Knowledge-Based Systems 24 (2011) 1380–1388

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

The random subspace binary logit (RSBL) model for bankruptcy prediction

Hui Li [a,d,*], Young-Chan Lee [b], Yan-Chun Zhou [c], Jie Sun [a]

[a] School of Economics and Management, Zhejiang Normal University, P.O. Box 62, 688 YingBinDaDao, Jinhua, Zhejiang 321004, China
[b] Division of Economics & Commerce, Dongguk University, GyeongJu Campus, Gyeongju, Gyeongbuk 780-714, Republic of Korea
[c] School of Business, Ningbo University, 818 FengHuaLu Road, Ningbo, Zhejiang 315211, China
[d] College of Engineering, The Ohio State University, 470 Hitchcock Hall, 2070 Neil Avenue, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 6 January 2011
Received in revised form 21 June 2011
Accepted 21 June 2011
Available online 29 June 2011

Keywords:
Bankruptcy prediction
Random subspace binary logit
Group decision of predictive models
Corporate failure prediction
Probit
Multivariate discriminant analysis

ABSTRACT

This paper proposes the random subspace binary logit (RSBL) model (or random subspace binary logistic regression analysis) by taking the random subspace approach and using the classical logit model to generate a group of diverse logit decision agents from various perspectives for predictive problem. These diverse logit models are then combined for a more accurate analysis. The proposed RSBL model takes advantage of both logit (or logistic regression) and random subspace approaches. The random subspace approach generates diverse sets of variables to represent the current problem as different masks. Different logit decision agents from these masks, instead of a single logit model, are constructed. To verify its performance, we used the proposed RSBL model to forecast corporate failure in China. The results indicate that this model significantly improves the predictive ability of classical statistical models such as multivariate discriminant analysis, logit model, and probit model. Thus, the proposed model should make logit model more suitable for predictive problems in academic and industrial uses.

1. Introduction

Investors, creditors, bankers, stockholders, and managers need effective tools for managing various risks associated with their decisions. Researches in this topic includes: Cho [9], Pai et al. [33], among others. One such tool is bankruptcy prediction, which refers to prediction of business failure through financial variables [11,14,15,26–28,34,36,38]. Early studies of bankruptcy prediction typically employed discriminant analysis. Fitzpatrick [10] provided an in-depth interpretation of bankruptcy variables and trends (i.e., he employed a type of multiple variable analysis); Beaver [3] proposed a framework for employing univariate analysis for bankruptcy prediction; and Altman [1] employed multivariate discriminant analysis (MDA) to predict bankruptcy. Because discriminant analysis is easy to understand, interpret, and explain, it has been suitable for industrial use. MDA assumes dichotomous data, a multivariate normal distribution, equal variance-covariance matrices across two groups, a specified prior probability of two groups, and the absence of multicollinearity [2]. However, the assumption of multivariate normality is often violated for bankruptcy data without effective solutions (e.g., only several financial variables distribute normally for bankruptcy data from China). Further, univariate normality is not a sufficient condition for multivariate normality. To address this problem, Martin [25] and Ohlson [29] used a conditional probability model to forecast bankruptcy. This type of model uses the nonlinear maximum likelihood method for estimating probability of bankruptcy. According to the assumption about probability distribution, logit and probit models assume a logistic distribution and a cumulative normal distribution, respectively. The relationship between financial variables and the probability of bankruptcy is typically assumed to be linear. Thus, linear logit model has typically been used. These two classical statistical models have been widely used for forecasting bankruptcy for many years [17,18,39,16]. Even now, firms frequently employ logit model to calculate probability of bankruptcy for their customers. Previous studies have attempted to improve the effectiveness of these two classical models by optimizing single model. For example, Refs. [32,31] proposed the Tabu method, a type of feature selection method, for discriminant analysis and logit model.

In the past two decades, a number of studies have investigated performance of intelligent models on bankruptcy prediction (e.g., case-based reasoning with the k-nearest neighbor as the heart, classification and regression tree, support vector machine, among others). However, the two statistical models are still very popular (particularly for industrial use) because they are well-known models for bankruptcy prediction and are easy to model, interpret, and explain. Logit model is used more frequently because it is less demanding than MDA. However, logit model has a drawback. Most of the recent studies (e.g., [22,23,40,19]) have demonstrated that predictive abilities of classical statistical models (e.g., logit model) are relative weak.

* Corresponding author at: School of Economics and Management, Zhejiang Normal University, P.O. Box 62, 688 YingBinDaDao, Jinhua, Zhejiang 321004, China. Tel.: +86 579 8229 8602.
E-mail address: lihuihit@gmail.com (H. Li).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.06.015

Predictive accuracy is one of the most important indicators of model effectiveness. If logit model is relative weak in predictive accuracy, the loss saved by using logit in bankruptcy prediction in industry will be smaller. However, it is difficult to introduce new models for industrial users for the following reasons. (1) Intelligent models are complex approaches for the problem, compared with statistical models. (2) Industrial practitioners are already familiar with the classical approaches. One is not likely to replace a familiar tool as long as it continues to perform reasonably well. Thus, the shortcoming of logit model—relative poor performance in terms of bankruptcy prediction—should be addressed to make it more suitable for industrial use. One may argue that the loss resulting from the use of logit model instead of intelligent models for bankruptcy prediction is not substantial because the difference in reductions in predictive error between statistical models and intelligent models is commonly 3%. However, a decision tool that can reduce predictive error by 3% could potentially save the industry approximately $1.2 billion annually [41]. Because logit model is widely used for bankruptcy prediction, we need to enhance its predictive ability for forecasting bankruptcy while preserving its advantages.

