Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 33 (2007) 847–856
Expert Systems with Applications
www.elsevier.com/locate/eswa

Credit scoring with a data mining approach based on support vector machines

Cheng-Lung Huang [a,*], Mu-Chen Chen [b], Chieh-Jen Wang [c]

[a] National Kaohsiung First University of Science and Technology, Department of Information Management, 2, Juoyue Road, Nantz District, Kaohsiung 811, Taiwan
[b] Institute of Traffic and Transportation, National Chiao Tung University, 4F, No. 118, Section 1, Chung Hsiao W. Road, Taipei 10012, Taiwan, ROC
[c] Department of Information Management, Huafan University, 1, Huafan Rd., Shihtin Hsiang, Taipei Hsien 223, Taiwan, ROC

Abstract

The credit card industry has been growing rapidly recently, and thus huge numbers of consumers' credit data are collected by the credit department of the bank. The credit scoring manager often evaluates the consumer's credit with intuitive experience. However, with the support of the credit classification model, the manager can accurately evaluate the applicant's credit score. Support Vector Machine (SVM) classification is currently an active research area and successfully solves classification problems in many domains. This study used three strategies to construct the hybrid SVM-based credit scoring models to evaluate the applicant's credit score from the applicant's input features. Two credit datasets in UCI database are selected as the experimental data to demonstrate the accuracy of the SVM classifier. Compared with neural networks, genetic programming, and decision tree classifiers, the SVM classifier achieved an identical classificatory accuracy with relatively few input features. Additionally, combining genetic algorithms with SVM classifier, the proposed hybrid GA-SVM strategy can simultaneously perform feature selection task and model parameters optimization. Experimental results show that SVM is a promising addition to the existing data mining methods.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Credit scoring; Support vector machine; Genetic programming; Neural networks; Decision tree; Data mining; Classification

1. Introduction

Recently, competition in the consumer credit market has become severe. With the rapid growth in the credit industry, credit scoring models have been extensively used for the credit admission evaluation (Thomas, 2000). In the last two decades, several quantitative methods have been developed for the credit admission decision. The credit scoring models are developed to categorize applicants as either accepted or rejected with respect to the applicants' characteristics such as age, income, and marital condition. Credit officers are faced with the problem of trying to increase credit volume without excessively increasing their exposure to default. Therefore, to screen credit applications, new techniques should be developed to help predict credits more accurately. The benefits of credit scoring involve reducing the credit analysis cost, enabling faster credit decisions, closer monitoring of existing accounts and prioritizing credit collections (Brill, 1998).

In the credit and banking area, a number of articles have been published, which herald the role of automatic approaches in helping creditors and bankers make loans, develop markets, assess creditworthiness and detect fraud. Creditors accept the credit application provided that the applicant is expected to repay the financial obligation. Creditors construct the credit classification rules (credit

* Corresponding author. Tel.: +886 7 6011000x4127; fax: +886 7 6011042.
E-mail address: clhuang@ccms.nkfust.edu.tw (C.-L. Huang).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.07.007

scoring models) based on the data of the previous accepted and rejected applicants. With sizeable loan portfolios, even a slight improvement in credit scoring accuracy can reduce the creditors' risk and translate considerably into future savings.

The modern data mining techniques, which have made a significant contribution to the field of information science (Chen & Liu, 2004), can be adopted to construct the credit scoring models. Practitioners and researchers have developed a variety of traditional statistical models and data mining tools for credit scoring, which involve linear discriminant models (Reichert, Cho, & Wagner, 1983), logistic regression models (Henley, 1995), *k*-nearest neighbor models (Henley & Hand, 1996), decision tree models (Davis, Edelman, & Gammerman, 1992), neural network models (Desai, Crook, & Overstreet, 1996; Malhotra & Malhotra, 2002; West, 2000), and genetic programming models (Ong, Huang, & Tzeng, 2005). From the computational results made by Tam and Kiang (1992), the neural network is most accurate in bank failure prediction, followed by linear discriminant analysis, logistic regression, decision trees, and *k*-nearest neighbor. In comparison with other techniques, they concluded that neural network models are more accurate, adaptive and robust.

