128 papers and listed more than 500 different variables that have been used for bankruptcy prediction. To obtain models with better predictive performance, many quantitative techniques and methods from statistics and data mining have been employed, such as discriminant analysis [3,5], linear regression [6], decision tree [7], neural networks [8–11], support vector machines [12] and a wide variety of hybrid methods [13–17]. In addition, some new hybrid methods based on fuzzy theory can be potential alternatives to predict corporate bankruptcy [18,19].

As shown in Fig. 1, the performance of bankruptcy prediction model is not only dependent on what features are selected and what quantitative methods are employed, but also dependent on

E-mail address: mrlgzhou@gmail.com

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.12.007

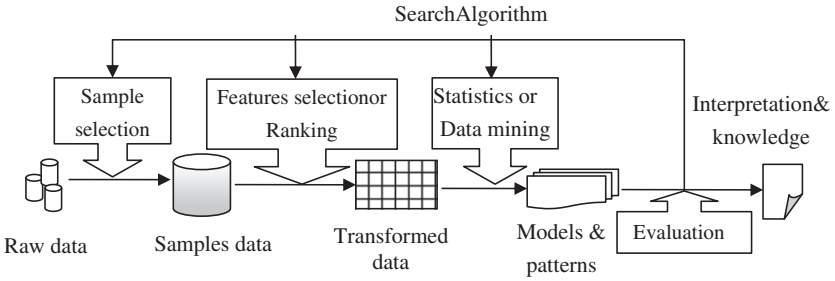

Fig. 1. Development processes of empirical bankruptcy prediction models.

what samples are selected and used for fitting the models. The form of a model is determined by the type of the quantitative approach, linear or nonlinear, implicit or explicit, but the parameters in the models are determined by the data-fitting process and therefore are determined by the selection of samples. Zhou et al. [1] investigated the performance of more than 20 models constructed by different quantitative methods with different features sets selected by six different features ranking techniques. The study just tried to explore what features should be selected and what quantitative methods should be employed in the development of corporate bankruptcy prediction models. It used paired samples as what most research in bankruptcy prediction did. The numbers of bankrupt and non-bankrupt observations are the same in the data set by randomly undersampling the non-bankrupt observations in the original data set. This study is to investigate how the performance of widely used bankruptcy prediction models is affected by different training sample sets which are used to estimate the models and are selected by different sampling strategies.

In the process of sample selection, one simple strategy is to use all available samples. For a large dataset, it will cause the failure of many quantitative approaches due to the unacceptable computational time and space. Another simple strategy is random sampling. If the sample size is not large, it has no problem of computational time and space. However, in practice, the bankrupt cases is very rare, while the number of non-bankrupt cases is very large, therefore, the proportion of bankrupt companies is very close to zero, which lead to a seriously imbalanced classification problem. Both of above two simple strategies without special treatment on the imbalanced samples may cause the quantitative models which always seek an accurate performance over training samples to classify all the test samples into the non-bankrupt class, which is not helpful for decision making.

The classification on imbalanced datasets has received great attention in recent research of data mining because of its wide real applications [20–24]. García et al. investigated the influence of both the imbalance ratio and classifier on the performance of four resampling strategies to deal with imbalanced data sets and found that over-sampling the minority class consistently outperforms under-sampling the majority class when data sets are strongly imbalanced [24]. However, the imbalance property in real datasets for bankruptcy prediction problem has been largely ignored by most literature about bankruptcy prediction. Most research uses dataset with paired samples for training and testing bankruptcy prediction models, in which the number of bankrupt companies is the same as that of the non-bankrupt companies. One important advantage of this strategy is that there is no class bias in the training samples and testing samples, the simple performance measure: classification accuracy on bankrupt and non-bankrupt instances can be used to evaluate the performance of models, which makes the objective of minimizing classification error consistent in the process of model construction (training process) and model evaluation (testing process). However, in real world bankruptcy prediction, the ratio of bankrupt firms to non-bankrupt firms, i.e. degree of imbalance, can be approximately as low as 1 to 100 or even 1 to 1000. There are only a few articles discussing the imbalance problem in bankruptcy prediction. Wilson and Sharda [10] made a comparison of predictive capabilities between neural networks and multivariate discriminant analysis with different degree of imbalance: 50/50, 20/80, and 10/90. They concluded that neural networks outperformed discriminant analysis in classification accuracy and neural network was shown to perform well in predicting both bankrupt firms and non-bankrupt firms when presented with equal numbers of examples in the learning phase. Neves and Vieira [25] tested the effect of different proportions of non-bankrupt firms in the sample to show the performance of an improved neural network model. They only tested three different degrees of imbalance: 50/50, 36/64, 28/72 and selected classification accuracy as the performance measure. Alfaro-Cid et al. [26] tested genetic programming approach incorporating cost matrix for bankruptcy prediction using a highly imbalanced dataset in which about 5–6% of companies went bankrupt. They selected 10 bankrupt firms and 150 non-bankrupt firms as the training set from the total 484 Spanish companies in the database and selected other firms with various numbers of bankrupt and non-bankrupt cases as the training set for different years. As pointed out by the authors, the highly unbalance complicates the classification, but it is an accurate reflection of the real world. Mathiasi Horta et al. [27] discussed some of the main problems in the preparation of models for bankruptcy prediction with the application of data mining techniques and pointed out that the first problem is the class imbalance which causes a poor classification performance and used ensemble strategy to deal with the imbalance problem.

Although Alfaro-Cid et al. [26] and Mathiasi Horta et al. [27] used cost matrix strategy and ensemble strategy to handle imbalance problem in bankruptcy prediction separately, for a real situation of CBPMs construction, with a large imbalanced dataset, the analyst need to know the answer to following questions: what strategies should be used to select the training sample; if the performance of CBPMs will be affected by the sampling strategy and how much will it be affected by the sampling strategy. Just like what Alfaro-Cid et al. [26] did in the selection of training samples, most research in bankruptcy prediction randomly selected a fixed number of cases in the training samples, few of them discuss the effect of sampling for training set on performance on real imbalanced test set.

The main purpose of this paper is to explore the effect of different sampling methods on the performance of CBPMs on real highly imbalanced datasets and make a comparison among several commonly used CBPMs in a real situation. The outline of this paper is as follows. Section 2 describes some popular sampling strategies and brought forward two new sampling strategies that will be used in this research. Section 3 introduces performance measures for imbalanced datasets. Section 4 reports the results of empirical study on two datasets with different sampling methods and quantitative methods. Section 5 concludes the paper and gives some discussion.

