Expert Systems with Applications 36 (2009) 7515–7518

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Integrating nonlinear graph based dimensionality reduction schemes with SVMs for credit rating forecasting

Shian-Chang Huang *

Department of Business Administration, National Changhua University of Education, College of Management, No. 2, Shi-Da Road, Changhua 500, Taiwan

ARTICLE INFO

Keywords:
Kernel graph embedding
Dimensionality reduction
Support vector machine
Multi-class classification
Credit rating

ABSTRACT

By integrating graph based nonlinear dimensionality reduction with support vector machines (SVMs), this study develops a novel prediction model for credit ratings forecasting. SVMs have been successfully applied in numerous areas, and have demonstrated excellent performance. However, due to the high dimensionality and nonlinear distribution of the input data, this study employed a kernel graph embedding (KGE) scheme to reduce the dimensionality of input data, and enhance the performance of SVM classifiers. Empirical results indicated that one-vs-one SVM with KGE outperforms other multi-class SVMs and traditional classifiers. Compared with other dimensionality reduction methods the performance improvement owing to KGE is significant.

1. Introduction

Credit rating assesses the credit worthiness of an individual, corporation, or even a country. Typically, credit rating tells a lender or investor the probability of the borrower being able to pay back a loan. Consequently, credit ratings are important determinants of risk premiums and even the marketability of corporate bonds.

Recently, credit rating forecasting had been a critical issue in the banking industry. All banking institutes and their regulators attempt to search for a precise internal credit system to model the credit quality of their evaluation borrowers. Furthermore, subprimemortgage crisis in the later half of 2007 profoundly impacts the banking sector of US. The bank with the most accurate estimation of its credit risk will be the most profitable. The objective of this study is thus to develop a reliable and accurate prediction models for risk assessment.

The development of the corporate credit rating prediction model has attracted lots of research interests in academic and business community. Many researchers have attempted to construct automatic classification systems using methods from data mining, such as statistical and artificial intelligence techniques. However, due the high dimensionality of input variables (both financial and non-financial information), this study combined a kernel graph embedding (KGE) scheme proposed by Yan et al. (2007) with multi-class SVMs to enhance our predictions.

Numerous classification techniques have been adopted for credit scoring. These techniques include (1) traditional statistical methods; for example, discriminant analysis, logistic regression (Steenackers & Goovaerts, 1989; Stepanova & Thomas, 2001), and Bayesian network, (2) non-parametric statistical models, such as k-nearest neighbor (Henley & Hand, 1997), (3) decision trees (Yobas, Crook, & Ross, 2000), and (4) neural networks (Desai, Crook, & Overstreet, 1996; West, 2000; Yobas et al., 2000). Recently, the support vector machine (SVM) method (Cristianini & Shawe-Taylor, 2000; Schoelkopf, Burges, & Smola, 1999; Vapnik, 1999), another form of neural networks, has become increasingly popular and is currently regarded as the state-of-the-art technique for regression and classification applications. The formulation of SVM is believed to embody the structural risk minimization principle (a maximum margin classifier), and thus to combine excellent generalization properties with a sparse model representation.

SVMs exploit the idea of mapping input data into a high dimensional reproducing kernel Hilbert space (RKHS) where linear classification is performed. However, owing to the large amounts of data from public financial statements which can be used for corporate credit rating predictions, the large scale input data will make SVM classifiers infeasible due to the curse of dimensionality. Consequently, one needs to select key features from the raw data to reduce the dimensionality of the classification problem. Dimensionality reduction have been studied extensively in both the statistics and machine learning communities during recent decades. Among dimensionality reduction, the linear algorithms principal component analysis (PCA) and linear discriminant analysis (LDA) have been the two most popular because of their relative simplicity and effectiveness.

However, as indicated by Yan et al. (2007), in many real world problems there is no evidence that the data is sampled from a linear subspace. This motivates researchers to consider manifold based techniques for dimensionality reduction. Recently, various

* Tel.: +886 4 7232105 7420, mobile: +886 953092968; fax: +886 4 7211292.
E-mail address: shhuang@cc.ncue.edu.tw

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.09.047

manifold learning techniques, such as ISOMAP (Tenenbaum, de Silva, & Langford, 2000), locally linear embedding (LLE) (Roweis & Saul, 2000) and Laplacian eigenmap (Belkin & Niyogi, 2001) have been proposed which reduce the dimensionality of a fixed input data set in a way that maximally preserve certain inter-point relationships. This research adopted the general framework of Yan et al. (2007) called kernel graph embedding (KGE) for dimensionality reduction. Their framework offers a unified view for understanding and explaining many of the popular dimensionality reduction algorithms. The kernelization of graph embedding applies the kernel trick on the linear graph embedding algorithm. Thus it can handle data with nonlinear distributions.

To handle the high dimensionality of the input data, this study combined KGE with SVMs to increase the rating accuracy. KGE will reduce the dimensionality of input data and simultaneously eliminate irrelevant features. This combination will reduce the computational loading of SVMs and enhance the forecasting accuracy. Moreover, this study applies three types of multi-class SVMs, one-vs-one, one-vs-all, and multi-class SVMs to classify enterprize credit rating, and compares these SVM classifiers with traditional classifiers. Empirical results indicated that the performance of SVMs with KGE are promising. The performance improvement owing to KGE is significant. The method developed here will help financial institutions make good assessments about their credit risks, and substantially reduce their losses.

