Expert Systems with Applications 36 (2009) 7562–7566

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Prediction model building with clustering-launched classification and support vector machines in credit scoring

Shu-Ting Luo [a,*], Bor-Wen Cheng [a], Chun-Hung Hsieh [b]

[a] Graduate School of Industry Engineering and Management, National Yunlin University of Science and Technology, 123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan
[b] National Taichung Institute of Technology, 129 Section 3, San-min Road, Taichung 40401, Taiwan

ARTICLE INFO

Keywords:
Credit scoring
Support vector machine
Clustering-launched classification

ABSTRACT

Recently, credit scoring has become a very important task as credit cards are now widely used by customers. A method that can accurately predict credit scoring is greatly needed and good prediction techniques can help to predict credit more accurately. One powerful classifier, the support vector machine (SVM), was successfully applied to a wide range of domains. In recent years, researchers have applied the SVM-based in the prediction of credit scoring, and the results have been shown it to be effective. In this study, two real world credit datasets in the University of California Irvine Machine Learning Repository were selected. SVM and a new classifier, clustering-launched classification (CLC), were employed to predict the accuracy of credit scoring. The advantages of using CLC are that it can classify data efficiently and only need one parameter needs to be decided. In substance, the results show that CLC is better than SVM. Therefore, CLC is an effective tool to predict credit scoring.

1. Introduction

Payments using credit cards have become more and more popular around the world. The credit card scoring manager often evaluates the consumer's credit with intuitive experience. Thus, credit scoring models have been extensively used by financial institutions to determine if loan customers belong to either a good applicant group or a bad applicant group. The advantages of using credit scoring models can be described as the benefit from reducing the cost of credit analysis, enabling faster credit decisions, insuring credit collections, and reducing possible risk (Lee, Chiu, Lu, & Chen, 2002; West, 2000). However, if hundreds of thousands, even millions of credit card or consumer loan applications need to be evaluated, the financial institutions will usually adopt models to assign scores to applicants rather than examining each one in detail. Hence various credit scoring models need to be developed for the purpose of efficient credit approval decisions. Therefore, to screen credit applications, new techniques should be developed to help predict credits more accurately.

In the past, many researchers have developed a variety of traditional statistical methods for credit scoring, with utilization of linear discriminant analysis (LDA) and logistic regression being the two most commonly used statistical techniques in building credit scoring models. However, Karels and Prakash (1987) and Reichert, Cho, and Wagner (1983) pointed that the application of LDA has often been challenged owing to its assumption of the categorical nature of the credit data and the fact that the covariance matrices of the good and bad credit classes are unlikely to be equal.

In addition to the LDA approach, logistic regression is another commonly used alternative to conduct credit scoring tasks. Logistic regression is a model used for prediction of the probability of occurrence of an event. It makes use of several predictor variables that may be either numerical or categories. Basically, the logistic regression model first appeared as the technique in predicting binary outcomes. Logistic regression does not require the multivariate normality assumption, however, the dependent variable accessible to a full linear relationship among independent variables in the exponent of the logistic function. Thomas (2000) and West (2000) indicated that both LDA and logistic regression are intended for the case when the underlying relationship between variables are linear and hence are reported to be lacking in sufficient credit scoring accuracy.

Friedman (1991) reported that multivariate adaptive regression splines (MARS) is another commonly discussed classification technique. MARS is widely accepted by researchers for the following reasons. Firstly, MARS is capable of modeling complex non-linear relationships among variables without strong model assumptions. Secondly, MARS can capture the relative importance of independent variables to the dependent variable when many potential independent variables are considered. Thirdly, the training process of MARS is simple and hence can save lots of model building time,

* Corresponding author.
E-mail address: g9521806@yuntech.edu.tw (S.-T. Luo).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.09.028

especially when the amount of data is huge. Finally, the resulting model of MARS can be more easily interpreted than can other classification techniques. The final fact for MARS is its important managerial and explanatory implications and can help to make appropriate decisions.

Recently, fashionable data mining techniques can be adopted to build credit scoring models. Desai, Crook, and Overstreet (1996) utilized neural networks (NN), LDA and logistic regression to build credit scoring models. The results revealed that NN shows promise if the performance measure is the percentage of bad loans accurately classified. However, LRA is as good as NN if the performance measure is the percentage of good and bad loans accurately classified. West (2000) compared the credit scoring accuracy of five neural network models, and reported that a hybrid architecture of neural network models should be considered for credit scoring applicants. In addition, Kuo, Ho, and Hu (2002) proposed a two-stage mining method, which uses the self-organizing map to determine the number of clusters and then employs the K-means algorithm to classify samples into clusters. The study effected hybrid utility of clustering and neural network techniques in the design of a credit scoring model. Malhotra and Malhotra (2002) compared the performance of artificial neuro-fuzzy inference systems (ANFIS) and multiple discriminant analysis models to screen potential defaulters on consumer loans. The result reported that the ANFIS performs better than the multiple discriminant analysis approach to identify bad credit applications.

