Expert Systems with Applications 39 (2012) 5325–5331

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine

Gang Wang [a,b,c,*], Jian Ma [c]

[a] School of Management, Hefei University of Technology, Hefei, Anhui 230009, PR China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei, Anhui, PR China
[c] Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Enterprise credit risk assessment
Ensemble learning
Bagging
Random subspace
SVM

ABSTRACT

Enterprise credit risk assessment has long been regarded as a critical topic and many statistical and intelligent methods have been explored for this issue. However there are no consistent conclusions on which methods are better. Recent researches suggest combining multiple classifiers, i.e., ensemble learning, may have a better performance. In this paper, we propose a new hybrid ensemble approach, called RSB-SVM, which is based on two popular ensemble strategies, i.e., bagging and random subspace and uses Support Vector Machine (SVM) as base learner. As there are two different factors, i.e., bootstrap selection of instances and random selection of features, encouraging diversity in RSB-SVM, it would be advantageous to get better performance. The enterprise credit risk dataset, which includes 239 companies' financial records and is collected by the Industrial and Commercial Bank of China, is selected to demonstrate the effectiveness and feasibility of proposed method. Experimental results reveal that RSB-SVM can be used as an alternative method for enterprise credit risk assessment.

1. Introduction

The enterprise credit risk assessment has long been regarded as important and widely studied issue in the academic and business community. In recent years, enterprise credit risk assessment has been become one of the primary ways for financial institutions to assess credit risk, improve cash flow, reduce possible risks and make managerial decisions (Huang, Chen, Hsu, Chen, & Wu, 2004; Huang, Chen, & Wang, 2007; Wang, Hao, Ma, & Jiang, 2011). For the enterprise credit risk assessment, the accuracy is quite significant to financial institutions' profitability. For example, the accuracy of assessment increases only one percent may retrieve a great loss for financial institutions (Hand & Henley, 1997).

Some statistical methods have been widely applied to build the enterprise credit risk assessment models, such as Linear Discriminant Analysis (LDA) (Karels & Prakash, 1987; Reichert, Cho, & Wagner, 1983), Logistic Regression Analysis (LRA) (Thomas, 2000; West, 2000), Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991). However, the problem with applying these statistical methods to enterprise credit risk assessment is that some assumptions, such as the multivariate normality assumptions for independent variables, are frequently violated in reality, which makes these methods theoretically invalid for finite samples (Huang et al., 2004).

In recent years, many studies have demonstrated that intelligent methods, such as Artificial Neural Network (ANN) (Desai, Crook, & Overstreet, 1996; West, 2000), Decision Tree (DT) (Hung & Chen, 2009; Makowski, 1985), Case Based Reasoning (CBR) (Shin & Han, 2001; Wheeler & Aitken, 2000) and Support Vector Machine (SVM) (Baesens et al., 2003; Huang et al., 2007; Schebesch & Stecking, 2005) can be used as alternative methods for enterprise credit risk assessment. In contrast with statistical methods, intelligent methods do not assume certain data distributions. These methods automatically extract knowledge from training data. According to previous studies, intelligent methods are superior to statistical methods in dealing with enterprise credit risk assessment problems, especially for nonlinear pattern classification (Huang et al., 2004). Among them, one of the most effective methods is SVM and has been successfully applied into enterprise credit risk assessment. However, the practical SVM has been implemented based on the approximation algorithm to reduce the cost of time and space. So, the obtained classification performance is far from the theoretically expected (Baesens et al., 2003; Huang et al., 2007; Schebesch & Stecking, 2005). Baesens et al. (2003) applied SVM, along with other classifiers to several enterprise credit risk datasets. They reported that SVM performs well in comparison with other algorithms, but do not always give the best performance. Schebesch and Stecking (2005) applied SVM to a database of applicants for building enterprise credit risk assessment model. They concluded

* Corresponding author at: Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong. Tel.: +852 9799 0955; fax: +852 2788 8694.
E-mail address: wgedison@gmail.com (G. Wang).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.11.003

that SVM performs slightly better than LRA, but not significantly so. Huang et al. (2007) found SVM classifies enterprise credit applications no more accurately than ANN, DT and Genetic Algorithms (GA), and compared the relative importance of using features selected by GA and SVM along with ANN and genetic programming.

To overcome these limitations, recently, integrating multiple classifiers into an aggregated output, i.e., ensemble method, has been turned out to be an efficient strategy for achieving high classification performance, especially when the base classifiers have different structures that lead to independent prediction errors (Breiman, 1996; Dietterich, 2000; Opitz & Maclin, 1999; Schapire, 1990; Wolpert, 1992). Yu, Wang, and Lai (2008) proposed a multi-stage neural network ensemble learning model to evaluate credit risk. Experimental results revealed the proposed neural network ensemble learning model can provide a promising solution to credit risk analysis. Tsai and Wu (2008) investigated the performance of a single classifier as the base learner to compare with multiple classifiers and diversified multiple classifiers by using neural networks. By comparing with the single classifier as the benchmark in terms of average accuracy, the ensemble method performs better. Nanni and Lumini (2009) investigated the performance of several systems based on ensemble methods for enterprise credit risk assessment. The results showed that ensemble methods may be used for boosting the performance of "stand-alone" classifier. Hung and Chen (2009) proposed a selective ensemble model of three classifiers, i.e., DT, ANN and SVM for enterprise credit risk assessment. Based on the expected probabilities of credit risk, this ensemble method provides an approach which inherits advantages and avoids disadvantages of different classification methods.