Table 1
A simple example of original training dataset.

ID.	F_1	F_2	⋯	F_M	Bankrupt
I_1	1.05	0.03	⋯	0.01	1
I_2	1.20	0.03	⋯	0.04	1
I_3	0.67	0.01	⋯	−0.12	1
I_4	0.57	0.21	⋯	−0.15	1
M_1	0.24	−0.57	⋯	−0.64	0
M_2	8.50	−0.13	⋯	0.88	0
M_3	12.47	−0.10	⋯	0.89	0
⋮	⋮	⋮	⋮	⋮	⋮
M_{10}	0.58	0.01	⋯	−0.10	0

2. Sampling strategies

Some research in machine learning community has addressed the strategy of resampling the original dataset to deal with the issue of class imbalance [28–32]. The commonly used resampling strategies include oversampling and undersampling. Oversampling is to sample the minority class over and over to achieve the balanced distribution of the two classes, while undersampling is to select a portion of the majority class to achieve the distribution balance of the two classes. In the original imbalanced training dataset, let the original sample set of minority class and majority class denoted by S_{mi} and S_{ma} separately, the size of minority class $|S_{mi}|$ is much less than the size of majority class $|S_{ma}|$. The training dataset is denoted by S. A simple example of an original training dataset for bankruptcy corporate prediction is shown in Table 1. This example is just for the purpose to explain sampling strategy and it is imbalanced but not highly imbalanced. Each observation has M features and an observed financial status: Bankrupt or Non-bankrupt in the next year. The bankrupt observation is marked with value 1 in the "Bankrupt" column while the non-bankrupt observation is marked with value 0. Each observation has a unique ID and the bankrupt observation with an ID staring with a letter "I" and non-bankrupt observation starting with a letter "M". In this example, bankrupt observations belong to the minority class and non-bankrupt observations belong to the majority class. Therefore, the set of minority class $S_{mi} = \{I_1, I_2, I_3, I_4\}$ and $|S_{mi}| = 4$, the set of majority class $S_{ma} = \{M_1, M_2, \ldots, M_{10}\}$ and $|S_{ma}| = 10$.

2.1. Oversampling

Two widely used oversampling methods: Random Oversampling with Replication (ROWR) and Synthetic Minority Over-sampling Technique (SMOTE) are employed in this study. ROWR is the method to balance class distribution through the random replication of minority class examples [32].

2.1.1. Random oversampling with replication (ROWR)

Algorithm 1. ROWR (S_{mi},S_{ma})

Input: The original sample set of minority class S_{mi} and sample set of majority class S_{ma}.
Output: sample set with balanced class S of size $2|S_{ma}|$.
1. $S = S_{ma}$;
2. **for** $i = 1$ ***to*** $|S_{ma}|$
   randomly selected an element $\boldsymbol{a}_i$ from sample set S_{mi}
   $S = S \cup \{\boldsymbol{a}_i\}$.
   endfor

For the original training dataset shown in Table 1, an example of the final training set S may like the follows: $\{M_1, M_2, \ldots, M_{10}, I_3, I_1, I_4, I_2, I_2, I_1, I_3, I_4, I_3, I_1\}$.

2.1.2. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE, proposed by Chawla et al. [32], is an improved oversampling approach in which the minority class is oversampled by creating "synthetic" examples rather than by oversampling with replacement. The main idea of SMOTE is to oversample the minority class by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbor. The detail of SMOTE is as follows [32]:

Algorithm 2. SMOTE ($|S_{mi}|, N, k$)

Input: Number of minority class samples $|S_{mi}|$; Amount of minority class being oversampled N%; Number of nearest neighbors k
Output: Union of $(N/100) \times |S_{mi}|$ synthetic minority class samples and the majority set S_{ma}
1. **if** $N < 100$ **then**
   Randomize the $|S_{mi}|$ minority class samples
   $T = (N/100) \times |S_{mi}|$
   $N = 100$
   endif
2. $N =$ $(int)(N/100)$
3. k = number of neighbors
4. numattrs = Number of attributes
5. Sample[][]: array for original minority class samples
6. newindex = 0//keeps a count of number of synthetic samples generated
7. Synthetic[][]: array for synthetic samples
8. **for** $i = 1$ *to* $|S_{mi}|$
   Compute k nearest neighbors for i, and save the indices in the *nnarray*
   Populate ($N, i, nnarray, newindex, Synthetic$)
   endfor
   Populate ($N, i, nnarray, newindex, Synthetic$)
9. **while** $N \neq 0$
   nn = random number between 1 and k
   for *attr* = 1 to *numattrs*
   $dif = Sample[nnarray[nn]][attr] - Sample[nnarray[i]][attr]$
   gap = random number between 0 and 1
   $Synthetic[newindex][attr] = Sample[i][attr] + gap \times dif$
   endfor
   newindex++
   $N = N - 1$
   Endwhile

For the original training dataset shown in Table 1, suppose the parameters in SMOTE algorithm are set as follows:$|S_{mi}| = 4$, $N = 300$, $k = 2$, then the output of SMOTE is a training set with a total of 22 observations including 12 bankrupt observations and 10 non-bankrupt observations $\{M_1, M_2, \ldots, M_{10}\}$. The most import step to generate the 12 bankrupt observations is the step 8 in above SMOTE. Suppose the i^{th} feature of observation with ID I_j is denoted by F_{j_i}. For each bankrupt observation, the step 8 generate another $N/100 = 3$ observations which are so called "synthetic" examples. For I_1, suppose its two nearest neighbors are I_2 and I_4 and the random number nn in step 9 in SMOTE is 2, then one synthetic example $\widehat{I}_1^1$ for I_1 is generated as follows:

$$F_{1-1} = 1.05, F_{1-2} = 0.03, \ldots, F_{1-M} = 0.01$$
$$F_{4-1} = 0.57, F_{4-2} = 0.21, \ldots, F_{4-M} = -0.15$$
$$\Delta F_1 = 0.57 - 1.05 = -0.58, \Delta F_2 = 0.21 - 0.03 = 0.18, \ldots,$$
$$\Delta F_i = F_{4-i} - F_{1-i}, \ldots, \Delta F_M = F_{4-M} - F_{1-M}$$
$$(F_1, F_2, \ldots, F_i, \ldots, F_M) = (1.05, 0.03, \ldots, F_{1-i}, \ldots, 0.01) + Rand() \times (\Delta F_1, \Delta F_2, \ldots, \Delta F_i, \ldots, \Delta F_M)$$

Rand() is a function to generate any real number between 0 and 1.

