Contents lists available at ScienceDirect

Expert Systems with Applications

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

Applying enhanced data mining approaches in predicting bank performance: A case of Taiwanese commercial banks

Shih-Wei Lin [a,*], Yeou-Ren Shiue [b], Shih-Chi Chen [c], Hui-Miao Cheng [d]

[a] Department of Information Management, Chang Gung University, Taiwan
[b] Department of Information Management, Huafan University, Taiwan
[c] Department of Industrial Management, National Taiwan University of Science and Technology, Taiwan
[d] Department of Information Management, China University of Technology, Taiwan

ARTICLE INFO

Keywords:
Bank performance
Data mining
Particle swarm optimization
Parameter optimization
Feature selection

ABSTRACT

The prediction of bank performance is an important issue. The bad performance of banks may first result in bankruptcy, which is expected to influence the economics of the country eventually. Since the early 1970s, many researchers had already made predictions on such issues. However, until recent years, most of them have used traditional statistics to build the prediction model. Because of the vigorous development of data mining techniques, many researchers have begun to apply those techniques to various fields, including performance prediction systems. However, data mining techniques have the problem of parameter settings. Therefore, this study applies particle swarm optimization (PSO) to obtain suitable parameter settings for support vector machine (SVM) and decision tree (DT), and to select a subset of beneficial features, without reducing the classification accuracy rate. In order to evaluate the proposed approaches, dataset collected from Taiwanese commercial banks are used as source data. The experimental results showed that the proposed approaches could obtain a better parameter setting, reduce unnecessary features, and improve the accuracy of classification significantly.

1. Introduction

The structure of Taiwanese financial institutions has significantly been affected by the Southeast Asian Financial Crisis of July, 1997. Not only had the average overdue loan climbed from 3% in 1996 to 8.16% by the end of 2001, but also the average loss on assets before tax had reached 109 billion by 2002 (Condition & Performance of Domestic Banks). Undoubtedly, the statistical results show that if these problems are solved promptly it will become a great threat to the country.

The bank performance prediction model uses scientific and systematic approaches to diagnose the financial operations of institutes. According to a precise and strict evaluation, the model can detect the weakness of institutions in advance and provide early warning signals to related financial governments. The performance prediction for financial firms, especially banks, has been an extensively researched area since the late 1960s (Altman, 1968). Creditors, auditors, stockholders, and senior management are all interested in performance prediction (Wilson & Sharda, 1994). The most precise way of monitoring banks is by on-site examinations made by the Financial Supervisory Institutions. In Taiwan, the primary government agencies responsible for the supervision of financial institutions are the Central Bank of China (Taiwan), the Ministry of Finance, and the Central Deposit Insurance Corporation of Taiwan. These three regulators utilize a six-category rating system to indicate the safety and soundness of institutions. This rating, referred to as the CAMELS rating, evaluates banks according to their basic functional areas including Capital adequacy, Asset quality, Management expertise, Earnings strength, Liquidity, and Sensitivity to market risk. This system is made of various financial ratios, obtained from the periodic reports of the entities under jurisdiction (Kao & Liu, 2004).

Early research on this topic applied statistical methods. Beaver (1966) was the first one who used univariate analysis to build the financial prediction model for banks. Altman (1968) pointed out the drawback of Beaver's model, and further utilized discriminant analysis (DA) to build the model. After several years, Altman, Haldeman, and Narayanan (1977) developed a bankruptcy classification model called Zeta analysis, which incorporated comprehensive inputs. Martin (1977) presented logistic regression to predict the probability of bank failure based on the data obtained from the Federal Reserve System. West (1985) used factor analysis and logistic regression to create composite variables to describe a bank's financial and operating characteristics. Both Cielen, Peeters, and Vanhoof (2004) and Kao and Liu (2004) used data envelopment analysis (DEA) to predict the bankruptcy of banks. A comprehensive survey on statistical methods for the prediction of bank performance can be found in Ravi Kumar and Ravi (2007). Statistical

* Corresponding author. Tel.: +886 3 2118800; fax: +886 3 2118020.
E-mail address: swlin@mail.cgu.edu.tw (S.-W. Lin).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.03.029

methods in performance prediction build simple models using a small set of financial variables, and these simple models could classify about two-thirds of a holdout sample. These statistical models were succinct and were easy to explain. However, due to strict assumption, such as linearity and normality, independence among predictable variables limits the application in real world. The major problem in applying these methods to the performance prediction is that the multivariate-normality assumptions for independent variables are frequently violated in financial data sets, which may make these methods theoretically invalid for finite samples (Berrt & Linoff, 1997; Huang, Chen, Hsu, Chen, & Wu, 2004).

