Expert Systems with Applications 38 (2011) 11261–11272

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Predicting corporate financial distress based on integration of decision tree classification and logistic regression

Mu-Yen Chen

Department of Information Management, National Taichung Institute of Technology, Taichung 404, Taiwan, ROC

ARTICLE INFO

Keywords:
Financial distress
Artificial intelligent
Decision tree classification
Logistic regression

ABSTRACT

Lately, stock and derivative securities markets continuously and rapidly evolve in the world. As quick market developments, enterprise operating status will be disclosed periodically on financial statement. Unfortunately, if executives of firms intentionally dress financial statements up, it will not be observed any financial distress possibility in the short or long run. Recently, there were occurred many financial crises in the international marketing, such as Enron, Kmart, Global Crossing, WorldCom and Lehman Brothers events. How these financial events affect world's business, especially for the financial service industry or investors has been public's concern. To improve the accuracy of the financial distress prediction model, this paper referred to the operating rules of the Taiwan Stock Exchange Corporation (TSEC) and collected 100 listed companies as the initial samples. Moreover, the empirical experiment with a total of 37 ratios which composed of financial and other non-financial ratios and used principle component analysis (PCA) to extract suitable variables. The decision tree (DT) classification methods (C5.0, CART, and CHAID) and logistic regression (LR) techniques were used to implement the financial distress prediction model. Finally, the experiments acquired a satisfying result, which testifies for the possibility and validity of our proposed methods for the financial distress prediction of listed companies.

This paper makes four critical contributions: (1) the more PCA we used, the less accuracy we obtained by the DT classification approach. However, the LR approach has no significant impact with PCA; (2) the closer we get to the actual occurrence of financial distress, the higher the accuracy we obtain in DT classification approach, with an 97.01% correct percentage for 2 seasons prior to the occurrence of financial distress; (3) our empirical results show that PCA increases the error of classifying companies that are in a financial crisis as normal companies; and (4) the DT classification approach obtains better prediction accuracy than the LR approach in short run (less one year). On the contrary, the LR approach gets better prediction accuracy in long run (above one and half year). Therefore, this paper proposes that the artificial intelligent (AI) approach could be a more suitable methodology than traditional statistics for predicting the potential financial distress of a company in short run.

1. Introduction

Recently, one of the most attractive business news is a series of financial crisis events related to the public companies. Some of these companies are famous and also at high stock prices, originally (e.g. Enron Corp., Kmart Corp., WorldCom Corp., Lehman Brothers Bank, etc.). In consequence of the financial crisis, it is always too late for many creditors to withdraw their loans, as well as for investors to sell their own stocks, futures, or options. Therefore, corporate bankruptcy is a very important economic phenomenon and also affects the economy of every country. In Taiwan, domestic and foreign capital markets have developed rapidly in recent years, gradually giving people the idea of making a financial investment. Nevertheless, Procomp Corp. and Cdbank Corp. bankruptcy events have also caused tremendous disorder in the financial market and related industries are also affected by these economic shocks in Taiwan. The number of bankruptcy firms is important for the economy of a country and it can be viewed as an indictor of the development and robustness of the economy (Zopounidis & Dimitras, 1998). The high individual, economic, and social costs encountered in corporate failures or bankruptcies have spurred searches for better understanding and prediction capability (McKee & Lensberg, 2002). Therefore, forecasting corporate financial distress plays an increasingly important role in today's society since it has a significant impact on lending decisions and the profitability of financial institutions.

A common methodology to bankruptcy prediction is to summarize the literature to search a large set of potential predictive financial and/or non-financial variables and then reduce a set of not significant variables, through traditional mathematical analysis

E-mail address: mychen@ntit.edu.tw

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.173

that will predict bankruptcy (Lensberg, Eilifsen, & McKee, 2006). Many traditional classification techniques have been presented to predict financial distress using ratios, e.g., univariate approaches (Beaver, 1966), multivariate approaches, linear multiple discriminant approaches (MDA) (Altman, 1968; Altman, Edward, Haldeman, & Narayanan, 1977), multiple regression (Meyer & Pifer, 1970), logistic regression (Dimitras, Zanakis, & Zopounidis, 1996), factor analysis (Blum, 1974), and stepwise (Laitinen & Laitinen, 2000). However strict assumptions of traditional statistics such as linearity, normality, independence among predictor variables and pre-existing functional form relating to the criterion variable and the predictor variable limit their application in the real world (Hua, Wang, Xu, Zhang, & Liang, 2007).

Therefore, this paper proposes a model of financial distress prediction comparing decision tree (DT) classification and logistic regression (LR) techniques. The main objectives of this paper are to (1) adopt DT and LR techniques to construct a financial distress prediction model, (2) use financial and non-financial ratios to enhance the accuracy of the financial distress prediction model, (3) employ a traditional statistical method (principle component analysis, PCA) to compare the degree of accuracy with that of the artificial intelligent (AI) approach, and (4) to expand this model so that it will work within a financial distress prediction system to provide information to investors as well as investment monitoring organizations. The data for our experiment were collected from the Taiwan Stock Exchange Corporation (TSEC) database.