The remainder of this paper is organized as follows: Section 2 describes the multi-class SVMs. Section 3 introduces the KGE algorithm. Subsequently, Section 4 describes the study data and discusses the empirical findings. Conclusions are finally given in Section 5.

2. Support vector machines

The support vector machines (SVMs) were proposed by Vapnik (1999). Based on the structured risk minimization (SRM) principle, SVMs seek to minimize an upper bound of the generalization error instead of the empirical error as in other neural networks. SVM classifiers construct a hyperplane to separate the two classes (labeled $y \in \{-1, 1\}$) so that the margin (the distance between the hyperplane and the nearest point) is maximal. The SVM classification function is formulated as follows:

$$y = \mathrm{sign}(\mathbf{w}^T\phi(\mathbf{x}) + b), \tag{1}$$

where $\phi(\mathbf{x})$ is called the feature, which is a nonlinear mapping from the input space $\mathbf{x}$ to the future space. The coefficients $\mathbf{w}$ and b are estimated by the following optimization problem:

$$\min_{\mathbf{w},b} R(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i, \tag{2}$$

with

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b_i) \geqslant 1 - \xi_i, \quad i = 1, \dots, l \tag{3}$$

$$\xi_i \geqslant 0, \quad i = 1, \dots, l, \tag{4}$$

where C is a prescribed parameter, which evaluates the trade-off between the empirical risk and the smoothness of the model.

After taking the Lagrangian and conditions for optimality, the dual solution of this convex optimization problem can be formulated as follows:

$$\max_{\alpha} D(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

with constraints,

$$0 \leqslant \alpha_i \leqslant C, \quad i = 1, \dots, l \tag{6}$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0, \tag{7}$$

where α are Lagrangian multipliers, which are also the solution to the dual problem, and $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function. b follows from the complementarity Karush–Kuhn–Tucker (KKT) conditions. The decision function is given by

$$f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b\right). \tag{8}$$

The value of the kernel is equal to the inner product of two vectors $\mathbf{x}$ and $\mathbf{x}_i$ in the feature space, such that $K(\mathbf{x}, \mathbf{x}_i) = \phi(\mathbf{x})\phi(\mathbf{x}_i)$. Any function that satisfying Mercer's condition (Vapnik, 1999) can be used as the kernel function.

2.1. Multi-class support vector machine

One approach to solving multi-class classification problem is to consider the problem as a collection of binary classification problems. k classifiers can be constructed, one for each class. The nth classifier constructs a hyperplane between class n and the $k-1$ other classes. A majority vote across the classifiers or some other measure can then be applied to classify a new point. That is, a particular point is assigned to the class for which the distance from the margin, in the positive direction (i.e., in the direction in which class "one" lies rather than class "rest"), is maximal. This is one-against-rest method for multi-class classification.

Alternatively, $C_2^k = \frac{k(k-1)}{2}$ hyperplanes can be constructed, separating each class from each other class, and similarly some voting schemes can be applied. This is one-against-one method. The above two methods have been used widely in the support vector literature to solve multi-class classification problems.

Another way to solve multi-class problems is to construct a decision function by considering all classes at once (Weston & Watkins, 1999). One can generalize (2) to the following setting:

$$\min_{\mathbf{w},b} R(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}_m\|^2 + C\sum_{i=1}^{l}\sum_{m \neq y_i}\xi_i^m, \tag{9}$$

with

$$\mathbf{w}_{y_i}^T\phi(\mathbf{x}_i) + b_{y_i} \geqslant \mathbf{w}_m^T\phi(\mathbf{x}_i) + b_m + 2 - \xi_i^m \tag{10}$$

$$\xi_i^m \geqslant 0, \qquad i = 1, \dots, l \quad m \in \{1, \dots, k\} \setminus y_i. \tag{11}$$

This gives the decision function:

$$f(\mathbf{x}) = \arg\max_{k}(\mathbf{w}_i^T\phi(\mathbf{x}) + b_i), \quad i = 1, \dots, k. \tag{12}$$

One can also find the solution to this optimization problem in dual variables by finding the saddle point of the Lagrangian. This method is termed as MSVM.

3. Kernel graph embedding

In this section, we present the dimensionality reduction method of Yan et al. (2007) and Cai et al. (2007). Given m samples $x_i|_{i=1}^m \in R^n$, dimensionality reduction aims at finding $y_i|_{i=1}^m \in R^d$, $d \ll n$, where y_i can represents x_i. In the past decades, many algorithms, either supervised or unsupervised, have been proposed to solve this problem. These algorithms can all be interpreted in a general graph embedding framework of Yan et al. (2007).

Given a graph G with m vertices, each vertex represents a data point. Let W be a symmetric $m \times m$ matrix with W_{ij} having the weight of the edge joining vertices i and j. The G and W can be defined to characterize certain statistical or geometric properties of the data set. The purpose of graph embedding is to represent each