2.2. Undersampling

2.2.1. Random Undersampling (RU)

Random under-sampling method is to balance class distribution through the random elimination of majority class examples [33]. RU employed in this study is to randomly select part of the majority class to achieve the balance which obtains the equivalent result.

Algorithm 3. RU (S_{mi},S_{ma})

Input: The original sample set of minority class S_{mi} and sample set of majority class S_{ma}.
Output: sample set with balanced class S of size $2|S_{mi}|$
1. $S = S_{mi}$;
2. **for** $i = 1$ ***to*** $|S_{mi}|$
   randomly selected an element $\boldsymbol{a}_i$ from sample set S_{ma}
   $S = S \cup \{\boldsymbol{a}_i\}$.
   endfor

For the original dataset in Table 1, one training set S generated by RU like follows: $\{I_1,I_2,I_3,I_4,M_8,M_3,M_7,M_5\}$ and its size is 8.

2.2.2. Undersampling Based on Clustering from Nearest Neighbor (UBOCFNN)

The idea of UBOCFNN is inspired by clustering problem to partition the points in sample space into K clusters and select the point which is the nearest to the central point of each cluster to represent the whole cluster. The detail of UBOCFNN is as follows:

Algorithm 4. UBOCFNN (S_{mi},S_{ma})

Input: The original sample set of minority class S_{mi} and sample set of majority class S_{ma}.
Output: sample set with balanced class S of size $2|S_{mi}|$.
1. $S = S_{ma}$
2. Partition the points in S_{ma} into $|S_{mi}|$ clusters with following two phases [34]:
   2.1. Uses batch updates, where each iteration consists of reassigning points to their nearest cluster centroid, all at once, followed by recalculation of cluster centroids.
   2.2. Uses online updates, where points are individually reassigned if doing so will reduce the sum of distances, and cluster centroids are recomputed after each reassignment.
3. **for** $i = 1$ **to** $|S_{mi}|$
   Find the central point $\boldsymbol{c}_i$ of cluster i
   Select the point $\boldsymbol{a}_i$ which is the nearest point in cluster i to the point $\boldsymbol{c}_i$
   $S = S \cup \{\boldsymbol{a}_i\}$
   endfor

For the original dataset in Table 1, suppose the non-bankrupt observations are clustered into four clusters as follows $\{M_2,M_3\}$, $\{M_1,M_{10},M_7\}$, $\{M_4,M_6,M_8,M_9\}$, $\{M_5\}$. It is easy to calculate the central point of each cluster, the coordinates of the central point is the average corresponding coordinates of all points in the cluster. Then the nearest point to the central point in each cluster can be found. One example of the final training set may like $S = \{I_1,I_2,I_3,I_4,M_2,M_7,M_6,M_5\}$.

2.2.3. Undersampling Based on Clustering from Gaussian Mixture Distribution (UBOCFGMD)

The idea of UBOCFGMD is similar to UBOCFNN, except that the clustering is based on Gaussian mixture distribution. The Expectation Maximization (EM) algorithm is adopted to estimate the parameters in a Gaussian mixture model with K components for data in S_{ma} then the data in S_{ma} will be partitioned into K clusters in terms of the K components of Gaussian mixture distribution [34]. In this study, we set K in UBOCFGMD and UBOCFNN to be $|S_{mi}|$. For UBOCFNN will select the point which is nearest to the central point of each cluster and then we can obtain a training set with size of $2|S_{mi}|$. But for UBOCFGMD, it may happen that there is no points partitioned into one cluster. In this case, for each cluster with points, we randomly select one to represent that cluster and for the clusters without points, we randomly select one from the nearest cluster with points. Here the nearest cluster is the one whose cluster number is closest to the cluster number of the concerned cluster.

For the original dataset in Table 1, suppose the non-bankrupt observations are clustered into four clusters based on Gaussian mixture distribution as follows: $\{M_1,M_3\}$, $\{M_2,M_{10},M_8\}$, $\{M_4,M_6,M_5,M_7,M_9\}$, $\{\}$. The fourth cluster happens to be null. The points selected for each cluster is similar to the process in UBOCFNN and for the fourth cluster, a point from cluster 3 is randomly selected to represent the fourth cluster. One example of S may be $\{I_1,I_2,I_3,I_4,M_1,M_{10},M_6,M_5\}$.

3. Performance measures

Five common performance measures as follows are selected to evaluate the performance of the models:

(1) Sensitivity (Sen) $= \frac{NN}{NB+NN}$
(2) Specificity (Spe) $= \frac{BB}{BN+BB}$
(3) Accuracy (Acc) $= \frac{NN+BB}{NB+NN+BB+BN}$
   where NN: non-bankruptcy classified as bankruptcy, NB: non-bankruptcy classified as bankruptcy, BB: bankruptcy classified as bankruptcy, BN: bankruptcy classified as non-bankruptcy.
(4) F-measure (F) $= \frac{2\times Precision\times Recall}{(Precision+Recall)}$ [35]
   where Precision $= \frac{NN}{NN+BN}$, Recall = Sensitivity. Precision shows what proportion of observations classified as non-bankruptcy by a model is real non-bankrupt, while Recall/Sensitivity shows what proportion of the real non-bankrupt observations has been correctly classified as non-bankruptcy by the models.
(5) Area under ROC curve (AUC): ROC graphs are two-dimensional graph in which Sensitivity is plotted on the Y axis and 1-Specificity is plotted on X axis. An ROC graph depicts relative trade-off between benefits (true non-bankruptcy) and costs (false non-bankruptcy), which is useful for organizing classifiers and visualizing their performance especially in the domains with skewed class distribution and unequal classification error costs. The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [36].

4. Empirical study

4.1. USA Bankruptcy Dataset (USABD)

The initial USABD consists of all firms from non-financial industry with observed financial status (Non-bankrupt or Bankrupt) from 1981 to 2009 in Compustat North America dataset provided

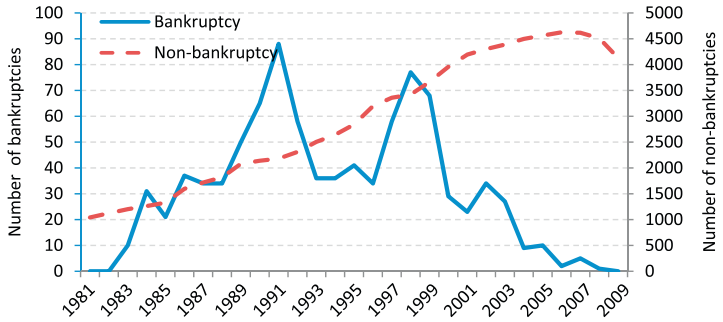

Fig. 2. Bankruptcy and Non-bankruptcy by year in USABD.

