Knowledge-Based Systems 85 (2015) 52–61

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

The performance of corporate financial distress prediction models with features selection guided by domain knowledge and data mining approaches

Ligang Zhou [a,*], Dong Lu [b], Hamido Fujita [c]

[a] School of Business, Macau University of Science and Technology, Taipa, Macau
[b] School of Business, SiChuan Normal University, SiChuan Province, PR China
[c] Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan

ARTICLE INFO

Article history:
Received 18 April 2014
Received in revised form 17 March 2015
Accepted 20 April 2015
Available online 27 April 2015

Keywords:
Financial distress prediction
Features selection
Domain knowledge
Data mining

ABSTRACT

Experts in finance and accounting select feature subset for corporate financial distress prediction according to their professional understanding of the characteristics of the features, while researchers in data mining often believe that data alone can tell everything and they use various mining techniques to search the feature subset without considering the financial and accounting meanings of the features. This paper investigates the performance of different financial distress prediction models with features selection approaches based on domain knowledge or data mining techniques. The empirical results show that there is no significant difference between the best classification performance of models with features selection guided by data mining techniques and that by domain knowledge. However, the combination of domain knowledge and genetic algorithm based features selection method can outperform unique domain knowledge and unique data mining based features selection method on AUC performance.

1. Introduction

Corporate financial distress prediction (CFDP) is very important for investors, credit lenders and company's partners, such as suppliers or retailers. The investors and credit lenders need to evaluate the financial distress risk of a company before they make any investment or credit granting decisions on the company in order to avoid suffering a great loss. A company's suppliers or retailers always conduct credit transaction with the company and they also need to fully understand the company's financial status and make decisions on the credit transaction.

To correctly predict a company's financial distress is a great concern for many stake holders of a company. This practical significance has driven a lot of studies on the issue of corporate financial distress prediction. Most of these studies often focused on introducing or improving the quantitative approaches from statistics and data mining discipline to develop corporate financial distress prediction models (CFDPM) with the objective of increasing the prediction accuracy. The preliminary study of CFDPM with a multivariate framework proposed by Altman [1] was based on the discriminant analysis approach. Thereafter, many other complex statistical and data mining methods were introduced to develop the CFDPM, such as neural networks [2,3], decision trees [4], and support vector machines [5]. In addition, the fuzzy theory can also be used for developing CFDPM [6,7]. Most recent research mainly focuses on the development of hybrid models with the combination of two or more than two methods [8–10]. Although the empirical results in these studies often showed that hybrid models could outperform the single models, the computation always consumes more time and the theory or reason for the combinations is not always known and explained, which prevent their wide applications in practice to some degree.

The problem of corporate financial distress prediction is to take advantage of all currently available information related to the company to predict if it will fall into the condition of default or financial difficulty. Consequently, the performance of the CFDPM is determined not only by the model or methods that is used for the prediction but also by the selection of available information. In practice, some credit rating agencies just use their experiences and judgments to select the relevant information to evaluate the credit risk of a particular company or individual with a simple scorecard instead of complex statistical models [11]. However, the information related to a company is huge, including macroeconomic situations, company characteristics, financial status and market information, and most studies have demonstrated that financial and marketing information is the most effective in

* Corresponding author.
E-mail addresses: mrlgzhou@gmail.com (L. Zhou), dlu@sicnu.edu.cn (D. Lu), issam@iwate-pu.ac.jp (H. Fujita).

http://dx.doi.org/10.1016/j.knosys.2015.04.017
0950-7051/

financial distress prediction. What financial and marketing information should be considered in the development of corporate financial distress prediction models?

There are often two research streams in the feature subset selection for corporate financial distress prediction models. One is based on the domain knowledge from financial and accounting theory. The main characteristic of the features selected by domain knowledge is that the effect of the features on the financial distress can be evaluated to some degree in terms of financial and accounting theory. Altman [1] investigated a set of twenty-two financial and economic ratios in the prediction of corporate bankruptcy and found that the subset of the following variables is useful for financial distress prediction: working capital/total assets, retained earnings/total assets, earnings before interest and taxes/total assets, market value equity/book value of total debt. Altman et al. [12] observed the distinct difference in the accounting procedures and the quality of financial documents between the firms in China and those in the western world, and considered variables that were widely accepted in China and deemed contributive in previous studies. They investigated fifteen variables that reflect various aspects of a company, such as profitability, liquidity and solvency, and asset management efficiency and capital structure and financial leverage. After considering a large number of combinations of the 15 characteristic variables, they found that the following feature subset yielded the best performance: total liabilities/total assets, net profit/average total assets, working capital/total assets, and retained earnings/total assets. Shumway [13] developed a simple hazard model and compared the performance of Altman's variables [1] and Zmijiewski's variables [14] and a new set of variables including accounting and three market-driven variables. The empirical result shows that the new accounting and market-driven variables set outperforms other two alternative models in out-of-sample forecasts. The accounting and market-driven feature subset includes: net income/total asset, total liabilities/total asset, relative size (market capitalization/total size of the corresponding market), the firm's past excess returns and the idiosyncratic standard deviation of the firm's stock returns. Ravi and Ravi [15] reviewed 128 papers in bankruptcy prediction and listed more than 500 different variables used by these different papers. Almost all of these 128 papers used different subsets of features. It is perhaps natural that different experts have different opinions in determining what information should be considered in the prediction of financial distress of a company.

