Bankruptcy Prediction Using Data Mining Tools

Wikil Kwak
University of Nebraska at Omaha, USA

INTRODUCTION

Data mining is a widely used business tool such as credit rating, consumer buying pattern prediction, and firm bankruptcy prediction studies. The purpose of this chapter is to summarize a multiple criteria linear programming (MCLP) data mining approach as well as other current data mining tools for firm bankruptcy prediction. The results of applications of data mining tools in a bankruptcy prediction study are promising as this approach performs better than traditional multiple discriminant analysis or logit analysis using financial data. There are trade-offs for applications of data mining for bankruptcy prediction studies with higher prediction rates. However, overall prediction rates of bankruptcy studies are test period sensitive from results of previous empirical studies (for example, Begley, et. al., 1996). Our empirical results show the prediction rates of U.S. data studies are similar to those of Japanese or Korean data studies. But traditional logit analysis can still add incremental information for financial characteristics of a firm.

Data mining can be achieved by association, classification, clustering, and prediction. From the mathematical point of view, the algorithms of data mining can be implemented by binary or induction decision tree, statistics, and neural networks. Our model uses one of the prediction mechanisms in data mining. Our bankruptcy prediction studies are to predict future outcomes based on past financial and other data. Firm bankruptcy prediction is an interesting topic to stakeholders such as investors, bankers, managers, employees, and the general public.

BACKGROUND

Altman (1968) used multiple discriminant analysis (MDA) by using financial ratios to predict a firm bankruptcy model. Altman et al. (1977) later proposed the ZETA model, but their assumptions of data being normally distributed can be a problem when applying this model.

Ohlson (1980) used a logit model without prior assumptions about the probability of bankruptcy or the distribution of predictor variables. However, hold-out sample tests, in general, are potentially upwardly biased (Grice and Dugan, 2001) and the differences in the macro economic factors are sensitive to specific time periods. For example, Grice and Ingram (2001) empirically tested Altman's model using the 1988-1991 test period and reported that the overall correct classification rate dropped to 57.8%. Therefore, the test period may be sensitive to overall prediction rates.

Freed and Glover (1986) proposed linear programming (LP) to minimize misclassifications in linear discriminant analysis. Gupta et al. (1990) also proposed linear goal programming as an alternative to discriminant analysis. However, these approaches may not be manageable for the large-scale databases at that time. Pompe and Feelders (1997) compared machine learning, neural networks, and statistics using experiments to predict firm bankruptcy. However, their results are not conclusive to which methods outperform the other methods.

Shin and Lee (2002) proposed a genetic algorithm (GA) in bankruptcy prediction modeling using Korean manufacturing companies from 1995

DOI: 10.4018/978-1-4666-5202-6.ch021

to 1997. Their approach is capable of extracting rules that are easy to understand for users like expert systems, but by using only the manufacturing industry, there may be an upward bias for the prediction accuracy rate of 80.8%. Recent bankruptcy studies classified and predicted the final bankruptcy resolution using longitudinal study (for example, Barniv et al., 2002), but our focus should be prediction accuracy, flexibility of the model, and easy to apply using real world data without difficulty of computational complexity. The most prominent bankruptcy studies in accounting and finance are Altman's and Ohlson's and our study uses their financial variables. These variables are pulled from each bankrupt firm's financial statements the year before they filed bankruptcy. If we want to improve the prediction accuracy, we could easily add a cash flows variable or non-financial variable such as stock market returns, missing dividend, and auditor changes.

Recently, other intelligent techniques such as neural network or support vector machines were used in firm bankruptcy studies (Kumar and Ravi, 2007). Min and Lee (2005) proposed a support vector machine for the firm bankruptcy prediction study and their holdout data classification rate was 83.06%. They concluded that their model is better than multiple discriminant analysis and logit models. MCLP data mining approach uses multiple measures to separate each data from different classes, but the support vector machine searches the minority of the data (support vectors) to represent the majority in classifying the data (Shi et al., 2005).

DATA MINING TOOLS

Multiple Criteria Linear Programming Data Mining Approach

The multiple criteria linear programming (MCLP) data mining approach (see Kou, Liu, Peng, Shi, Wise, & Xu, 2003) is our primary model of interest in Kwak et al. (2005) and Kwak et al. (2006) to predict firm bankruptcy. Following is the summary of the basic concept of our MCLP data mining model.

Given a set of k variables about the event or state to be predicted or classified (such as bankruptcy, auditor change, or internal control weakness), in database $a = (a_1, \ldots, a_k)$, let $A_i = (A_{i1}, \ldots, A_{ik}) \in R^k$ be the sample observations of data for the variables, where $i = 1, \ldots, n$ and n is the sample size. We want to determine the coefficients of the variables, denoted by $X = (x_1, \ldots, x_k)^T$, and a boundary value of b to separate two classes: B (such as Bankruptcy in Kwak et al. (2005) and in Kwak et al. (2006)) and NB (such as Non-bankruptcy in Kwak et al. (2005) and in Kwak et al. (2006)) or AC (such as Auditor change in Kwak et al. (2011) and internal control weakness in Kwak et al. (2009)) and NAC (such as Non-auditor change in Kwak et al. (2011) and non-internal control weakness in Kwak et al. (2009)); that is,

$$\boldsymbol{A}_i\boldsymbol{X} \leq b, \boldsymbol{A}_i \in \text{B or AC and } \boldsymbol{A}_i\boldsymbol{X} > b, \boldsymbol{A}_i \in \text{NB or NAC.} \tag{1}$$

Consider now two kinds of measurements for better separation of B or AC and NB or NAC firms. Let α_i be the overlapping degree with respect to $\boldsymbol{A}_i$, and β_i be the distance from $\boldsymbol{A}_i$ to its adjusted boundary. In addition, we define α to be the maximum overlapping of two-class boundary for all cases $\boldsymbol{A}_i$ ($\alpha_i < \alpha$) and β to be the minimum distance for all cases $\boldsymbol{A}_i$ from its adjusted boundary ($\beta_i > \beta$). Our goal is to minimize the sum of α_i and maximize the sum of β_i simultaneously. By adding α_i into (1) above, we have:

$$\boldsymbol{A}_i\boldsymbol{X} \leq b + \alpha_i, \boldsymbol{A}_i \in \text{B or AC and } \boldsymbol{A}_i\boldsymbol{X} > b - \alpha_i, \boldsymbol{A}_i \in \text{NB or NAC.} \tag{2}$$

However, by considering β_i, we can rewrite (2) as

$$\boldsymbol{A}_i\boldsymbol{X} = b + \alpha_i - \beta_i, \boldsymbol{A}_i \in \text{B or AC and } \boldsymbol{A}_i\boldsymbol{X} = b - \alpha_i + \beta_i, \boldsymbol{A}_i \in \text{NB or NAC.} \tag{3}$$

Related Content