Table 2
Sample size of training set and test set by different sampling methods for USABD.

Sampling methods		ROWR	SMOTE *N = 100*	RU	UBOCFNN & UBOCFGMD	AS
Training set size	Bankruptcy	49,581	1660	830	830	830
	Non-bankruptcy	49,581	49,581	830	830	49,581
Test set size	Bankruptcy	88	88	88	88	88
	Non-bankruptcy	35,630	35,630	35,630/88	35,630	35,630

Table 3
Performance of models on USABD by sampling methods.

Models	ROWR					SMOTE				
	Sen	Spe	Acc	F	AUC	Sen	Spe	Acc	F	AUC
LDA	0.7710	0.5997	0.7706	0.8702	0.7159	0.7329	0.4943	0.7323	0.8452	0.6687
LOGR	0.6578	0.7311	0.6580	0.7933	0.7334	0.9970	0.0170	0.9946	0.9973	0.6862
DT	0.9764	0.1098	0.9743	0.9870	0.5433	0.9539	0.1818	0.9520	0.9751	0.6482
NN	–	–	–	–	–	0.9574	0.2386	0.9556	0.9769	0.7762
SVM	–	–	–	–	–	–	–	–	–	–
	RU (on imbalanced test set)					RU (on balanced test set)				
LDA	0.7756	0.5739	0.7751	0.8730	0.7047	0.7125	0.5864	0.6494	0.6703	0.6852
LOGR	0.4762	0.7682	0.4769	0.5830	0.6428	0.6341	0.6705	0.6523	0.6102	0.6746
DT	0.6639	0.7614	0.6641	0.7967	0.7460	0.6250	0.7875	0.7063	0.6805	0.7424
NN	0.7020	0.7750	0.7021	0.8240	0.7881	0.6636	0.7545	0.7091	0.6956	0.7519
SVM	0.7639	0.7250	0.7638	0.8658	0.7965	0.7398	0.7364	0.7381	0.7385	0.7739
	UBOCFNN					UBOCFGMD				
LDA	0.0196	0.9659	0.0219	0.0384	0.6181	0.7851	0.6250	0.7847	0.8792	0.7437
LOGR	0.1809	0.8182	0.1824	0.3062	0.6226	0.1523	0.9432	0.1543	0.2643	0.5477
DT	0.5587	0.6250	0.5589	0.7165	0.6169	0.6472	0.8182	0.6476	0.7856	0.7286
NN	0.6988	0.5341	0.6984	0.8222	0.6232	0.6739	0.8182	0.6742	0.8049	0.7805
SVM	0.1107	0.8409	0.1125	0.1993	0.5445	0.7479	0.7386	0.7478	0.8554	0.7971

	AS				
	Sen	Spe	Acc	F	AUC
LDA	0.7109	0.4886	0.7103	0.8304	0.6385
LOGR	0.9993	0.0000	0.9968	0.9984	0.7220
DT	1.0000	0.0000	0.9975	0.9988	0.5000
NN	0.9999	0.0000	0.9975	0.9987	0.7622
SVM	–	–	–	–	–

– Denotes failure of model construction due to the overflow of memory.

by Wharton Research Data Service. The bankrupt company is defined as one whose reason for deletion is marked as "bankruptcy" or "liquidation" in the original Compustat North America dataset.

There is a wide variety of financial information declared by the company, therefore, it is always challenging to identify which group of information is effective in predicting bankruptcy. Zhou et al. explored what information should be selected for bankruptcy prediction with different features ranking strategies for bankruptcy prediction [1]. Tsai [37] explored some features selection methods based on statistical characteristics for bankruptcy prediction, while Pacheco [38,39] and Unler [40] employed heuristic optimization methods to select features for models in financial applications. However, we only restrict this study to those variables which are well accepted as explanatory variable in

Table 4
p Values of Wilcoxon signed rank test between pairs of sampling methods on USABD.

	ROWR	SMOTE	RU	UBOCFNN	UBOCFGMD	AS
ROWR	1.000	1.000	1.000	0.500	1.000	0.250
SMOTE		1.000	0.625	0.125	0.875	0.625
RU			1.000	0.063*	0.625	0.625
UBOCFNN				1.000	0.125	0.625
UBOCFGMD					1.000	0.625
AS						1.000

* Indicates significant difference at 0.1 significance level.

bankruptcy prediction models. The explanatory variables are those used with the highest frequency in the 128 reviewed papers by Ravi Kumar and Ravi [4]. All these variables are financial ratios which can be computed in terms of financial items declared in the financial statements. The 10 explanatory variables selected in this study include: net income/total assets (NI/TA), current ratio (CR), retained earnings/total assets (RE/TA), working capital/total assets (WC/TA), EBIT/total assets (EBIT/TA), sales/total assets (S/TA), cash/total assets (C/TA), current assets/total assets (CA/TA), stock holder's equity/total debt (SHE/TD), cash/current liabilities (C/CL). This set of variables includes the five important and widely used variables proposed by Altman [3].

The empirical study of this research is consistent with most of the previous literature. All prediction is conducted on base of yearly financial reports. The available financial data from a firm in the most recent fiscal year is used to do the bankruptcy prediction at the time of estimation. For the selected 10 variables, in the case of missing value, the missing value will be imputed with the previous available value. If there is no available value in all the previous years, then this case will be removed.

Fig. 2 shows the number of bankruptcies and non-bankruptcies by year over the sample period. As seen, the number of bankruptcies has been increasing over time, with the largest number of bankruptcies occurring in the late 1980s and early 1990s and in 1997 and 1998. Although subprime crisis happened in USA in 2007, since most bankrupt firms are in financial industries and some firms were officially declared as bankruptcy after 2009, there are only a couple of bankrupt firms after 2007 in this data set. Finally, there are a total of 918 bankrupt observations and 85,211 non-bankrupt observations from observed year 1981 to 2009. It shows that the dataset is highly imbalanced and the degree of imbalance ranges from 0/4128 (0) to 88/2181 (0.0403) over years.