In recent years, it has been shown that data mining techniques are effective and efficient compared with the statistical methods in finance fields. Data mining techniques can find out the potential and significant information needed from the enormous data by rapid and deep exploration. For examples, Mochón, Quintana, and Sáez (2007) provided the rationale for using soft computing techniques in finance and presented several applications. Quintana, Saez, and Mochon (2007) proposed the evolutionary nearest neighbor classifier (evolutionary nearest neighbor classifier, ENPC) for early bankruptcy prediction. Oleda and Fernandez (1997) provided several classifiers to predict the case of bankruptcy. Their results pointed out that C4.5 and NN have better performance than DA and Logit. Ahn, Cho, and Kim (2000) proposed a hybrid intelligent system by combining a rough set approach and a neural network (NN) to predict the failure of firms based on the past financial performance data. The result showed that the proposed hybrid models outperform both DA and neural network models. de Andres, Landajo, and Lorca (2005) made a comparative analysis on forecasting business profitability of a Spanish case by various classification techniques. Their results pointed out that the artificial intelligence-based approaches are better than traditional statistical methods, such as LDA and Logit models. Ryu and Yue (2005) introduced a method called isotonic separation to the prediction of firm bankruptcy. They used feature reduction methods to reduce the ratios used in the prediction and then various classification methods, such as DA, decision tree (DT), neural networks (NN), learning vector quantization, rough sets, and isotonic separation, were used with reduced ratios. Their experimental results showed that the isotonic separation method is a promising technique, performing better than other methods for bankruptcy prediction. Rastogi and Shim (2000) developed an approach, called the PUBLIC, an improved decision tree classifier that integrated the second "pruning" phase with the initial "building" phase. Experimental results demonstrate the effectiveness of PUBLIC. Although the PUBLIC approach can reduce number of nodes and execution time, it is not designed for achieving the best classification accuracy.

To predict the performance of banks, feature data is required. Selecting the right set of features for classification is a difficult problem when designing a good classifier. Typically, one does not know a priori in which features are relevant for a particular classification task. One common approach is to collect as many features as possible prior to the learning and data-modeling phase. However, in most classification problems, given a large set of potential features, identifying a small subset to classify data object is generally necessary. Data without feature selection might be redundant or noisy, and decrease the classification efficiency.

Lee, Han, and Kwon (1996) used Korean bankruptcy data from 1979–1992 to build hybrid neural network models for bankruptcy prediction. In order to enhance the performance, they used multiple discriminant analysis (MDA) and DT methods, respectively, to reduce the number of input variables. Alam, Booth, Lee, and Thordarson (2000) adapted fuzzy clustering and self-organization neural networks for identifying potentially failing banks. The results showed that both the fuzzy clustering and self-organizing neural networks are promising classification tools for identifying potentially failing banks. Lin and McClean (2001) used four classifiers – DA, logistic regression, neural networks and DT in bankruptcy prediction. Each used two feature selection methods for predicting corporate failure: human judgment, based on financial theory, and the ANOVA statistical method. Park and Han (2002) applied the case-based reasoning (CBR) and feature weights to build the bankruptcy model. Huang et al. (2004) utilized support vector machine (SVM) and back propagation neural network (BPN) to build a model in a related research – credit analysis. Since there are many financial variables, they ran the ANOVA to test whether the differences are significant or not.

Tung, Quek, and Cheng (2004) proposed the use of a new neural fuzzy system to predict banking failure. Their experimental results reveal that hybrid approach has better performance to predict banking failure. Becerra, Galvao, and Abou-Seads (2005) proposed neural and wavelet network models for financial distress. The result showed that their approach is a valid alternative to the classical DA models. Moreover, wavelet networks may have advantages over the conventional multi-layer perceptron structures employed in neural network frameworks. However, feature selection problem was not investigated in their study. Ravi Kumar and Ravi (2006) proposed an ensemble classifier using simple majority voting scheme to bankruptcy prediction in Banks. The experimental results showed that ensemble classifier could perform better than stand-alone classifier. Shin, Lee, and Kim (2005) compared the predictive accuracy of SVM with that of artificial neural network (ANN), MDA and learning vector quantization (LVQ). The performance of SVM was found to be the highest. They applied two stages of input variable selection process. At the first stage, they selected 52 variables among more than 250 financial ratios by an independent-sample *t*-test. In the second stage, they selected 10 variables using an MDA stepwise method to reduce dimensionality. Min and Lee (2005) utilized a grid search method (Hsu, Chang, & Lin, 2003) to find the appropriate parameter setting for SVM and used principal component analysis (PCA) to reduce the number of multi-dimensional financial ratios to two factors. They compared the SVM model with MDA, logistic regression and BPN, and showed that the SVM model has the best accuracy rate. Nevertheless, these approaches are usually applied either to the statistical feature selection or to parameters pretest, or applied them in sequentially. To the best of our knowledge, few studies have simultaneously considered the feature selection and the optimal parameter setting in predicting bank performance.

The purpose of this study is to apply particle swarm optimization (PSO) for performing a feature selection and a parameter determination for two well-known data mining techniques: SVM and DT, and they are called PSO–SVM and PSO–DT, respectively. In order to verify the performance of the proposed approaches, the dataset collected from the Taiwanese commercial banks is used to predict the bank performance.

The remainder of this paper is organized as follows: Section 2 provides an overview of SVM, DT, feature selection and particle swarm optimization (PSO). Section 3 then introduces how the proposed PSO-based approaches are used to perform feature selection and parameter determination for SVM and DT. Experimental results are compared with some of the existing approaches in Section 4. Conclusions are finally drawn in Section 5, along with recommendations for future research.

2. Research background

2.1. Support vector machine

SVM can be briefly described as follows (Burgers, 1998; Cristianini & Shawe-Taylor, 2000). Let $(x_1,y_1\ ),\dots,(x_m,y_m)\in X\times\{\pm1\}$.