From the perspective of management science and ensemble learning, using various decision agents with diverse opinions is effective in improving performance of predictive model [30,20]. The predictive ability of a committee of decision agents may exceed that of a single decision agent. Logit model can be regarded as a decision agent for bankruptcy prediction. The current problem can be described and represented by some variables as a mask from which a decision agent will be constructed. The random subspace method is an approach for producing various representations (i.e., masks) that can be used to generate different decision agents. This approach injects randomness into problem representation by randomly selecting variables with replacement. This means that different variable sets are used to construct logit model.

Thus, to improve the analysis performance of logit model, the present study combines random subspace approach with binary logit model to generate the simple random subspace binary logit (RSBL) model that takes into account different decision agents' opinions. The results of practical application verifying its ability to forecast corporate failure in China indicate that the proposed model can forecast corporate failure significantly better than classical statistical models (i.e., MDA, logit model, and probit model). The paper is organized as follows. Section 2 introduces binary logit model and random subspace approach. Section 3 proposes the RSBL model. Section 4 uses the proposed model to forecast corporate failure in China, and Section 5 concludes.

2. Binary logit model and random subspace approach

2.1. Binary logit model for bankruptcy prediction

Logit model is used to forecast probability of an event by fitting data (represented by some variables) to the logistic curve. The term "logit," introduced by Berkson [4], is borrowed from probit model, a similar model introduced by Bliss [5]. In the field of bankruptcy prediction, the occurrence of an event refers to corporate failure. Further, the variables include financial variables calculated from firm's public financial statements. For example, the probability of a firm declaring bankruptcy in the future can be determined by analyzing various variables for the firm's profitability (e.g., various financial ratios). Assume that the result of bankruptcy prediction is either corporate failure or non-failure. The number of X_i is known, and the probability of pro_i is unknown. Here X_i refers to the ith object (observation), and a total of m observations exist. For each observation, a set of variables can be used to inform the probability of an event (e.g., a bankruptcy). Assume that there are k variables, namely $x_1, x_2, \ldots, x_k$. The logit value of the unknown binomial probability is modeled as a linear function:

$$\begin{aligned}\text{logit}(probability_i) &= \ln\left(\frac{probability_i}{1-probability_i}\right)\\ &= \beta_0+\beta_1x_1^i+\beta_2x_2^i+\cdots+\beta_kx_k^i.\end{aligned} \tag{1}$$

The unknown parameter β_j can be estimated through maximum likelihood estimation of generalized linear models. The greater the value of β_j is, the more the jth variable contributes to prediction. Further, β_0 refers to an intercept, and β_j ($j = 1, 1, \ldots, k$), the regression coefficient of the jth variable. Finally,

$$probability_i = \frac{1}{1+e^{-(\beta_0+\beta_1x_1^i+\beta_2x_2^i+\cdots+\beta_kx_k^i)}}. \tag{2}$$

Assume that

$$z = \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k. \tag{3}$$

The following function, which is dependent on z, is referred to as logistic regression:

$$f(z) = \frac{1}{1+e^{-z}}, \tag{4}$$

where $f(z) \in (0, 1)$ represents the probability of an event. Typically, the cutoff value is 0.5. Logistic regression is useful for describing the relationship between financial variables and the probability of corporate failure.

2.2. Random subspace approach

Random subspace approach refers to the construction of decision models through random selection of a number of variables from a given set of variables [13]. This approach reflects a type of ensemble method and is known to be able to improve predictive accuracy [6,7,8]. Each time a random subspace is generated, a decision agent will be produced by constructing a model on top of the representation of the current problem. Finally, all decision agents are integrated by a simple vote by the committee. This approach is described as follows. Consider the condition with m observations in a k-dimensional space:

$$\{(x_{1j}, x_{2j}, \ldots, x_{ij}, \ldots, x_{kj})|x_{ij}\}, \quad \text{where } i \in \{1, m\},\ j \in \{1, k\}. \tag{5}$$

i is the number of observations; j is the number of variables; and x_{ij} refers to the value of the ith observation for the jth variable. The random subspace represented by the randomly selected variables is expressed as follows:

$$\begin{aligned}&\{(x_{1j}, x_{2j}, \ldots, x_{mj})|x_{ij}\},\\ &\text{where}\\ &x_{ij} = x_{ij} \text{ for } i \in I, \text{ and}\\ &x_{ij} = Null \text{ for } i \notin I.\end{aligned} \tag{6}$$

I is the k'-dimensional subset of $\{1, 2, \ldots, k\}$; and $k' \leqslant k$. Here we assume that $k' = k$. Thus, the total number of variables in subspaces is the same as that of those variables in the initial space. Because some of the same variables are selected, the actual number of variables is less than k. The creation of random space is repeated P times. Each time, a random generator from 1 to k' is used to select a variable to be used in the subspace, and the process is repeated k' times. The source of randomness is based on re-sampling P times with replacement the blocks of m observations for each k variable. As a result, P random subspaces are created as different masks for the current problem. The same algorithm is implemented on top of these masks to generate diverse decision agents. According to Ho [13], this approach reflects a type of stochastic discrimination to increase