To test the performance of different sampling and quantitative models, the total sample space is split into training sample set and test sample set. The training set consists of all observations from observed year 1981 to 2001. All models are estimated with the same training set and are used in predicting bankruptcies for the period 2002–2009.

The employed quantitative methods include linear discriminant analysis (LDA), logistic regression (LOGR), decision tree C4.5 (DT), neural network (NN) and support vector machines (SVM). The decision tree C4.5 is implemented with the J48 function provided by Weka [35]. Other models are implemented with Matlab on PC with 2G RAM. The neural network model is a three layer back propagation network. The support vector machine model adopts the least square SVM proposed by Suykens, et al. [41]. The parameter k is set to be 5 in the SMOTE sampling methods while N is tested with value of 100, 500, 1000, 2000. The SMOTE methods with different N are denoted by SMOTE100, SMOTE500, SMOTE1K and SMOTE2K respectively.

For ROWR method, it just randomly duplicates the instances in minority class until the balance of two classes in the final training sample set. SMOTE generates new instances of minority class in terms of the parameter N. Both underampling method UBOCFNN and UBOCFGMD select a proportion of majority class to keep balance of two classes in the final training sample set. The simple sampling strategy to select all observations directly as the training set is also tested. This strategy is denoted by AS (all sample). The

Table 5
p Values of Wilcoxon signed rank test on performance measures of Groups 1 and 2 on USABD.

	Sen	Spe	Acc	F	AUC
p Values	0.0049*	0.6611	0.1909	0.0000*	0.0043*

* Indicates significant different at 0.1 significance level.

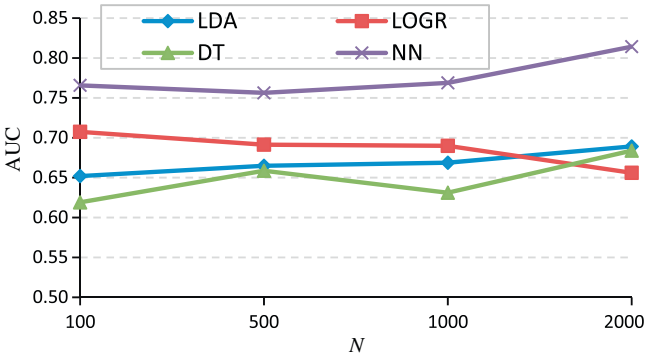

Fig. 3. AUC performance of models trained by different training set obtained by SMOTE sampling method with different N on USABD.

Table 6
Summary statistics of AUC performance on imbalanced test of five methods trained by 10 different sample sets obtained with RU method on USABD.

	Min	Max	Mean	Median	Std
LDA	0.6619	0.7265	0.7047	0.7082	0.0197
LOGR	0.4955	0.7813	0.6428	0.6381	0.1162
DT	0.6931	0.7737	0.7460	0.7485	0.0234
NN	0.7546	0.8173	0.7881	0.7897	0.0188
SVM	0.7679	0.8080	0.7965	0.7999	0.0112

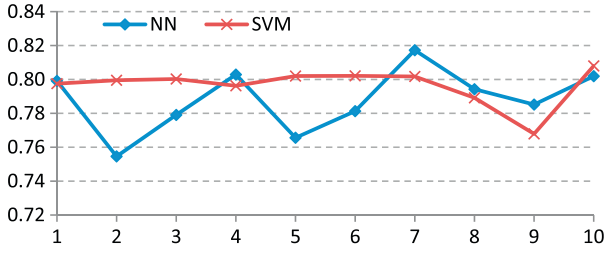

Fig. 4. AUC of NN and SVM trained by sample set obtained with RU sampling method on USABD.

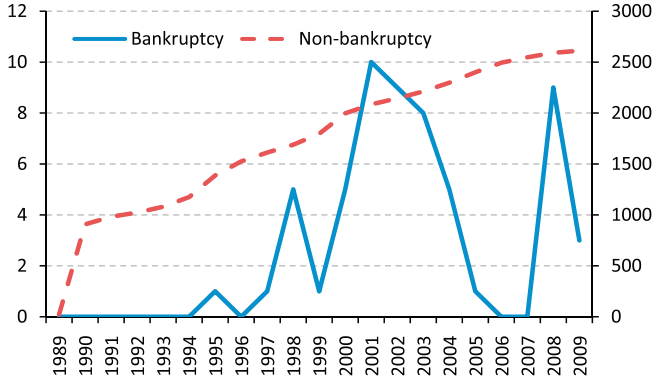

Fig. 5. Bankruptcy and Non-bankruptcy by year in JPNBD.

observations in or before 2001 compose the training sample set and others make the test sample set. The number of sample in the training set and test set from each sampling method is shown as Table 2. Since most previous literature uses random undersampling strategy and tests models on paired samples, to make the comparison of performance of the models on paired test sample set and highly imbalanced test set in real situation, the RU sampling strategy is tested on test sample set with paired sample and highly imbalanced sample set.

Table 7
Sample size of training set and test set by different sampling method for JPNBD.

Sampling methods		ROWR	SMOTE $N = 100$	RU	UBOCFNN & UBOCFGMD	AS
Training set size	Bankruptcy	17,280	46	23	23	23
	Non-bankruptcy	17,280	17,280	23	23	17,280
Test set size	Bankruptcy	35	35	35	35	35
	Non-bankruptcy	19,298	19,298	19,298/35	19,298	19,298

Table 8
Performance of models on JPNBD by sampling methods.

Models	ROWR					SMOTE				
	Sen	Spe	Acc	F	AUC	Sen	Spe	Acc	F	AUC
LDA	0.8462	0.7714	0.8461	0.9165	0.8955	0.9851	0.3714	0.9840	0.9919	0.8401
LOGR	0.9185	0.7238	0.9181	0.9573	0.9195	0.9967	0.1429	0.9952	0.9976	0.9269
DT	0.9948	0.2413	0.9935	0.9967	0.6184	0.9954	0.1929	0.9940	0.9970	0.6339
NN	0.9597	0.4952	0.9588	0.9789	0.8113	0.9954	0.1857	0.9939	0.9969	0.7859
SVM	–	–	–	–	–	–	–	–	–	–
	RU (on imbalanced test set)					RU (on balanced test set)				
LDA0.7428	0.8400	0.7430	0.8508	0.8748	0.7143	0.8429	0.7786	0.7636	0.8682	
LOGR	0.8316	0.7229	0.8314	0.9061	0.8081	0.8943	0.7171	0.8057	0.8212	0.8717
DT	0.8064	0.8257	0.8065	0.8904	0.8058	0.7657	0.7971	0.7814	0.7768	0.7747
NN	0.8447	0.6714	0.8444	0.9145	0.7853	0.8429	0.6429	0.7429	0.7693	0.7737
SVM	0.7022	0.7857	0.7024	0.8200	0.8467	0.7657	0.7457	0.7557	0.7561	0.8571
	UBOCFNN					UBOCFGMD				
LDA	0.6307	0.8857	0.6311	0.7734	0.8873	0.6926	0.7429	0.6927	0.8182	0.7960
LOGR	0.8509	0.8857	0.8510	0.9194	0.8787	0.7072	0.7714	0.7073	0.8283	0.7393
DT0.6944	0.8286	0.6946	0.8195	0.7615	0.8693	0.8000	0.8692	0.9299	0.8347	
NN	0.8064	0.7429	0.8063	0.8926	0.8130	0.8247	0.6857	0.8245	0.9037	0.8467
SVM	0.6523	0.9429	0.6528	0.7895	0.8755	0.8054	0.7714	0.8054	0.8920	0.8569

	AS				
	Sen	Spe	Acc	F	AUC
LDA	0.9934	0.2000	0.9919	0.9959	0.8373
LOGR	0.9988	0.0571	0.9971	0.9985	0.9198
DT	0.9998	0.0571	0.9981	0.9991	0.6276
NN	0.9989	0.1143	0.9973	0.9987	0.6726
SVM	–	–	–	–	–

– Denotes failure of model construction due to the overflow of memory.

The results of predictive performance of above five quantitative models for test observations in year 2002–2009 are listed in Table 3. For ROWR and RU sampling strategies, the results are the average value of 10 iterations. For SMOTE with different values of *N*, the results are the average value of performance with different *N*.

To make a comparison among these sampling methods, Wilcoxon Signed-Rank test is employed. Unlike student test, this non-parametric test method does not require assumption of normal distribution of the random variable. Moreover, some empirical results suggest that it is also stronger than student test especially in the comparison of a pair of classifiers [42]. Although both F-measure and AUC are commonly used measures for imbalanced dataset, all groups of test in this experiment show that the correlation coefficient of these two measure is only 0.2960. F-measure, like Sen, Spe and Acc, is determined by a fixed threshold value which defines the boundary value to classify the instance into bankruptcy or non-bankruptcy. AUC shows the relationship of Sen and 1-Spe with the change of threshold, the point with coordination of (Sen, 1-Spe) is a just one point in the ROC graph with a default threshold in the quantitative models, such as 0 for logistic regression in this study. In data mining community, AUC is the most proper performance measure for the imbalanced dataset, therefore, statistical comparison of sampling methods is conducted in terms of AUC. Table 4 shows the *p* values of Wilcoxon signed rank tests between pairs of all the sampling methods that are tested on imbalanced test set with 88 bankrupt observations and 35,630 non-bankrupt observations. As seen, only AUC by UBOCFNN sampling method is significant less than that by RU. Since oversampling method ROWR and SMOTE and directly sampling all method (AS) cannot significantly increase performance of models, with a view to computational time, the undersampling methods:

Table 9
p Values of Wilcoxon signed rank test between pairs of sampling methods on USABD.

	ROWR	SMOTE	RU	UBOCFNN	UBOCFGMD	AS
ROWR	1.0000	0.6250	0.8750	1.0000	1.0000	0.6250
SMOTE		1.0000	0.8750	0.6250	0.8750	0.1250
RU			1.0000	0.4375	0.8125	0.3750
UBOCFNN				1.0000	0.6250	0.2500
UBOCFGMD					1.0000	0.6250
AS						1.0000

Table 10
p Values of Wilcoxon signed rank test on performance measures of Groups 1 and 2 on USABD.

	Sen	Spe	Acc	F	AUC
p Values	0.8431	0.3235	0.2690	0.0000*	0.8205

* Indicates significant different at 0.1 significance level.

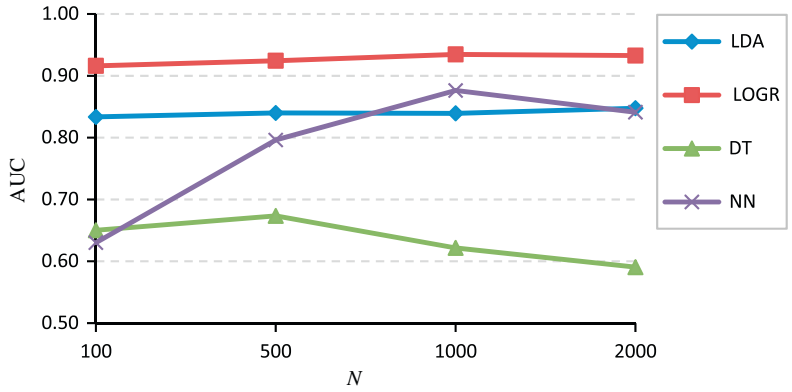

Fig. 6. AUC performance of models trained by different training set obtained by SMOTE sampling method with different *N* on JPNBD.

Table 11
Summary statistics of AUC performance on imbalanced test of five methods trained by 10 different sample sets obtained with RU method on USABD.

	Min	Max	Mean	Median	Std
LDA	0.7603	0.9342	0.8748	0.8868	0.0550
LOGR	0.6105	0.9216	0.8081	0.8207	0.0964
DT	0.6822	0.8738	0.8058	0.8167	0.0554
NN	0.6578	0.8554	0.7853	0.7868	0.0609
SVM	0.7925	0.8958	0.8467	0.8383	0.0339

RU, UBOCFGMD both are better choices. All five models perform badly with UBOCFNN sampling method. Among the five quantitative methods, NN and SVM perform better than other methods with the RU and UBOCFGMD sampling methods.