Over the last few years, the application of a new classification technique; the support vector machine (SVM), was introduced to deal with the classification problem. Many researchers have applied SVM in credit scoring and financial risk predictions and the results appeared promising (Baesens et al., 2003; Schebesch & Stecking, 2005). In addition, Hung, Chen, and Wang (2007) adopted three strategies to build the hybrid SVM-based credit scoring models to evaluate the applicant's credit score from the applicant's input features.

This study aims to efficiently obtain the discriminant function; the data set is preprocessed by deliberation. Two real-world cases were used below to compare the accuracy rate to two classification models including the SVM and clustering-launched classification (CLC). A new classifier, CLC, can combine clustering with classification. It was developed by the Graduate School of Computer Science and Information Technology, National Taichung Institute of Technology in Taiwan (Chen, Lin, Chiu, Lin, & Chen, 2006). The advantage of the CLC is that it can classify data efficiently and easily. The use of CLC is simple and only one parameter needs to be decided. This study is divided into six parts: Section 2 illustrates SVM and CLC; Section3 introduces of the data analysis in this study; the results are shown in Section 4 and discussed in Section 5. Finally, conclusions are presented in Section 6.

2. Support vector machines and clustering-launched classification

2.1. Support vector machines

The support vector machine (SVM) developed by Vapnik (1995) is an emerging powerful machine learning technique to classify and do regression. SVM is used for a wide variety of problems and it has already been successful in pattern recognition, bio-informatics, and more (Berry & Linoff, 1997; Rüping, 2000). SVM can produce regression or classification functions from a set of training data. The basic procedure for applying SVMs to a classification model can be stated briefly as follows. First, map the input vectors into a feature space, which is possible with a higher dimension. The mapping is either linear or non-linear, depending on the kernel function selected. Then, within the feature space, seek an optimized division, i.e. construct a hyper-plane that separates two or more classes. It, therefore, has the ability to deal with a large number of features. The decision function (or hyper-plane) determined by an SVM is composed of a set of support vectors, which are selected from the training samples. Generally, the main idea of SVM comes from binary classification, namely to find a hyperplane as a segmentation of the two classes to minimize the classification error.

A simple description of the SVM algorithm is provided as follows. Given a training set with input vectors and target labels (x_i, y_i), $i = 1,\ldots,l$, $x \in R^n$, $y \in \{+1, -1\}$, the following conditions:

$$\begin{aligned} x_i + b &\geqslant +1 \quad \text{for} \quad y_i = +1 \\ x_i + b &\leqslant -1 \quad \text{for} \quad y_i = -1 \end{aligned} \tag{1}$$

which is equivalent to

$$y_i(x_i \cdot w + b) - 1 \geqslant 0 \quad \forall i = 1,\ldots,n. \tag{2}$$

This technique looks for a hyperplane $w \cdot x_i + b = 0$ to separate the data from classes +1 and −1 with a maximal margin in the feature space with the margin width between both hyperplanes equal to $\frac{2}{\|w\|^2}$. The maximization of the margin is equivalent to minimize the norm of w. As visualized in Fig. 1, it is a typical two-dimensional case.

In primal weight space, the classifier takes the decision function form Eq. (3). Thus, Cristianini and Taylor (2000) indicated that SVM was trained to solve the following optimization problem:

$$f(x) = \text{sign}(w \cdot x) + b) \tag{3}$$

$$\text{Minimize} \quad \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{N} \xi_i \tag{4}$$

$$\text{Subject to } y_i(w \cdot x_i + b) \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0 \text{ for } i = 1,\ldots,n, \tag{5}$$

where C is a regularization parameter that imposes a trade-off between training error and the variables χ_i are slack variables which are needed in order to allow misclassifications in the set of inequalities (e.g. due to overlapping distributions). The restrictions are imposed to ensure that no training pattern should be within the margins. However, they are relaxed by the slack variables to avoid noisy data. The first part of the objective function tries to maximize the margin between both classes in the feature space, whereas the second part minimizes the misclassification error.

The classifier represented in Eq. (5) is still restricted by the fact that it performs only a linear separation of the data. This can be overcome by mapping the input examples to a high-dimensional space, where they can be efficiently separated by a linear SVM. This mapping is performed with the use of Kernel functions, which allow the access to spaces of high dimensions without knowing the mapping function explicitly, which usually is very complex. The Kernel functions compute dot products between any pair of

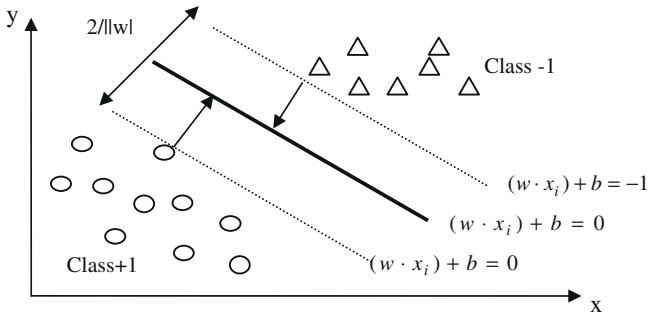

Fig. 1. Illustration of SVM optimization of the margin in the feature space.