Base on the above motivation, we propose a new hybrid ensemble approach, called RSB-SVM, which is based on two popular ensemble strategies, i.e., bagging and random subspace and use SVM as base learner for enterprise credit risk assessment. Both theoretical and experimental researches show that combining a set of accurate and diverse classifiers will lead to a powerful classification system (Breiman, 1996; Dietterich, 2000; Opitz & Maclin, 1999; Schapire, 1990). For the first condition, accuracy, we choose SVM as the base learner. And for the diversity, among the diverse ensemble methods that are available, bagging and random subspace are two more often used methods and have been found to be accurate, computationally feasible across various data domain. In addition, it has been observed that an important prerequisite for ensemble methods to reduce the test error is that it generates a diversity of ensemble members (Breiman, 1996; Dietterich, 2000; Opitz & Maclin, 1999). However, for bagging, the only factor encouraging diversity is the proportion of different objects in the training samples. Although the classifier techniques used in bagging are sensitive to small changes in data, the bootstrap sampling appears to lead to ensembles of low diversity compared to other ensemble methods, e.g., boosting. In order to encourage diversity, we can use random subspace strategy to select a subset of features as input. As a result, we introduce random subspace strategy into Bagging SVM and get RSB-SVM. As there are two different factors, i.e., bootstrap selection of instances and random selection of features, encouraging diversity in RSB-SVM, it would be advantageous to get better performance. For the testing and illustration purposes, the enterprise credit risk dataset, which includes 239 companies' financial records from China and is collected by the Industrial and Commercial Bank of China, is selected to demonstrate the effectiveness and feasibility of proposed method. Experimental results reveal that RSB-SVM gets the best performance among eight methods, i.e., SVM, Bagging SVM, Random Subspace SVM, Boosting SVM, LRA, DT and ANN. And the non-linear kernel of SVM is more feasible than the linear kernel in the enterprise credit risk assessment practice. All these results illustrate that RSB-SVM can be used as an alternative method for enterprise credit risk assessment.

The remainder of the paper is organized as follows. In Section 2, the background of SVM, bagging and random subspace are presented. In Section 3, we propose a new hybrid ensemble approach, RSB-SVM, based on the bagging and the random subspace for enterprise credit risk assessment. In Section 4, we present the details of experiment design and report experimental results. Based on the observations and results of these experiments, Section 5 draws conclusions and future research directions.

2. Background

2.1. Support Vector Machine

As a relatively new class of machine learning techniques based on statistical learning theory (Cortes & Vapnik, 1995; Vapnik, 1995), SVM for enterprise credit risk assessment has obtained several state-of-art results in classification accuracy. In SVM, original input space is mapped into a high-dimensional dot product space called a feature space, and in the feature space the optimal hyperplane is determined to maximize the generalization ability of the classifier. The optimal hyperplane is found by exploiting the optimization theory, and respecting insights provide by the statistical learning theory. Fig. 1 shows an illustration of the idea of an optimal hyperplane for linearly separable patterns.

Given a set of training samples $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i \in R^n$ is the vector space pattern, $y_i \in \{-1, 1\}$ is the class label for a 2-class problem, SVM for classification attempts to find a classifier $f(x)$, which minimizes the expected misclassification rate. A linear classifier $f(x)$ is a hyperplane, and can be represented as $f(x) = \mathrm{sgn}(w^T x + b)$. Finding the optimal classifier $f(x)$ in SVM is equivalent to solving a convex quadratic optimization problem in (1):

$$\max_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{1a}$$

$$\text{Subject to } y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \quad (\xi_i \geq 0, i = 1, \ldots, N) \tag{1b}$$

where C is called the regularization parameter, and is used to balance the classifier's complexity and classification accuracy on the training set D. This quadratic problem is generally solved through its dual formulation. Simple replacing the involved vector inner-product with a non-linear kernel function converts linear SVM into a more flexible non-linear SVM, which is essence of the famous kernel trick. Any function satisfying Mercer's condition can be used as the kernel function (Vapnik, 1995). Some typical kernel function are:

$$\text{Linear}: K(x_i, x_j) = x_i^T x_j \tag{2}$$

$$\text{Polynomial}: K(x_i, x_j) = (x_i^T x_j + 1)^d \tag{3}$$

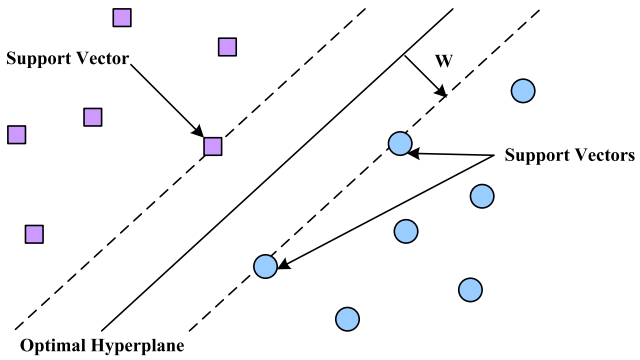

Fig. 1. A linear separable Support Vector Machine.