Most previous literature trains and tests models with balanced sample by random undersampling method, but in real world, the test sample set is highly imbalanced. Therefore, to see if there is any difference on performance of models trained by the balanced sample but tested on balanced and highly imbalanced test set, Wilcoxon signed rank test is employ to make a comparison between two groups of results: Group 1 consists of all performance results from five models (LDA, LOGR, DT, NN, SVM) trained by 1660 instances (830 bankruptcies and 830 non-bankruptcies) and tested on 35,718 instances (88 bankruptcies and 35,630 non-bankruptcies); Group 2 consists of all performance results from the same five models trained by same training set as Group 1 but tested on 176 instances (88 bankruptcies and 88 non-bankruptcies). The *p* values of Wilcoxon signed rank test on different performance measures of these two groups are listed in Table 5. It shows that there is significant difference on the performance of Sensitivity, F-measure and AUC. It is interesting to observe in this experiment that above three performance measures of models tested on paired sample is underestimated when compared to those tested on highly imbalanced sample. Although the difference between them is slight, it is statistically significant.

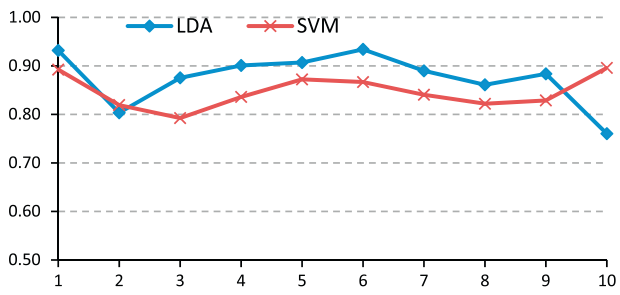

Fig. 7. AUC of NN and SVM trained by sample set obtained with RU sampling method on USABD.

Fig. 3 shows the AUC of four models trained with sample set obtained by SMOTE sampling method with different parameter *N*. There is no clear relationship between the parameter *N* and AUC performance, but NN achieves the biggest AUC (0.8142) with $N = 2000$.

The summary statistics of AUC performance on imbalanced test set of five methods trained by 10 different sample sets obtained with RU method is show in Table 6. It shows that SVM model gets the best mean AUC with the lowest standard deviation in the 10 iterations of test with different training sample sets. Fig. 4 shows the AUC of NN and SVM of the 10 iterations of tests. It indicates that both NN and SVM can perform stably even they are trained with different randomly paired samples.

4.2. Japanese Bankruptcy Dataset (JPNBD)

Japanese Bankruptcy Dataset is smaller than USABD. The samples in JPNBDS are retrieved from Compustat Global, Wharton Research Data Service. Only non-financial firms are included. JPNBDS include samples with observed financial status (Non-Bankrupt or Bankrupt) from 1989 to 2009. The bankrupt company is defined as "bankruptcy" or "liquidation" in the original database. Fig. 5 shows the number of bankruptcies and non-bankruptcies by year over the sample period. Finally, there are a total of 58 bankrupt observations and 36,578 non-bankrupt observations from observed year 1989 to 2009. It shows that the dataset is highly imbalanced and the degree of imbalance ranges from 0/2547 (0) to 10/2085 (0.0048) over years.

The variables selection and models setting on JPNBD is the same as that for USABD. The observations in or before 2001 compose the training sample space and others make the test sample set. The number of sample in the training set and test set from each sampling method is shown as Table 7. The number of final selected sample by the undersampling methods is only 46 in the training sample set including 23 bankruptcies and 23 non-bankruptcies; therefore, it is a classification problem with small-size sample.

The results of predictive performance of the five quantitative models for test observations in year 2002–2009 are listed in Table 8. For ROWR and RU sampling strategies, the results are the average value of 10 runs. For SMOTE with different values of *N*, the results are the average value of performance with different *N*. Table 9 shows the *p* values of Wilcoxon signed rank tests between two of all the sampling methods that tested on imbalanced test set with 35 bankrupt observations and 19,298 non-bankrupt observations. It indicates that there is no statistically significant difference among these sampling methods at 0.1 significance level. Table 10

shows the difference of performance between two groups of test similar to that in USABD. Group 1 consists of all performance results from five models (LDA, LOGR, DT, NN, SVM) trained by 46 instances (23 bankruptcies and 23 non-bankruptcies) and tested on 19,333 instances (35 bankruptcies and 19,298 non-bankruptcies); Group 2 consists of all performance results from the same five models trained by the same training set as Group 1 but tested on 70 instances (35 bankruptcies and 35 non-bankruptcies). As seen in Table 10, there is no significant difference between the two group test on all performance measure except F-measure, which indicates that it would be proper to use the paired training sample and test sample in the models development and models selection. Among all test with different sampling methods and quantitative models, the LOGR with SMOTE sampling strategy achieves the best AUC performance (0.9269) and Fig. 6 shows the AUC of four models trained with sample set obtained by SMOTE sampling method with different parameter *N*. There is no clear relationship between the parameter *N* and AUC performance, but LOGR and LDA perform stably with different *N*. Therefore, in the case that there is only dozens of minority class sample, oversampling method could be better choice for the model development. In this experiment, SVM failed in the oversampling method due to high computational space requirement, a combination of SMOTE and undersampling strategy may reduce the total sample size and therefore reduce the space requirement in SVM model construction.

The summary statistics of AUC performance on imbalanced test set of five methods trained by 10 different sample sets obtained with RU method is show in Table 11. It shows that LDA model get the greatest mean AUC and followed by SVM in the 10 runs of test with different training sample sets. Fig. 7 shows the AUC of LDA and SVM of the 10 runs of tests. It indicates that both LDA and SVM can perform stably even they are trained with different paired samples obtained by RU method.

5. Conclusion

This paper investigates the effect of six different sampling methods on the performance of five quantitative bankruptcy prediction models. Each sampling method and quantitative model is tested on two datasets. The experimental results shows that when there are hundreds of bankrupt observations in the dataset, undersampling method is better than oversampling method because there is no significant difference on performance but oversampling method consumes more computational time. When there are only dozens of bankrupt cases in the dataset, oversampling method SMOTE is a better choice and if the training sample size is too large to cause the failure of model construction, the combination of SMOTE and undersampling maybe an alternative. In the test on both datasets, it is interesting to observe that the difference of AUC performance of all models, trained by sample set obtained by random undersample method, tested on random paired sample and real highly imbalanced sample is very slight or not significant, therefore, in the bankruptcy prediction model selection, the models can be evaluated and measured on their performance on random balanced sample set instead of the real highly imbalanced test sample.

