Expert Systems with Applications 37 (2010) 1535–1541

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A hybrid approach of DEA, rough set and support vector machines for business failure prediction

Ching-Chiang Yeh [a], Der-Jang Chi [b,*], Ming-Fu Hsu [c]

[a] Department of Business Administration, National Taipei College of Business, Taipei, Taiwan, ROC
[b] Department of Accounting, Chinese Culture University, Taipei 11114, Taiwan, ROC
[c] Department of International Business Studies, National Chi Nan University, Nantou County, Taiwan, ROC

ARTICLE INFO

Keywords:
Business failure
Financial ratios
DEA
Rough set
Support vector machines

ABSTRACT

The prediction of business failure is an important and challenging issue that has served as the impetus for many academic studies over the past three decades. While the efficiency of a corporation's management is generally acknowledged to be a key contributor to corporation's bankrupt, it is usually excluded from early prediction models. The objective of the study is to use efficiency as predictive variables with a proposed novel model to integrate rough set theory (RST) with support vector machines (SVM) technique to increase the accuracy of the prediction of business failure. In the proposed method (RST–SVM), data envelopment analysis (DEA) is employed as a tool to evaluate the input/output efficiency. Furthermore, by RST approach, the redundant attributes in multi-attribute information table can be reduced, which showed that the number of independent variables was reduced with no information loss, is utilized as a preprocessor to improve business failure prediction capability by SVM. The effectiveness of the methodology was verified by experiments comparing back-propagation neural networks (BPN) approach with the hybrid approach (RST–BPN). The results shows that DEA do provide valuable information in business failure predictions and the proposed RST–SVM model provides better classification results than RST–BPN model, no matter when only considering financial ratios or the model including both financial ratios and DEA.

1. Introduction

The prediction of business failure is an important and challenging issue that has served as the impetus for many academic studies over the past three decades (Altman, 1968; Beaver, 1966; Bryant, 1997; Ohlson, 1980). Business failure is a general term and, according to a widespread definition, is the situation that a firm cannot pay lenders, preferred stock shareholders, suppliers, etc., or a bill is overdrawn, or the firm is bankrupt according to the law (Ahn, Cho, & Kim, 2000). Widely identified causes and symptoms of business failure include poor management, autocratic leadership and difficulties in operating successfully in the market. As the world's economy has been experiencing severe challenges during the past decade, more and more companies, no matter large or small, are facing the problems of filing bankruptcy. Thus, accurate business failure prediction models have drawn serious attention from both researchers and practitioners aiming to provide on time signals for better investment and government decisions with timely warnings.

Many different useful techniques have already been investigated in the course of comparative studies related in several review articles (Altman, 1984; Dimitras, Zanakis, & Zopounidis, 1996; Jones, 1987; Keasey & Watson, 1991; Scott, 1981; Zavgren, 1983) in order to solve the problems involved during the evaluation process. Recently, Kumar and Ravi (2007) gave a complete review of methods used for the prediction of business failure and of new trends in this area. Basically, the business failure prediction models use appropriate independent variables to "predict" a company is a healthy company or a bankrupt one. Therefore, the business failure prediction problems are in the scope of the more general and widely discussed discrimination and classification problems (Johnson & Wichern, 2002). However, while these well-established techniques are there to solve business failure prediction problems and applications, two main problems arise.

First, after Beaver (1966) and Altman (1968) used the financial ratios methodology in conducting business failure predictions, most of the studies only considered financial ratios as independent (input) variables. Although financial ratios, originated in a corporation's financial statements, can reflect some characteristics of a

* Corresponding author. Address: Department of Accounting, Chinese Culture University, 55, Hwa-Kang Road, Yang-Ming-Shan, Taipei 11114, Taiwan, ROC. Tel.: +886 2 2861 0511x35505; fax: +886 2 2861 4177.
E-mail address: djchi@ms43.hinet.net (D.-J. Chi).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.06.088

corporation from various aspects to a certain extent. While the efficiency of a corporation's management is generally acknowledged to be a key contributor to corporation's bankrupt (Gestel et al., 2006; Seballos & Thomson, 1990; Secrist, 1938), it is usually excluded from early prediction models. Therefore, in this study, we believe efficiency which reflects the status of the management of a corporation in business failure predictions will be decisive factors affecting the predictive capability.

For a typical efficiency measurements, for example, research on operational efficiency—the most widely studied efficiency issue—assumes the resources of a corporation as inputs (e.g., personnel, technology, space, etc.) and some measurable form of the services provided as output (e.g., number of accounts serviced or loans and other transactions processed, etc.). However, it is hard to evaluate the efficiency of a corporation directly from its financial statements. An approach known as data envelopment analysis (DEA) may serve to offer useful insights into the manner by incorporating multiple inputs and outputs; DEA is able to provide measures for the efficiency of a corporation.

Secondly, early studies of business failure prediction used statistical techniques such as univariate statistical methods, multiple discriminant analysis (MDA), linear probability models, and logit and probit analysis have mainly been used for business classification problems (Altman, 1968; Altman, Haldeman, & Narayanan, 1977; Collins & Green, 1972). These conventional statistical methods, however, have some restrictive assumptions such as the linearity, normality and independence among predictor or input variables. Considering that the violation of these assumptions for independent variables frequently occurs with financial data (Deakin, 1972), the methods can have limitations to obtain the effectiveness and validity.

