Expert Systems with Applications 38 (2011) 6244–6253

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Empirical research of hybridizing principal component analysis with multivariate discriminant analysis and logistic regression for business failure prediction

Hui Li *, Jie Sun

School of Economics and Management, Zhejiang Normal University, 91 subbox in P.O. Box 62, YingBinDaDao 688, Jinhua City 321004, Zhejiang Province, PR China

ARTICLE INFO

Keywords:
Business failure prediction (BFP)
Principal component analysis (PCA)
Multivariate discriminant analysis (MDA)
Logistic regression (logit)

ABSTRACT

Predicting business failure of listed companies is a hot topic because of the emergency of financial crisis in developed countries recently. The two classical statistical methods of multivariate discriminant analysis (MDA) and logistic regression (logit) have taken a key role in the area of business failure prediction (BFP). However, they are frequently criticized for the relative low predictive performance compared with some newly developed predictive methods. The present research contributes to the construction of a new hybrid method for BFP by integrating principal component analysis (PCA) with MDA and logit to help them produce dominating predictive performance, a pilot study demonstrating the effectiveness of employing PCA with MDA and logit for BFP, even a pilot study demonstrating the usefulness of PCA in BFP. The hybrid method is implemented by using stepwise method of MDA, stepwise method of logit, and independent sample *t* test as the preprocessing procedure, on the basis of which PCA is further used to extract features for MDA and logit. The optimal feature set for a specific task of BFP is determined by an empirical means of splitting all available data for thirty times to obtain the optimal preprocessing procedure for PCA. In the empirical research, this study attempts to investigate whether or not the new hybrid method can produce dominating performance in short-term BFP of Chinese listed companies. For comparison, the most preferred filter approach of stepwise method of MDA for short-term BFP in China was employed to select optimal features for MDA and logit. Meanwhile, the use of PCA on all available data to extract features for MDA and logit to make predictions is also employed to make a comparison. Empirical results indicate that the new method, namely: the hybridization of PCA with MDA and logit, can produce dominating predictive performance in short-term BFP of Chinese listed companies. The preferred preprocessing procedure of PCA for short-term BFP of Chinese listed companies is stepwise method of MDA and corresponding performance of MDA and logit respectively outperforms all the other methods. Meanwhile, predictive performance of MDA and logit with the optimal PCA procedure is not significantly different.

1. Introduction and motivation

Business failure occurs when a company has serious losses and when the company becomes insolvent with liabilities. This phenomenon causes great loss to stock holders, managers, investors, and employees because of the rapid change in global economy. Industrials and academies need effective methods that can predict business failure accurately. Business failure prediction (BFP) is a specific application of nonlinear analysis and linear analysis in real world. The inside principle of BFP is to find hidden patterns in data by using intelligent and statistical techniques. Financial ratios and cash flow information are chief variables in BFP (McGurr & DeVaney, 1998).

The research mode of BFP is to find more accurate predictive models and predictors. Data for BFP is usually collected from public information of companies. Financial ratios take a key role in predicting business failure as various predictors. Thus, collecting as much financial ratios as possible as predictors to make a prediction is a fundamental step in the area. If financial ratios are at hand, methods of feature selection or feature extraction are commonly used to find optimal predictors for a specific task of BFP in a standalone means. Assume that all optimal features and data are ready, predictive models can be constructed on the basis of the data by using methodologies and technologies from statistic, artificial intelligence, data mining, and business computing.

When constructing a method for BFP, the problem of finding optimal features and finding optimal models should both be taken into consideration. Models used in this area chiefly belong to the two categories of statistical models and intelligent models (Ahn & Kim, 2009; Chauhan, Ravi, & Chandra, 2009; Cho, Hong, & Ha, 2010; Hardle, Lee, Schafer, & Yeh, 2009; Hu & Ansell, 2009; Kim & Kang, 2010; Kumar & Ravi, 2007; Li, Adeli, Sun, & Han, 2011;

* Corresponding author. Tel.: +86 158 8899 3616.
E-mail address: lihuihit@gmail.com (H. Li).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.11.043

Li, Lin, Yeh, & Lee, 2011; Li & Sun, 2010, 2011; Li, Sun, & Wu, 2010; Ravisankar & Ravi, 2010; Sun & Li, 2009; Tsai, 2009; Tseng & Hu, 2010; Yang, Platt, & Platt, 1999). Thus, appliers classify methods for BFP into two groups, that is, the group of statistical method and the group of intelligent method. Till now, classical statistical methods of multivariate discriminant analysis (MDA) and logistic regression (logit) are the most popular in the area of BFP (Altman & Saunders, 1998). However, lots of researches criticize the relative low predictive performance of the two classical statistical methods when comparing their predictive performance with that of newly developed intelligent predictive method.

