Expert Systems with Applications 41 (2014) 2353–2361

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An improved boosting based on feature selection for corporate bankruptcy prediction

Gang Wang [a,b,c,*], Jian Ma [c], Shanlin Yang [a,b]

[a] School of Management, Hefei University of Technology, Hefei, Anhui 230009, PR China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei, Anhui, PR China
[c] Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Corporate bankruptcy prediction
Ensemble learning
Boosting
Feature selection

ABSTRACT

With the recent financial crisis and European debt crisis, corporate bankruptcy prediction has become an increasingly important issue for financial institutions. Many statistical and intelligent methods have been proposed, however, there is no overall best method has been used in predicting corporate bankruptcy. Recent studies suggest ensemble learning methods may have potential applicability in corporate bankruptcy prediction. In this paper, a new and improved Boosting, FS-Boosting, is proposed to predict corporate bankruptcy. Through injecting feature selection strategy into Boosting, FS-Booting can get better performance as base learners in FS-Boosting could get more accuracy and diversity. For the testing and illustration purposes, two real world bankruptcy datasets were selected to demonstrate the effectiveness and feasibility of FS-Boosting. Experimental results reveal that FS-Boosting could be used as an alternative method for the corporate bankruptcy prediction.

1. Introduction

Predicting corporation bankruptcy is an important management science problem and its main goal is to differentiate those corporations with a probability of distress from healthy corporations. Moreover, incorrect decision-making in financial institutions may run into financial difficulty or distress and cause many social costs affecting owners or shareholders, managers, government, etc. As a result, how to predict corporate bankruptcy has become a hot topic for both industrial application and academic research (Li, Andina, & Sun, 2012; Olson, Delen, & Meng, 2012; Zhou, Lai, & Yen, 2014).

As there are no mature theories of corporate bankruptcy, studies in corporate bankruptcy have largely been based on trial and error iterative processes of selecting features and predictive models (Li & Sun, 2009; Zhou, Lai, & Yen, 2014). With the development of statistics, artificial intelligence (AI), some statistical methods and intelligent methods have been proposed for corporate bankruptcy prediction. The statistical methods applied in corporate bankruptcy prediction mainly include Linear Discriminant Analysis (LDA), Multivariate Discriminant Analysis (MDA), Quadratic Discriminant Analysis (QDA), Logistic Regression Analysis (LRA), and Factor Analysis (FA) (Li & Sun, 2009; Zmijewski, 1984). However, the problem with applying these statistical techniques to corporate bankruptcy prediction is that some assumptions, such as the multivariate normality assumptions for independent variables, are frequently violated in the practice, which makes these techniques theoretically invalid for finite samples (Shin & Lee, 2002). In recent years, many studies have demonstrated that intelligent techniques such as Artificial Neural Networks (ANN), Decision Tree (DT), Case-Based Reasoning (CBR), Support Vector Machine (SVM) can be used as alternative methods for corporate bankruptcy prediction (Olson et al., 2012; Tsai & Wu, 2008). In contrast with statistical techniques, intelligent techniques do not assume certain data distributions and automatically extract knowledge from training samples (Wang, Ma, Huang, & Xu, 2012).

However, there is still no overall best intelligent methods used in predicting corporate bankruptcy. The performance of prediction depends on the details of the problem, the data structure, the used characteristics, the extent to which it is possible to segregate the classes by using those characteristics, and the objective of the classification (Duéñez-Guzmán & Vose, 2013). Recently, integrating multiple predictors into an aggregated output, i.e. ensemble methods, has been demonstrated to be an efficient strategy for achieving high prediction performance, especially when the component predictors have different structures that lead to independent prediction errors (Breiman, 1996; Polikar, 2006). Moreover, latest studies have shown that such ensemble techniques have performed better than single intelligent technique in financial distress prediction (Deligianni & Kotsiantis, 2012; Sun & Li, 2012). In this paper, a novel and improved Boosting, FS-Boosting, is proposed to predict corporate bankruptcy. Through injecting feature

* Corresponding author at: Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong. Tel.: +852 9799 0955; fax: +852 2788 8694.
E-mail address: wgedison@gmail.com (G. Wang).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.033

selection strategy into Boosting, FS-Booting can get better performance as base learners in FS-Boosting could get more accuracy and diversity. For the testing and illustration purposes, two real world bankruptcy datasets were selected to demonstrate the effectiveness and feasibility of FS-Boosting. Among eight methods, FS-Boosting gets the best prediction accuracy on two datasets. Experimental results reveal that FS-Boosting can be used as an alternative method for the corporate bankruptcy prediction.

The remainder of the paper is organized as follows. In Section 2, we review the related work of corporate bankruptcy prediction. In Section 3, an improved boosting, FS-Boosting, is proposed for corporate bankruptcy prediction. In Section 4, we present the details of experiment design. Sections 5 and 6 summarize and analyze empirical results and discussion. Based on the results and observations of these experiments, Section 7 draws conclusions and future research directions.

2. Related work

Many techniques have been proposed by prior research. In this study, we classified these techniques into statistical techniques and intelligent techniques.