Artificial intelligence approaches that are less vulnerable to these assumptions, such as inductive learning, Neural networks (NN) can be alternative methodologies for classification problems to which traditional statistical methods have long been applied. NN have shown to have better predictive capability than MDA and logistic regression in business failure prediction problems (Coleman, Graettinger, & Lawrence, 1991, Rahimian Salchengerger, Cinar, & Lash, 1993, Salchengerger et al., 1992, Sharda & Wilson, 1996, Tam & Kiang, 1992, Wilson & Sharda, 1994, Zhang, Hu, Patuwo, & Indro, 1999). Recently, support vector machines (SVM), developed by Vapnik (1995), have gained popularity due to many attractive features and excellent generalization performance on a wide range of problems. Also, SVM embody the structural risk minimization principle (SRM), which has been shown to be superior to traditional empirical risk minimization principle (ERM) employed by conventional neural networks. It has been demonstrated by Min and Lee (2005) that SVM outperform NN, MDA and logistic regression in business failure prediction.

While there are several arguments that variable selection, also called feature selection, is a fundamental problem that has significant impact on the prediction accuracy of the models. Many methods have been developed to create the best preparation for data inputs. For doing good classification process in SVM, the preparation of data inputs for classifier needs special treatment to guarantee the good performance in the classifier. It is therefore not surprising that much research has been done on dimensionality reduction (Dash & Liu, 1997; Kira & Rendell, 1992; Langley, 1994). A technique that can reduce dimensionality using information contained within the data set and preserving the meaning of the features is clearly desirable. Rough set theory (RST) can be used as such a tool to discover data dependencies and reduce the number of attributes contained in a dataset by purely structural methods (Pawlak, 1991), have successfully been applied to real world classification problems (Ahn et al., 2000; Siegel, de Korvin, & Omer, 1993; Slowinski & Zopounidis, 1995).

The objective of the study is using efficiency as predictive variables and proposed a novel model to integrate RST with SVM technique, named RST–SVM, to increase the accuracy for the prediction of business failure. By RST approach, the redundant attributes in multi-attribute information table can be reduced, which showed that the number of independent variables was reduced with no information loss, is utilized as a preprocessor to improve business failure prediction capability by SVM. In the first stage, RST is selected for doing variable selection because of its reliability to obtain the significant independent variables. The second stage of the study will use the obtained significant independent variables from RST as inputs of SVM models. The obtained results can then be compared to see whether the one including efficiency variable will give better classification accuracy or not. In the proposed method, DEA is employed as a tool to evaluate the input/output efficiency. The effectiveness of the methodology was verified by experiments comparing back-propagation neural networks (BPN) approach with the hybrid approach (RST–BPN).

This paper is organized as follows. We will give a brief review of the DEA model used to evaluate the efficiency of a corporation in Section 2. Section 3 describes classification techniques used in previous researches concerned with our paper: RST and SVM, respectively. In Section 4, the proposed data preprocessing algorithm by RST and hybrid models is described. In Section 5, we analyze and compare the results of each model. Finally, discussion and conclusions are provided in Section 6.

2. Using DEA for evaluating the efficiencies

Data envelopment analysis (DEA) is an evaluation tool for decision making units (DMUs) and it solves many decision-making problems by integrating multiple inputs and outputs simultaneously. DEA is a non-parametric data analytic technique that is extensively used by various research communities (e.g., Hong, Ha, Shin, Park, & Kim, 1999; Seol, Choi, Park, & Park, 2007; Sohn & Moon, 2004). The basic ideas behind DEA date back to Farrell (1957) but the recent series of discussions started with the article by Charnes, Cooper, and Rhodes (1978). We give very briefly the salient features of DEA. More detailed information can be obtained elsewhere (Banker, Charnes, & Cooper, 1984; Charnes, Cooper, Lewin, & Seiford, 1993).

The DEA ration form, proposed by Charnes, Cooper and Rhodes (CCR) (1978), is designed to measure the relative efficiency or productivity of a specific DMUk. The DEA formulation is given as follows. Suppose that there is a set of n DMUs to be analyzed, each of which uses m common inputs and s common outputs. Let k ($k = 1,\ldots,n$) denote the DMU whose relative efficiency or productivity is to be maximized.

$$\begin{aligned} \text{Maximize} \quad & h_k = \frac{\sum_{r=1}^{s} u_{rk} Y_{rk}}{\sum_{i=1}^{m} v_{ik} X_{ik}} \\ \text{Subject to} \quad & \frac{\sum_{r=1}^{s} u_r Y_{rj}}{\sum_{i=1}^{m} v_i X_{ij}} \leqslant 1 \\ & u_r, v_i \geqslant 0 \\ & i = 1, 2, \ldots, m \\ & r = 1, 2, \ldots, s \\ & j = 1, 2, \ldots, n \end{aligned} \tag{1}$$

where u_{rk} is the variable weights of given to the rth output of the kth DMU, v_{ik} is the variable weights of given to the ith input of the kth DMU, u_{rk} and v_{ik} are decision variables determining the relative efficiency of DMU_k, Y_{rj} is the rth output of the jth DMU, and X_{ij} is the ith input of the jth DMU. It also assumes that all Y_{rj} and X_{ij} are positive. h_k is the efficiency score and is less than or equal to 1. When efficiency score of h_k is 1, DMU_k is called the efficient frontier.