This paper mainly focuses on the sampling method for bankruptcy prediction model construction with highly imbalanced dataset. All tested quantitative models just adopt the fundamental form and have no parameters and model optimization. The performance of models varies with the sampling method, but SVM can achieve good performance in most scenarios. There are a lot of bankruptcy prediction models, in practice, model selection process should be conducted since no model can always perform well. How sample distribution affects the power of prediction models? Can we identify the difficult and easy observation for test in terms of characteristics of training sample and prediction models? These problems will be our future research.

References

[1] L. Zhou, K.K. Lai, J. Yen, Empirical models based on features ranking techniques for corporate financial distress prediction, Computers & Mathematics with Applications 64 (2012) 2484–2496.
[2] W.H. Beaver, Financial ratios as predictors of failure, Journal of Accounting Research 4 (1966) 71–111.
[3] E.I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, Journal of Finance 23 (1968) 589–609.
[4] P. Ravi Kumar, V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review, European Journal of Operational Research 180 (2007) 1–28.
[5] R.A. Collins, R.D. Green, Statistical methods for bankruptcy forecasting, Journal of Economics and Business 34 (1982) 349–354.
[6] R. Collins, An empirical comparison of bankruptcy prediction models, Financial Management 9 (1980) 52–57.
[7] A. Gepp, K. Kumar, S. Bhattacharya, Business failure prediction using decision trees, Journal of Forecasting 29 (2009) 536–555.
[8] K.Y. Tam, Neural network models and the prediction of bank bankruptcy, Omega 19 (1991) 429–445.
[9] P. Coats, L. Fant, Recognizing financial distress patterns using a neural network tool, Financial Management 22 (1993) 142–155.
[10]] R. Wilson, R. Sharda, Bankruptcy prediction using neural networks, Decision Support Systems 11 (1994) 545–557.
[11] M. Leshno, Y. Spector, Neural network prediction analysis: the bankruptcy case, Neurocomputing 10 (1996) 125–147.
[12] K.S. Shin, K.J. Lee, H.J. Kim, Support vector machines approach to pattern detection in bankruptcy prediction and its contingency, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2004) 1254–1259.
[13] H. Ahn, K.-J. Kim, Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach, Applied Soft Computing 9 (2009) 599–607.
[14] E. Alfaro, N. García, M. Gámez, D. Elizondo, Bankruptcy forecasting: an empirical comparison of AdaBoost and neural networks, Decision Support Systems 45 (2008) 110–122.
[15] S.H. Min, J. Lee, I. Han, Hybrid genetic algorithms and support vector machines for bankruptcy prediction, Expert Systems with Applications 31 (2006) 652–660.
[16] H. Jo, I. Han, Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction, Expert Systems with Applications 11 (1996) 415–422.
[17] H.L. Chen, B. Yang, G. Wang, J. Liu, X. Xu, S.J. Wang, D.Y. Liu, A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method, Knowledge-Based Systems 24 (2011) 1348–1359.
[18] Y.-C. Hu, Determining membership functions and minimum fuzzy support in finding fuzzy association rules for classification problems, Knowledge-Based Systems 19 (2006) 57–66.
[19] Y.-C. Ko, H. Fujita, G.-H. Tzeng, An extended fuzzy measure on competitiveness correlation based on WCY 2011, Knowledge-Based Systems 37 (2013) 86–93.
[20] M.D. Pérez-Godoy, A. Fernández, A.J. Rivera, M.J. del Jesus, Analysis of an evolutionary RBFN design algorithm* CO2RBFN, for imbalanced data sets, Pattern Recognition Letters 31 (2010) 2375–2388.
[21] J. Van Hulse, T. Khoshgoftaar, Knowledge discovery from imbalanced and noisy data, Data & Knowledge Engineering 68 (2009) 1513–1542.
[22] Y. Sun, M.S. Kamel, A.K.C. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition 40 (2007) 3358–3378.
[23] S. García, J. Derrac, I. Triguero, C.J. Carmona, F. Herrera, Evolutionary-based selection of generalized instances for imbalanced classification, Knowledge-Based Systems 25 (2012) 3–12.
[24] V. García, J.S. Sánchez, R.A. Mollineda, On the effectiveness of preprocessing methods when dealing with different levels of class imbalance, Knowledge-Based Systems 25 (2012) 13–21.
[25] J. Neves, A. Vieira, Improving bankruptcy prediction with Hidden Layer Learning Vector Quantization, European Accounting Review 15 (2006) 253–271.
[26] E. Alfaro-Cid, K. Sharman, A. Esparcia-Alcázar, A genetic programming approach for bankruptcy prediction using a highly unbalanced database, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 4448 (2007) 169–178.
[27] R.A. Mathiasi Horta, B.S.L. Pires De Lima, C.C.H. Borges, A semi-deterministic ensemble strategy for imbalanced datasets (SDEID) applied to bankruptcy prediction, WIT Transactions on Information and Communication Technologies 40 (2008) 205–213.
[28] S.J. Yen, Y.S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Systems with Applications 36 (2009) 5718–5727.
[29] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (2009) 1263–1284.

[30] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, Handling imbalanced datasets: a review, GESTS International Transactions on Computer Science and Engineering 30 (2006).
[31] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, Computational Intelligence 20 (2004) 18–36.
[32] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.
[33] S. Kotsiantis, P. Pintelas, Mixture of expert agents for handling imbalanced data sets, Annals of Mathematics, Computing & TeleInformatics 1 (2003) 46–55.
[34] MathWorks, Statistics Toolbox User's Guide, The MathWorks, Massachusetts, 2012.
[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explorations 11 (2009).
[36] T. Fawcett, Roc Graphs: Notes and Practical Considerations for Data Mining Researchers, HP Laboratories, Palo Alto, CA, USA, January 2003.
[37] C.-F. Tsai, Feature selection in bankruptcy prediction, Knowledge-Based Systems 22 (2009) 120–127.
[38] J. Pacheco, S. Casado, L. Núñez, Use of VNS and TS in classification: variable selection and determination of the linear discrimination function coefficients, IMA Journal of Management Mathematics 18 (2007) 191–206.
[39] J. Pacheco, S. Casado, L. Núñez, A variable selection method based in tabu search for logistic regression models, European Journal of Operational Research 199 (2009) 506–511.
[40] A. Unler, A. Murat, A discrete particle swarm optimization method for feature selection in binary classification problems, European Journal of Operational Research 206 (2010) 528–539.
[41] J.A.K. Suykens, T.V. Gestel, J.D. Brabanter, B.D. Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002.
[42] J. Demšar, Statistical comparisons of classifiers over multiple data set, Journal of Machine Learning Research (2006) 1–30.