The application of classical statistical methods in predicting business failure witnesses the foundation of the BFP. The two classical statistical methods of MDA and logit were applied to solve the problem of BFP long time ago and achieved satisfactory performance. Altman (1968) firstly emploied the method of MDA to predict business failure. Ohlson (1980) pioneered the study of using logit to predict business failure. Since the use of MDA and logit to predict business failure, the two classical statistical methods have all along served as challenge methods for the problem. For example, Taffler (1983) emploied the Z-score model of MDA with the data from UK to predict short-term business failure. Grice and Ingram (2001) examined three research questions using Altman's Z-score method, that is, (1) Is the method useful for predicting business failure in recent periods; (2) Is the method useful for predicting business failure of non-manufacturing companies; (3) Is the method useful for predicting financial stress conditions, with the conclusion that Altman's Z-score method is still useful to predict financial distress conditions. Recently, the two statistical methods are still vital in predicting business failure, for example, Yim and Mitchell (2005) used logit models and MDA as baseline methods for compassion with hybrid networks in predicting business failure. Tseng and Lin (2005) proposed a quadratic interval logit model by using a quadratic programming approach to deal with binary features in BFP. Jones and Hensher (2007) attempted to evaluate the theoretical and empirical significance of the presented multinomial nested logit model for the explanation and prediction of business failure. Hensher, Jones, and Greene (2007) used a multinomial error component logit model, an extension of the more familiar mixed logit model, to predict business failure.

In addition to the use of the two classical statistical methods of MDA and logit to predict business failure in standalone way, some researches also attempt to integrate them with newly developed predictive methods; for example, Laitinen and Laitinen (2000) attempted to combine logit model with Taylor's series expansion to predict short-term business failure. Cheng, Chen, and Fu (2006) attempted to predict business failure by combining logit model with neural network, which combination retains advantages of both methods and avoids disadvantages of both methods. Hua, Wang, Xu, Zhang, and Liang (2007) utilized logistic regression to decrease the empirical risk of support vector machine in solving the task of BFP. Sun and Li (2008b) used MDA and logit as components to construct multi-classifiers system for BFP.

This research addresses prediction of business failure by hybridizing principal component analysis (PCA) with the two classical statistical methods of MDA and logit to construct a new hybrid predictive method. The three filters of stepwise method of MDA, stepwise method of logit, and independent sample *t* test are integrated into the hybrid method as preprocessing procedure of PCA. The optimal feature set for a specific task of BFP is determined by an empirical means of splitting all available data for thirty times to obtain the optimal preprocessing procedure for PCA.

This study attempts to make an investigation on whether or not the employment of PCA can help MDA and logit produce dominating predictive performance. What is of interest is whether the feature extraction method of PCA can produce more optimal features in helping MDA and logit use more information inside data to make a prediction. In order to fulfill this objective, this study uses PCA in three means. The commonly used filter approaches in BFP of Chinese listed companies are stepwise method of MDA, stepwise method of logit, ANOVA, and independent sample *t* test. However, the last two filters produce the same features with the data for short-term BFP of Chinese listed companies. Thus, the study uses PCA respectively on preprocessed features selected by stepwise method of MDA, preprocessed features selected by stepwise method of logit, and preprocessed features selected by independent sample *t* test. Corresponding features extracted are respectively named MPCA features, LPCA features, and TPCA features. The way of using PCA on all available features is also employed to make a comparison. Corresponding features are called PCA features. Stepwise method of MDA is the most preferable filter approach of short-term BFP of Chinese listed companies (Li & Sun, 2008, 2009). Corresponding feature set selected by stepwise method of MDA is called MDA feature set (MDAFS). Predictive performance of MDA and logit on the four PCA feature sets is to be compared with that of MDA and logit on MDAFS. This study is a pilot research focusing on hybridizing PCA with MDA and logit to predict business failure. Till now, seldom researches demonstrate the effectiveness of PCA in the area of BFP, which guarantees the originality of this research.

The paper is organized as follows. Section 2 presents the new hybrid method for BFP by integrating PCA with MDA and logit. Section 3 presents the design and results of the empirical research of the use of the new hybrid method with data from Chinese listed companies. Section 4 addresses detailed analysis and discussion on the results. Section 5 makes conclusion and outlines further research directions.

2. Research methodology

Principal component analysis (PCA) is a useful statistical technique for feature extraction. PCA can help a classifier produce more accurate predictive performance (Avci & Turkoglu, 2009). PCA is on the assumption that most information about classification is contained in the directions along which the feature values are the largest (Polat & Güneş, 2008). The results of a PCA are usually discussed in terms of component scores and loadings (Shaw, 2003). In PCA, the eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix is calculated, usually after mean centering the data for each attribute. Some researches call PCA as the discrete Karhunen–Loève transform (KLT), the Hotelling transform or proper orthogonal decomposition (POD) in different fields of application. When carrying out dimensionality reduction using PCA, those characteristics of the data set that contribute most to its variance are retained. Meanwhile, lower-order principal components are kept and higher-order ones are ignored.

The hybrid method by the integration of PCA with MDA and logit is illustrated as Fig. 1, which indicates that the hybrid method includes three levels, that is, data level, the level of hybrid method, and prediction level. The hybrid method is implemented by integrating PCA as the component of feature extraction. Stepwise method of MDA, stepwise method of logit, and independent sample *t* test are used to produce preprocessed results for the implementation of PCA. The extracted features by PCA are respectively inputted into MDA and logit to make a prediction. The key problem of the hybrid method is to find the optimal PCA procedure, which can be solved by an empirical means of splitting the whole data into training and testing parts for thirty times. The PCA procedures that respectively generate the best predictive performance for MDA and logit on the thirty testing data sets are regarded as the optimal ones. Finally, the optimal model of the hybrid method derives by a trial-and-error process between MDA methods and logit methods.