Expert Systems with Applications 38 (2011) 8336–8342

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Using partial least squares and support vector machines for bankruptcy prediction

Zijiang Yang [a], Wenjie You [b,c], Guoli Ji [b,*]

[a] School of Information Technology, York University, Toronto, Canada M3J 1P3
[b] Department of Automation, Xiamen University, 361005 Xiamen, China
[c] Department of Mathematics and Computer Science, Fujian Normal University, Fujian 350300, China

ARTICLE INFO

Keywords:
Partial least squares
Support vector machine
Bankruptcy prediction

ABSTRACT

The evaluation of corporate financial distress has attracted significant global attention as a result of the increasing number of worldwide corporate failures. There is an immediate and compelling need for more effective financial distress prediction models. This paper presents a novel method to predict bankruptcy. The proposed method combines the partial least squares (PLS) based feature selection with support vector machine (SVM) for information fusion. PLS can successfully identify the complex nonlinearity and correlations among the financial indicators. The experimental results demonstrate its superior predictive ability. On the one hand, the proposed model can select the most relevant financial indicators to predict bankruptcy and at the same time identify the role of each variable in the prediction process. On the other hand, the proposed model's high levels of prediction accuracy can translate into benefits to financial organizations through such activities as credit approval, and loan portfolio and security management.

1. Introduction

There have been many attempts to create a prediction model of business failure. The first publication appeared in 1968 and was authored by Beaver (1968) who created a univariate discriminant model using financial ratios selected by a dichotomous classification test. Now, bankruptcy prediction models utilize both statistical analysis and data mining techniques to refine the decision support tools and improve decision making. In addition to discriminant analysis, the traditional statistical methods include regression, logistic models, factor analysis, etc. The more recent data mining techniques include decision trees, neural networks (NNs), fuzzy logic, genetic algorithm (GA), support vector machine (SVM) and so forth. The statistical applications, although enhanced over time, were restricted by the rigorous assumptions of traditional statistics such as the linearity, normality, independence among predictor variables and pre-existing functional form related to criterion and predictor variables.

Balcaen and Ooghe (2006) presented an overview of the classic statistical methods for predicting business failure developed thus far and provided a detailed analysis of four types: (1) univariate analysis, (2) risk index models, (3) multivariate discriminant analysis, and (4) conditional probability model (logit, probit, linear probability models). The authors identified a number of problems in the application of these methods to bankruptcy prediction which include data anomalies and data instability, inappropriate sample and independent variable selection, and ignorance of time horizon. This study was limited to the statistical models only; the data mining approach has not been included in the scope of the paper.

Berg (2007) evaluated several models for bankruptcy forecasting such as linear discriminant analysis, generalized linear models, neural networks (NNs) and generalized additive model (GAM). His analysis used out-of-sample and out-of-time validation and demonstrated that GAM performed significantly better than other models at all risk levels. Moreover, the tests indicated that GAM was sensitive to the default horizon. The limitation of the model lies in its implementation since the method is considered computationally intensive, non-intuitive and requires approximations and dummy variables.

The data mining approach introduced pattern recognition and pattern classification abilities due to the nonlinear, non-parametric adaptive-learning properties of neural networks. Now, hybrid NN models for predicting bankruptcy using statistical and inductive learning methods and support vector machines have proven to yield superior predictive performance.

Min and Lee (2005) were one of the first to apply SVM to the problem of bankruptcy prediction. By mapping input vectors onto a high-dimensional feature space, the study showed that SVM transformed complex problems into simpler ones to which linear discriminant functions could be applied. At a performance level similar to or better than back propagation neural networks (BPNNs), SVM was able to conduct classification learning with a relatively small amount of data. The study implemented a grid-search technique using 5-fold cross-validation to choose optimal parameter values and the kernel

* Corresponding author.
E-mail address: glji@xmu.edu.cn (G. Ji).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.01.021

function. To validate the high prediction accuracy of this model a comparison of the model has been conducted against BPNs, MDA, and logit, respectively. The results of empirical analysis showed that SVM outperformed the other methods.

Min, Lee, and Han (2006) proposed a hybrid intelligence approach by integrating genetic algorisms and SVM to maximize the SVM predictive performance. Two of its aspects were concurrently targeted: feature subset selection and parameter optimization. The application of the proposed hybrid model showed its effectiveness in optimizing feature subset and parameters of SVM and therefore enhanced the accuracy of bankruptcy prediction. The study uncovered the underlying correlation between the feature subset and kernel parameters however did not implement or incorporate this relationship into the model.

Hung and Chen (2009) proposed a selective ensemble of three classifiers, i.e. the decision tree, the back propagation neural network and the support vector machine. Based on the expected probabilities of both bankruptcy and non-bankruptcy, this ensemble provides an approach which inherits advantages and avoids disadvantages of different classification techniques.

Tsai (2009) compared five well-known feature selection methods used in bankruptcy prediction, which are *t*-test, correlation matrix, stepwise regression, principle component analysis (PCA) and factor analysis (FA) to examine their predictive performance.

Min and Jeong (2009) propose a new binary classification method for predicting corporate failure based on a genetic algorithm, and to validate its prediction power through empirical analysis.

Another attempt at hybrid intelligent systems for bankruptcy prediction was made by Chandra, Ravi, and Bose (2009) who presented a novel hybrid intelligent system in the framework of soft computing to predict the failure of dotcom companies. The hybrid intelligent system comprises the techniques such as multilayer perceptrons (MLP), random forest (RF), logistic regression (LR), SVM, classification and regression trees (CART). In addition, Tsai and Wu (2008), Etemadi, Rostamy, and Dehkordi (2009), Ahn and Kim (2009), Cho, Kim, and Bae (2009), Ravi and Pramodh (2008), Nanni and Lumini (2009), Hu (2009), Nam, Kim, Park, and Lee (2008), Hu and Tseng (2007), and Tsakonas, Dounias, Doumpos, and Zopounidis (2006), to mention a few, also studied different models for bankruptcy prediction.

