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

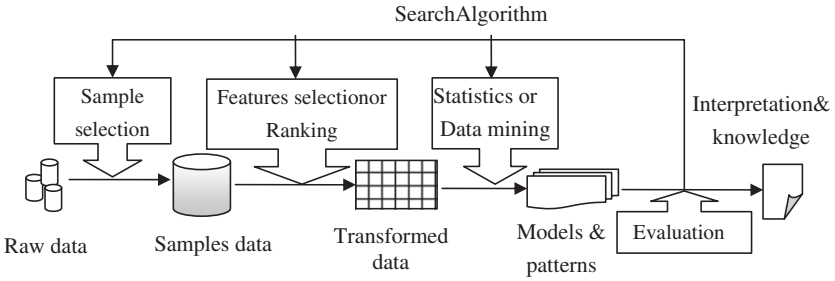

Fig. 1. Development processes of empirical bankruptcy prediction models.

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.