The rest of this paper is organized as follows. A literature review of related techniques is provided in Section 2. We describe our proposed approach and its capabilities of each step in Section 3. Section 4 presents the process for choosing appropriate variables by PCA. In Section 5, we analyzed the prediction performance of our approach and fulfilled several experiments. Moreover, we compared our results with the DT, and LR approaches in Section 6. Finally, we inference our conclusions and discuss future research in Section 7.

2. Classification techniques

2.1. Decision trees algorithm

Data mining (DM), also known as "knowledge discovery in databases" (KDD), is the process of discovering meaningful patterns in huge databases (Han & Kamber, 2001). In addition, it is also an application that can provide significant competitive advantages for making the right decision (Huang, Chen, & Lee, 2007). The more common model functions in the current data mining process include the classification, regression, clustering, association rules, summarization, dependency modeling and sequence analysis (Mitra, Pal, & Mitra, 2002). Decision tree is one of common DM methodologies that provide both classification and predictive functions simultaneously. Focusing on the data provided, it produces a model of tree-shaped structure using inductive reasoning (Chang & Chen, 2009). Many scholars also use artificial neural network (ANN) techniques to solve classification and prediction problems. However, there are three weaknesses of neural networks mostly. Firstly, neural networks are not guaranteed to converge to a global optimal solution. Secondly, neural networks have the well-known 'over-training problem.' Last, neural networks have the 'black-box' phenomenon that lacks the ability to explain their behavior (Roiger & Geatz, 2003). The first two problems have been solved by setting the number of hidden nodes and learning parameters. However, it is difficult to explain how neural networks act, and how they make the decisions through their layers. Decision trees are a well-known technique and have had many successful applications to real-world problems (Kumar & Ravi, 2007). In addition, decision trees have the ability to build models with datasets including numerical and categorical data (Witten & Frank, 2005).

Major algorithms of decision tree analysis model include ID3 (Interactive Dichotomiser 3), C5.0, classification and regression trees (CRAT), and chi-squared automatic interactive detector (CHAID) models. In the late 1970s, Ross (1993) proposed an algorithm named ID3 to generate decision trees. Based on the theory of information gain, ID3 algorithm chooses the optimal information gain to as a first attribute for branching of decision trees and thus constructs a simple trees structure. However, ID3 algorithm still has its shortcoming when using information gain as a rule to select attributes for segmentation will result in bias over attributes of higher values. Therefore, in the condition when only one data remains in the sub-tree after data set segmentation, its information gain is the highest, indicating a less meaningful segmentation (Ross, 1993).

The C4.5 algorithm improves ID3 with regard to the splitting rule and the calculation method (Quinlan, 1993). It uses gain-ratio index instead as a measurement method to segment attributes and thus can reduce the influence of ID3 drawback that segmentation nodes prefer too many sub-trees. C5.0 algorithm is a commercial version of C4.5, such as Clementine and RuleQuest (Quinlan, 1997), and it improves the rule generation of C4.5. It can be skilled in processing enormous datasets particularly. Besides, C5.0 algorithm is also faster in speed and more memory efficient than C4.5 due to Boosting method adopting.

In CART (Breiman, Friedman, Olshen, & Stone, 1984), the building of the tree classifier is also accomplished by recursively splitting the instance space in smaller subparts. CART algorithm generates a binary decision tree, unlike ID3 which only creates two children. Both of CART and CHAID provide a set of rules that can be used to an unclassified dataset to predict which records will have a given result. CART segments a dataset by creating two-way splits, but CHAID segments a dataset by chi-square test to create multi-way splits. Generally, CART needs less data preparation than CHAID.

CHAID algorithm is based on the chi-square test and constructed by repeatedly splitting subsets of the space into two or more child nodes with the entire datasets (Michael & Gordon, 1997). To decide the best split at any node, any allowable pair of categories of the predictor variables is merged until there is no statistically significant difference within the pair with respect to the target variable. This CHIAD algorithm surely handles interactions between the independent variables that are directly available from an examination of the tree. Although there is no optimal method to obtain the best segment size, in fact, CHAID can assist researchers to compromise variances against segment size to discovery the most adequate one. Certainly, CHAID algorithm clearly proves which segmentation variable must come first for the large datasets.

2.2. Statistical algorithms

Most of the broadly used traditional statistical algorithm applied for prediction and diagnosis in many disciplines are discriminant analysis, Logistic regression, Bayesian approach, and multiple regression. These models have been proven to be very effective, however, for solving relatively less complex problems. LR is a regression method for predicting a dichotomous dependent variable. In producing the LR equation, the maximum-likelihood ratio was used to determine the statistical significance of the variables (Hosmer & Lemeshow, 2000). In logistic regression models, dependent variable is always in categorical form and has two or more levels. Independent variables may be in numerical or categorical form (Camdeviren, Yazici, Akkus, Bugdayci, & Sungur, 2007).We consider the situation where we observe a binary outcome variable y and a vector $x = (1, x_1, x_2, \ldots, x_k)$ of covariates for each of N