Expert Systems with Applications 38 (2011) 8607–8611

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Using Gaussian process based kernel classifiers for credit rating forecasting

Shian-Chang Huang *

Department of Business Administration, National Changhua University of Education, Changhua, Taiwan

ARTICLE INFO

Keywords:
Gaussian process
Support vector machine
Kernel classifier
Multi-class classification
Credit rating forecasting

ABSTRACT

The subprime mortgage crisis have triggered a significant economic decline over the world. Credit rating forecasting has been a critical issue in the global banking systems. The study trained a Gaussian process based multi-class classifier (GPC), a highly flexible probabilistic kernel machine, using variational Bayesian methods. GPC provides full predictive distributions and model selection simultaneously. During training process, the input features are automatically weighted by their relevances with respect to the output labels. Benefiting from the inherent feature scaling scheme, GPCs outperformed convectional multi-class classifiers and support vector machines (SVMs). In the second stage, conventional SVMs enhanced by feature selection and dimensionality reduction schemes were also compared with GPCs. Empirical results indicated that GPCs still performed the best.

1. Introduction

The subprime mortgage crisis is an ongoing financial crisis in the United States. This caused a ripple effect across the financial markets and global banking systems. In the crisis, credit risk assessment profoundly impacts the banking sector. The bank with the most accurate estimation of its credit risk will be the most profitable. On the other hand, corporate credit ratings are typically very costly to obtain, Standard & Poors in terms of time and human resources to perform deep analysis of a company's risk status based on various aspects ranging from strategic competitiveness to operational details. As a result, not all companies can afford annually updated credit ratings from these agencies, making credit rating prediction valuable to the investment community and banks.

Although rating agencies claim that both financial and non-financial information is considered in the rating decision process, their rating criteria are not explicit. Consequently, many researchers have attempted to construct automatic classification systems using methods from data mining, such as statistical and artificial intelligence (AI) techniques. The objective of this study is to develop reliable prediction models based on a new method developed by AI, the Gaussian process based classifiers (GPCs). Due the high dimensionality of input variables (relevant or irrelevant), this study trained the GPCs by a fast variational Bayesian algorithm proposed by Girolami and Rogers (2006) to reduce the computational loading of our predictions.

Nowadays, financial institutions' loan portfolios expand rapidly. These institutions are actively investigating various alternatives to improve the accuracy of their credit scoring practices. Improving scoring accuracy by even a fraction of a percent can translate into significant future savings. Thus, 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, kernel classifiers (such as support vector machines, SVMs) exploit the idea of mapping input data into a high dimensional reproducing kernel Hilbert space (RKHS) where linear classification is performed. SVMs (e.g., Cristianini & Shawe-Taylor, 2000; Vapnik, 1995), another form of neural networks, have been gaining popularity and has been regarded as the state-of-the-art technique for regression and classification applications with many successful applications (Schölkopf, Burges, & Smola, 1999; Schölkopf & Smola, 2002; Huang, 2008; Huang & Wu, 2008; Huang, Chuang, Wu, & Lai, 2010). It is believed that the formulation of SVM embodies the structural risk minimization principle, thus combining excellent generalization properties with a sparse model representation. Despite of these attractive features and many good empirical results obtained using SVMs, some data modeling participants have begun to realize that the ability for the SVM method to produce sparse

* Address: College of Management, National Changhua University of Education, No. 2, Shi-Da Road, Changhua 500, Taiwan. Tel.: +886 4 7232105 7420; fax: +886 4 7211292; mobile: +886 953092968.

E-mail address: shhuang@cc.ncue.edu.tw

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.01.064

models has perhaps been overstated. For example, it has been shown that the standard SVM technique is not always able to construct parsimonious models in system identification and financial forecasting (e.g., Drezet & Harrison, 1998; Huang & Wu, 2010). On the other hand, large amounts of data from public financial statements can be used for corporate credit rating predictions. The large scale of input data will make SVMs infeasible due to the curse of dimensionality.

The fact that the SVM technique may not guarantee a sufficiently sparse model is the motivation for probability or statistical kernel classifiers. Many probability kernel classifiers have recently received much attention from the machine learning community. Some popular probability kernel classifiers are the Bayes point machine (BPM, Herbrich, Graepel, & Campbell, 2001), and Gaussian process based classifiers. The SVM was proposed as a classifier maximizing the margin, which is the smallest distance between data points and the class boundary (Vapnik, 1995). The BPM is also a kernel classifier whose goal is to approximate Bayes-optimal classification by finding the center of the mass of version space, which is the set of hyperplanes in feature space that separate the data. It was also shown that SVMs can be viewed as a form of Bayes point machine.

In contrast, GPCs are a Bayesian kernel classifier derived from Gaussian process priors over probit or logistic functions (Gibbs & MacKay, 2000; Girolami & Rogers, 2006; Neal, 1997; Williams & Barber, 1998). Gaussian processes (GPs) provide a principled, practical, probabilistic approach to learning in kernel machines. Although these models have a long history in statistics, their potential has only become widely appreciated in the machine learning community during the past decade. With Bayesian inference, GPs provide full predictive distributions and model selection simultaneously. In GPCs, the target values are discrete class labels and so it is not appropriate to model them via a multivariate Gaussian density. In GPCs, one usually uses the Gaussian process as a latent function whose sign determines the class label for binary classification. For multi-class classification, one can use multiple GPs, a multivariate GP, multinomial logistic functions, or multinomial probit functions to determine the output labels.

The introduction of an individual hyperparameter for every class of the latent function is the key feature of the GPC method, and is ultimately responsible for the sparsity properties of the GPC classifier. During the Bayesian optimization process, many of these hyperparameters are driven to large values, so that the corresponding weights of input variables are effectively forced to be zero. Thus the variables are removed from the trained model. Thus GPC classifiers automatically adjust the weights (or importance) between input variables, and do not require any feature selection or dimensionality reduction scheme to resist noisy input data.

This study used various financial variables from public financial statements for the classification. These data contain considerable information regarding enterprize credit risk. This study compared the performance of GPC with convectional multi-class classifiers and kernel classifiers. Empirical results demonstrated that the GPC with Bayesian inference archived lowest prediction error rates. It is believed that irrelevant input variables will deteriorate the performance of conventional kernel classifiers. Consequently, feature selection and dimensionality reduction schemes were employed to extract features for these classifiers and then compared with GPCs. Empirical results indicated that the GPC still performs the best.

The remainder of this paper is organized as follows: Section 2 describes the conventional kernel classifiers. Section 3 introduces the Gaussian process based classifiers. Subsequently, Section 4 describes the study data and discusses the empirical findings. Conclusions are finally given in Section 5.

2. Related works

2.1. Two-class support vector machines

The support vector machines (SVMs) were proposed by Vapnik (1995). 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 (labelled $y \in \{-1, 1\}$) 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

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

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

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,

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

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

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, 1995) can be used as the Kernel function. The Gaussian kernel function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \tag{9}$$

is specified in this study, because Gaussian kernels tend to give good performance under general smoothness assumptions.

2.2. 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