The above sampling of recently published literature on predicting corporate failure shows a vast number of approaches applied in an attempt to refine the classification model. Most forecasts achieved accuracy between 65% and 85%. This paper proposes the combination of a partial least squares (PLS) based feature selection along with SVM for information fusion. This method provides an attractive alternative to predict bankruptcy. PLS can successfully identify the complex nonlinearity and correlations among the financial indicators. The experimental results demonstrate the superior predictive ability of this proposed model. It is the first time that PLS has been introduced into bankruptcy prediction. This paper also validates the idea of pattern recognition in financial diagnosis using PLS. The purposes of this paper are to (1) illustrate a method for bankruptcy prediction based on the selection of the fewest number of financial indicators. On the one hand, it will improve the accuracy level in comparison to traditional bankruptcy prediction models. On the other hand, it will significantly decrease the data collection effort; (2) investigate how each financial indicator affects the bankruptcy of the companies; (3) improve the accuracy level of the bankruptcy prediction.

The rest of the paper is organized as follows. Section 2 discusses the models and methodology utilized in this paper. Section 3 gives an overview of the dataset used to validate our proposed method. Section 4 presents our findings and shows the comparison between the results from our proposed model and the other models implicated in the literature. Finally, our conclusions are presented in Section 5.

2. Methodology

2.1. Partial least squares (PLS)

PLS is a supervised feature extraction method. By principal component analysis and the synthesis of variable extraction, the most comprehensive explanatory variables that predicted the variable *Y* were extracted. PLS can separate the information and noise of the examined system so that the appropriate models can be established. The compression of information features based on PLS compresses the explanatory variable *X* and also takes into account the correlation with the predicted variable *Y*. Its results will have more practical meanings.

PLS extracts the first latent variable t_1 from the variable set *X*. t_1 extracts as much variation information of *X* as possible. At the same time, it extracts the first latent variable u_1 from the variable set *Y* to ensure the largest correlation between t_1 and u_1. Then the regression equations with *Y* and t_1 and the equations with *X* and t_1 are established. The algorithm terminates if the regression equations meet the accuracy requirements. Otherwise, the second latent variable t_2 was extracted from the residual information which has been interpreted by t_1 of *X*, and u_2 was extracted from the residual information which has been interpreted by t_1 of *Y*. Repeat this process until it reaches the required precision.

Assuming $X = (X_1, X_2, \ldots, X_p)$ which is *n* samples of *p*-dimensional index and forecast variable *Y*, the optimization can be formulated as below:

$$\begin{cases} Max & cov(Xw_i,\ Yc_i) \\ s.t. & w_i'w_i = 1;\ c_i'c_i = 1; \\ & w_i' \sum_X w_j = 0; \\ & c_i' \sum_Y c_j = 0; \end{cases}$$

where the linear combination $t_i = Xw_i$ was the *i*th latent variables, and $\Sigma_X = X'X$, $\Sigma_Y = Y'Y$.

The solution of the above optimization problem (w_i, c_i) can be found below (Lorber, Wangen, & Kowalski, 1987; Wold, Ruhe, Wold, & Dunn, 1984):

$$w_i = \begin{cases} \Sigma_{XY}\Sigma_{YX} \text{ main eigenvector}, & i = 1 \\ (I - P_X)\Sigma_{XY}(I - P_Y)\Sigma_{YX} \text{ main eigenvector}, & i > 1 \end{cases},$$

$$c_i = \begin{cases} \Sigma_{YX}w_i, & i = 1 \\ (I - P_Y)\Sigma_{YX}w_i, & i > 1 \end{cases}$$

where $P_X = (\Sigma_X W)[(\Sigma_X W)^T(\Sigma_X W)]^{-1}(\Sigma_X W)^T$, $P_Y = (\Sigma_Y C)[(\Sigma_Y C)^T(\Sigma_Y C)]^{-1}(\Sigma_Y C)^T$, $W = (w_{ij})$, $C = (c_{ij})$.

On the one hand, the component t_h extracted in the calculation of PLS represent as much as possible the variation information of *X*. On the other hand, it is associated with *Y* as much as possible to explain the information of *Y*. To measure the explanatory power of t_h to *X* and *Y*, the definition of a variety of explanatory power is defined where $r(x_i, x_j)$ indicates the correlation coefficient between the two variables.

Definition 1 (*Variation Explanation and Accumulation of Variation Explanation*). We define the Variation Explanatory Power of t_h to Y: $Rd(y_k; t_h) = r^2(y_k, t_h)$ is the variable variation explanation between component t_h and dependent variable y_k. $Rd(Y; t_h) = \frac{1}{q}\sum_{k=1}^{q} Rd(y_k; t_h)$ is the variable variation explanation between component t_h and dependent variable *Y*. $Rd(Y; t_1, \ldots, t_m) = \sum_{h=1}^{m} Rd(Y; t_h)$ is the accumulation of variation explanation between component $t_1, t_2, \ldots, t_m$ and dependent variable *Y*. $Rd(y_k; t_1, \ldots, t_m) = \sum_{h=1}^{m} Rd(y_k; t_h)$ is the accumulation of variation explanation between component $t_1, t_2, \ldots, t_m$ and dependent variable y_k.