2.1. Statistical techniques for corporate bankruptcy prediction

In the past, many researchers have developed a variety of statistical techniques for corporate bankruptcy. The main statistical methods include LDA, MDA, QDA, LRA, and FA. One of the earliest techniques of corporate bankruptcy prediction were proposed by Beaver (1966) and Altman (1968). They used single discriminant analysis and multiple discriminant analysis, respectively, to identify corporate that would go bankrupt. Subsequently, due to the restrictive statistical requirement of normality for the explanatory variables and quality for the variance-covariance group matrices, logit and probit models were also applied (Ohlson, 1980; Zmijewski, 1984). West used the factor analysis to create composite variables to describe bank's financial and operating characteristics (West, 1985). Experimental results demonstrated that the combined method of factor analysis and logit was promising in evaluating bank's condition.

However, these conventional statistical techniques have some restrictive assumptions, such as the normality and independence among predictor or input variables. Considering that the violation of these assumptions for independent variables frequently occurs with financial data, the statistical techniques can have limitations to obtain the effectiveness and validity (Shin & Lee, 2002).

2.2. Intelligent techniques for corporate bankruptcy prediction

In recent years, many studies have demonstrated that intelligent techniques can be alternative methodologies to predict corporate bankruptcy. Intelligent techniques automatically extract knowledge from a dataset and construct different model representations to explain the data set. The major difference between intelligent techniques and statistical techniques is that statistical techniques usually need researchers to impose structures to different models, such as the linearity in the multiple regression analysis, and to construct the model by estimating parameters to fit the data or observation, while intelligent techniques allow learning the particular structure of the model from the data (Wang, Hao, Ma, & Jiang, 2011).

The intelligent techniques frequently used include DT (Shaw & Gentry, 1990), ANN (Tam & Kiang, 1992; Tang & Chi, 2005), CBR (Buta, 1994; Shin & Han, 2001), and SVM (Min & Lee, 2005; Van Gestel et al., 2003). Shaw and Gentry applied DT to risk classification applications and found that the performance of DT was better than probit or logit analysis (Shaw & Gentry, 1990). Tam and Kiang used ANN to predict bankruptcy risk and compared ANN with a linear discriminate model, a logit model, DT and KNN (Tam & Kiang, 1992). Tang and Chi proposed a means to collect and determine explanatory variables using a between-countries approach (Tang & Chi, 2005). In addition, they established a systematic experiment to investigate the influences of techniques for both network architecture selection and variable selection on neural network models' learning and prediction capability. CBR, which benefits from utilizing case specific knowledge of previous experienced problem situations, is also applied to predict corporate bankruptcy. A new problem is solved by finding a similar past case and reusing it in the new problem domain. Buta developed a CBR model using financial data of 1000 companies in the Standard and Poor's Compustat database (Buta, 1994). And Shin and Han proposed a CBR method which used nearest neighbor matching algorithms to retrieve cases (Shin & Han, 2001). Another widely used method is SVM whose formulation simultaneously embodies the structural risk and empirical risk minimization principles. Van Gestel et al. reported least squares SVM got significantly better results when contrasted with the classical methods (Van Gestel et al., 2003). Min and Lee proposed grid-search method using five-fold cross validation to find out the optimal parameter values of kernel function of SVM (Min & Lee, 2005) and found that SVM outperformed NN, MDA and logit models.

However, there is still no overall best intelligent techniques used in predicting corporate bankruptcy. Recently, latest studies have shown that ensemble techniques have performed better than single intelligent technique for corporate bankruptcy prediction (Alfaro, García, Gámez, & Elizondo, 2008; Deligianni & Kotsiantis, 2012; Sun & Li, 2012; Sánchez-Lasheras, de Andrés, Lorca, & de Cos Juez, 2012). For example, Alfaro et al. compared two classification methods, i.e., AdaBoost and ANN, and experimental results showed the improvement in accuracy that AdaBoost achieves against the ANN (Alfaro et al., 2008). Sun and Li proposed a new SVM ensemble method whose candidate single classifiers are trained by SVM algorithms with different kernel functions on different feature subsets of one initial dataset (Sun & Li, 2012). Deligianni and Kotsiantis found that an ensemble of classifiers could enable users to predict bankruptcies with satisfying precision long before the final bankruptcy (Deligianni & Kotsiantis, 2012). At the same time, some studies also found that ensemble methods are not always clearly superior to single classifiers (Alfaro-Cid et al., 2008; Nanni & Lumini, 2009; Tsai & Wu, 2008). It means that ensemble techniques should be adjusted according to the character of corporate bankruptcy prediction. In this paper, an improved Boosting based on feature selection is proposed to predict corporate bankruptcy.

3. Feature selection based Boosting for bankruptcy prediction

3.1. Feature selection

Feature selection has been an active research area in machine learning and data mining communities (Liu & Motoda, 1998). The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection reduces the dimensionality of feature space, and removes redundant, irrelevant, or noisy data. It brings the immediate effects for application: speeding up an algorithm, improving the data quality and thereof the performance of classifier (Blum & Langley, 1997).

Diverse feature selection techniques have been proposed in the machine learning and data mining literature. Feature selection