Another stream in feature subset selection is based on data mining techniques. Adherents to the data mining stream view believe that data will tell everything, and the approach uses some features selection methods in data mining to identify which feature subset can improve the prediction performance without considering the financial and accounting meanings of the features. Tsai [16] compared five well-known features selection methods used in bankruptcy prediction and used multi-layer perceptron neural networks to construct the prediction model, and found the *t*-test features selection method performs better than others. du Jardin [17] introduced a neural network based model using a set of variables selected by a criterion being adapted to the network for the bankruptcy prediction problem. Drezner et al. [18] reported that a tabu search based variables selection model can increase the predictability of corporate bankruptcy by up to 10 percentage points in comparison to Altman's *Z*-Score [1] model. Although most researchers in this stream like Cho, Mays, et al. [10,19] noticed that there were hundreds of financial variables and the model performance was affected by input variables selection, they only investigated a very small subset of variables guided by previous studies in the data set for empirical study without taking good advantage of the original data set from which the sample for training and testing model was retrieved. Few previous studies in financial distress prediction compare the performance of features selection with domain knowledge and data mining, together with investigating the difference of feature subset found by domain knowledge and data mining [2–4,8–10].

The contribution of this study is twofold. First, it compares the performance of domain knowledge and data mining based features selection methods in financial distress prediction on a data set with more than three hundred variables. The experimental result shows that the features selected by data mining methods can perform as well as those selected by domain knowledge of experts in finance or accounting. Second, it considers the combination of domain knowledge and data mining features selected approach in order to take good advantage of the experts' professional knowledge and the powerful mining capability of data mining techniques. The experimental result shows that the performance of the combined method can outperform unique domain knowledge and unique features selection method.

The outline of this paper is as follows. Section 2 introduces the important domain knowledge and data mining feature subset selection methods for financial distress prediction. Section 3 reports the empirical results and Section 4 gives the conclusion.

2. Domain knowledge vs. data mining in features selection

2.1. Features selection by domain knowledge

Financial ratio analysis is an important way to analyze financial statements. There are often hundreds of financial ratios measuring different aspects of a company, such as liquidity, long-term solvency, asset management, profitability, and market value. The meaning and usage of the financial variables has been widely discussed in finance [20,21]. It is impossible to investigate all financial ratios suggested for CFDPM by the researchers from finance and accounting. Only the ratios that are widely accepted and have been verified with great performance and have been taken as a benchmark in most previous research are considered. Therefore, a classical group of features selected from domain knowledge is based on the work from Altman [1], Altman [12] and Shumway [13]. The feature subset employed by Altman [1], Altman [12] and Shumway is denoted as FA1, FA2, and FS respectively. The union of these three feature subsets is denoted by FAAS. The detail of the ten features in FAAS is briefly described as follows.

1. Working capital to total assets (WCTA) measures the firm's liquidity or short-term solvency. High WCTA shows that the firm can match its account payable obligation on time and a low WCTA indicates that the firm may be unable to pay its suppliers and creditors.
2. Retained earnings to total assets (RETA) reflects a firm's strategy on its net earnings. If a firm needs more funds for the increase of business and it prefers to raise funds from inside, the firm would like to keep a higher RETA.
3. Earnings before interest and taxes to total assets (EBTITA) is an important measures of a firm's profitability. Higher EBITTA indicates higher profitability of a firm.
4. Sales to total assets (STA) is also a measures of a firm's profitability. A low ratio indicates that the total assets of the firm cannot provide adequate revenue.
5. Net income to total assets (NITA) is also known as return on assets (ROA). It indicates how efficient a firm's management is at using its assets to generate earnings. It is another important measure of a firm's profitability.
6. Total liabilities to total assets (TLTA) measures a firm's long-term solvency. It indicates a firm's financial risk by determining what ratio of company's assets is financed by debt. Higher TLTA means higher financial risk.