Desai et al. (1996) investigated neural networks, linear discriminant analysis and logistic regression for scoring credit decision. They concluded that neural networks outperform linear discriminant analysis in classifying loan applicants into good and bad credits, and logistic regression is comparable to neural networks. West (2000) investigated the credit scoring accuracy of several neural networks. Results were benchmarked against traditional statistical methods such as linear discriminant analysis, logistic regression, *k*-nearest neighbor and decision trees. Malhotra and Malhotra (2002) applied neuro-fuzzy models to analyze consumer loan applications and compared the advantages of neuro-fuzzy systems over traditional statistical techniques in credit-risk evaluation. Hoffmann, Baesens, Martens, Put, and Vanthienen (2002) applied a genetic fuzzy and a neuro-fuzzy classifier for credit scoring. Baesens et al. (2003) benchmarked state-of-the-art classification algorithms for credit scoring.

Recently, researchers have proposed the hybrid data mining approach in the design of an effective credit scoring model. Hsieh (2005) proposed a hybrid system based on clustering and neural network techniques; Lee and Chen (2005) proposed a two-stage hybrid modeling procedure with artificial neural networks and multivariate adaptive regression splines; Lee, Chiu, Lu, and Chen (2002) integrated the backpropagation neural networks with traditional discriminant analysis approach; Chen and Huang (2003) presents a work involving two interesting credit analysis problems and resolves them by applying neural networks and genetic algorithms techniques.

Since even a fraction of improvement in credit scoring accuracy may translate into noteworthy future savings, the major issue of previous studies focused on increasing the accuracy of credit decisions. For conventional statistical classification techniques, an underlying probability model must be assumed in order to calculate the posterior probability upon which the classification decision is made. The more recently developed data mining techniques such as neural networks, genetic programming (GP) and support vector machines (SVM) can perform the classification task without this limitation. Additionally, these artificial intelligence methods also achieved better performance than traditional statistical methods.

Support vector machines (SVM) were first suggested by Vapnik (1995) and have recently been used in a range of problems including pattern recognition (Pontil & Verri, 1998), bioinformatics (Yu, Ostrouchov, Geist, & Samatova, 2003), and text categorization (Joachims, 1998). Huang, Chen, Hsu, Chen, and Wu (2004) obtained prediction accuracy around 80% for both backpropagation neural networks and SVM methods for the United States and Taiwan markets. When using SVM, two problems are confronted: how to choose the optimal input feature subset for SVM and how to set the best kernel parameters. These two problems are crucial because the feature subset choice influences the appropriate kernel parameters and vice versa (Fröhlich & Chapelle, 2003). Therefore, this study proposed hybrid SVM-based approaches to optimize the input feature subset and model parameters.

Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model (Zhang, 2000). With a small feature set, the explanation of rationale for the classification decision can be easier realized. In addition to the feature selection, proper model parameters setting can improve the SVM classification accuracy. The parameters that should be optimized include penalty parameter C and the kernel function parameters such as the gamma (γ) for the radial basis function (RBF) kernel. To design a SVM, one must choose a kernel function, set the kernel parameters and determine a soft margin constant C. The grid algorithm is an alternative to finding the best C and gamma when using the RBF kernel function (Hsu & Lin, 2002). Besides the grid algorithm, other optimization tools such as genetic algorithm, which is adopted in this study, can also be applied to optimize the feature subset and model parameter. To successfully build credit scoring models, this study tried three SVM-based strategies: (1) using grid search to optimize model parameters, (2) using grid search to optimize model parameters and using *F*-score calculation to select input features, and (3) using genetic algorithm to simultaneously optimize model parameters and input features.

This paper is organized as follows. Section 2 describes basic SVM concepts. Section 3 describes three SVM-based strategies used in this research. Section 4 presents the experimental results from using the proposed method to classify two real world datasets. Section 5 gives remarks and provides a conclusion.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات