Engineering Applications of Artificial Intelligence 26 (2013) 848–862

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

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

Orthogonal support vector machine for credit scoring

Lu Han [a,b,*], Liyan Han [a], Hongwei Zhao [b]

[a] School of Economics and Management, Beihang University, Beijing 100191, China
[b] PBC School of Finance, Tsinghua University, Beijing 100083, China

ARTICLE INFO

Article history:
Received 8 December 2011
Received in revised form
25 September 2012
Accepted 8 October 2012
Available online 17 November 2012

Keywords:
Dimension curse
Orthogonal dimension reduction
Support vector machine
Logistic regression
Principal component analysis
Credit scoring

ABSTRACT

The most commonly used techniques for credit scoring is logistic regression, and more recent research has proposed that the support vector machine is a more effective method. However, both logistic regression and support vector machine suffers from curse of dimension. In this paper, we introduce a new way to address this problem which is defined as orthogonal dimension reduction. We discuss the related properties of this method in detail and test it against other common statistical approaches—principal component analysis and hybridizing logistic regression to better solve and evaluate the data. With experiments on German data set, there is also an interesting phenomenon with respect to the use of support vector machine, which we define as 'Dimensional interference', and discuss in general. Based on the results of cross-validation, it can be found that through the use of logistic regression filtering the dummy variables and orthogonal extracting feature, the support vector machine not only reduces complexity and accelerates convergence, but also achieves better performance.

1. Introduction

Credit risk based on the characteristics of the debtor is often divided into sovereign, corporate, retail, etc. Retail debt is centered on customer credit, which includes short-term and intermediate-term credit to finance the purchase of commodities and services for consumption or to refinance debt incurred for such purposes. Retail credit is characterized by three points: first, large amounts with small scale. At present in China, retail loans can account for a quarter of the total debt, with a speed of growth approaching 10%; second, the potential risk is high but the information is scattered and complicated. In the loan application form there are thousands of variables to describe and, even worse, is that different organizations always use different variables; and third, the efficiency of business processing requires highly developed decision-making techniques as competition is getting more and more intense. These characteristics determine the banks need to implement risk management evaluation methods based on quantitative analysis. A good credit risk evaluation tool can help to grant credit to more creditworthy applicants and thus increases profit. Moreover, it can deny credit for the noncredit worthy applicants and thus decreases losses.

Currently, credit scoring has become the primary method to develop a credit risk assessment tool. It is a method to evaluate the credit risk of loan applicants with their corresponding credit score that is obtained from a credit scoring model (Altman, 1998). A credit score is a number that can represent the creditworthiness of an applicant and it is based on the analysis of an applicant's characteristics from the application file using the credit scoring model. The credit scoring model (Thomas et al., 2002) is developed on the basis of historical data about the applicant's performance on previously made loans with the use of some quantitative techniques, such as statistics analysis, mathematical programming, artificial intelligence and data mining. A well-designed model should have higher classification accuracy to classify the new applicants or existing customers as good or bad and the model is the core of credit scoring.

The most popular methods adopted in credit scoring are statistical methods. The statistical principle discriminating different groups in a population can be traced back to 1936 in Fisher (1936) publication which used a linear model to calculate the distance between two classes as the decision factor. It is known as the Fisher's discrimination model. In 1977, Martin (1977) first introduced the logistic regression method to the bank crisis early warning classification. Martin chose to use data between 1970 and 1976, with 105 bankrupt companies and 2058 non-bankrupt companies in the matching sample, and analyzed the bankruptcy probability interval distribution, with two types of errors and the relationship between the split points, he then found that size, capital structure, and performance were key indexes for the judgment. Martin determined that the accuracy rate of the overall classification could reach 96.12%. Logistic regression analysis had

* Corresponding author. Tel.: +86 1861 166 7963.
E-mail addresses: hanluivy@126.com (L. Han), hanly@buaa.edu.cn (L.Y. Han), hongwei_zhao@yeah.net (H.W. Zhao).

0952-1976/$ - see front matter Crown Copyright © 2012 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2012.10.005

significant improvements over discriminant analysis with respect to the problem of classification. Martin also noted that logistic regression could overcome many of the issues with discriminant analysis, including the assumption that variables must be normally distributed. Wiginton (1980), was one of the first researchers to report credit scoring results with the logistic regression model. Although the result was not very impressive, the model was simple and could be illustrated easily. Then, at that point the logistic regression model had become the main approach for the practical credit scoring application. In 1997, Hand and Henley (1997) summarized statistical methods in credit scoring. These methods are relatively easy to implement and are able to generate straightforward results that can be readily interpreted. Nonetheless, as commonly known, there are also quite a few limitations associated with the applications of these statistical methods. First of all, they have the fatal problem called 'Curse of dimension' which suggests that if there are numerous variables to apply, because of multicollinearity between variables, the results are always erroneous and misleading. Therefore, before applying statistical methods, the process entailed tremendous data preprocessing efforts through variable selection. This strategy usually requires domain expert knowledge and an in-depth understanding of the data. In addition, all the statistical models are based on a hypothesis condition. In a real world application, a hypothesis such as that the dependent variable should follow logic normal distribution and so on, may not hold. Most importantly, based on these algorithms, these statistical models have difficulty in the automation of modeling processes and lack robustness. When environmental or population changes occur, the static models usually fail to adapt and need to be rebuilt again.

In response to the concern for classification accuracy in retail loans applications, researchers discovered the application of the support vector machine (SVM). The support vector machines (SVM) approach was first proposed by Cortes and Vapnik (1995). The main idea of SVM is to minimize the upper bound of the generalization error. SVM usually maps the input variables into a high-dimensional feature space through some nonlinear mapping. In that space, an optimal separating hyper plane, which is one that separates the data with the maximal margin, is constructed by solving a constrained quadratic optimization problem. Suykens et al. (2002) constructed the least squares support vector machine(LS-SVM) and used it for the credit rating of banks and reported the experimental results compared with ordinary least squares (OLS), ordinary logistic regression (OLR) and the multilayer perceptron (MLP). The result showed that the accuracy of the LS-SVM classifier was better than the other three methods. Schebesch and Stecking (2005) used a type of standard SVM proposed by Vanik with a linear and radial basis function (RBF) kernel for dividing credit applicants into subsets of 'typical' and 'critical' patterns which can be used for rejecting applicants. Schebesch and Stecking concluded these types of SVM should be widely used because of their performance. Gestel et al. (2003) discussed a benchmark study of seventeen different classification techniques on eight different real-life credit datasets. They used SVM and LS-SVM with linear and RBF kernels and adopted a grid search mechanism to tune the hyper parameters in their study. The experimental results indicated that six different methods were the best in terms of classification accuracy among the eight datasets — linear regression, logistic regression, linear programming, classification tree, neural networks and SVM. In addition, the experiments showed that the SVM classifiers can overall yield the best performance. Yang (2007) experimented with several kernel learning methods to apply adaptive credit scoring, and found that the results can be very impressive when using the SVM. Nevertheless the existing research findings have all focused on batch learning and the selection of parameters, as seen in the work of Yu et al. (2006,2008) which shows SVM's advantages in solving high dimensional problems. However, there are two obvious drawbacks to SVM (Min and Lee, 2005). One is that when the variables are not 'meaningful' and 'huge', SVM requires a long time to train and the hyper plane is not accurate, which we also define as curse of dimension. The drawback is a fatal flaw, although this method has good robustness and can always achieve higher accuracy, when applied to samples, SVM lacks the capability to explain its results. That is, the results obtained from SVM classifiers are not intuitive to humans and are difficult to illustrate comparing with logistic regression. This is a common problem that all machine learning methods are facing. Though the results with these methods have strong advantages in accuracy, the non-parameter results often lack of statistical theory, and so which cannot be directly corresponding to the realistic economic significance. Just as in regression analysis, regression coefficient directly represents the influence of independent variable acting on dependent variable, but in support vector machine (SVM), the relationship between independent variable and dependent variable cannot be explained directly. So this limits these methods in practical application, and at the same time this also is a cause for over fitting phenomenon.

Dimension curse (Anderson, 1962) can be defined as this phenomenon: as the number of variables increase, more and more variables will have multicollinearity, which can be described as when the correlation coefficient gets large, and is in a high dimensional space, the distribution of the sample points will become sparse. Statistical methods will prove to be erroneous with multicollinearity, and SVM will need a large amount of support vectors to construct hyper plane. Now, to solve the curse of dimensionality, researchers often use two methods to reduce variables. One method is feature selection, another is feature extraction. Feature selection is to select important variables closely related with the target in order to reduce the model's dimensions; feature extraction is to construct new variables which are not linearly dependent through structure transformation. The drawback of feature selection is in reducing information and the advantage is that it is easy to explain. Feature extraction is just the opposite. Many scholars have performed a lot of work to reduce dimensions. Sugiyama (2007) tried feature selection to reduce dimensions in Fisher discriminant analysis. Bellman (1961) is the first to note the curse of dimension in kernel classifiers. He stipulated that owing to the large amounts of data from public financial statements that can be used for bankruptcy predictions, the large scale of input data makes Kernel classifiers infeasible due to the curse of dimensionality. Consequently, one needs to transform the input data space to a suitable low dimensional subspace that optimally represents the data structure. In the studies of Huang (2009), he discussed the use of a nonlinear graph as a type of method for feature selection to reduce dimension. Han and Han (2010) have tried logistic regression to select meaningful variables for neural networks. The other methods regarding dimensionality reduction, linear algorithms such as principal component analysis and linear discriminant analysis, are the two most widely used methods, which can be found in the works of Gutierrez et al. (2010) and Hua et al. (2007).

Just based on the studies above, we want to improve the accuracy of credit scoring through dimension reduction. Our novel contribution is that we give these researchers in the field of application using logistic regression and support vector machine a new way to address dimension curse that we defined as 'Orthogonal dimension reduction' (ORD). Based on the experience of statistics, we compare the traditional way to address dimension curse—hybridizing with logistic regression (HLR) (Fukunaga, 1990) on behalf of feature selection and principal