Expert Systems with Applications 42 (2015) 741–750

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Credit scoring using the clustered support vector machine

Terry Harris *

Credit Research Unit, Department of Management Studies, The University of the West Indies, Cave Hill Campus, P.O. Box 64, Barbados

ARTICLE INFO

Article history:
Available online 6 September 2014

Keywords:
Credit risk
Credit scoring
Clustered support vector machine
Support vector machine

ABSTRACT

This work investigates the practice of credit scoring and introduces the use of the clustered support vector machine (CSVM) for credit scorecard development. This recently designed algorithm addresses some of the limitations noted in the literature that is associated with traditional nonlinear support vector machine (SVM) based methods for classification. Specifically, it is well known that as historical credit scoring datasets get large, these nonlinear approaches while highly accurate become computationally expensive. Accordingly, this study compares the CSVM with other nonlinear SVM based techniques and shows that the CSVM can achieve comparable levels of classification performance while remaining relatively cheap computationally.

1. Introduction

In recent years, credit risk assessment has attracted significant attention from managers at financial institutions around the world. This increased interest has been in no small part caused by the weaknesses of existing risk management techniques that have been revealed by the recent financial crisis and the growing demand for consumer credit (Wang, Yan, & Zhang, 2011). Addressing these concerns, over past decades credit scoring has become increasingly important as financial institutions move away from the traditional manual approaches to this more advanced method, which entails the building of complex statistical models (Huang, Chen, & Wang, 2007; Zhou, Lai, & Yu, 2010).

Many of the statistical methods used to build credit scorecards are based on traditional classification techniques such as logistic regression or discriminant analysis. However, in recent times non-linear approaches,[1] such as the kernel support vector machine, have been applied to credit scoring. These methods have helped to increase the accuracy and reliability of many credit scorecards (Bellotti & Crook, 2009; Yu, 2008). Nevertheless, despite these advances credit analyst at financial institutions are pressed to continually pursue improvements in classifier performance in an attempt to mitigate the credit risk faced by their institutions. However, many of the improvements in classifier performances remain unreported due to the proprietary nature of industry led credit scoring research which attempts to find more efficient and effective algorithms.

In the wider research community, the recent vintages of non-linear classifiers (e.g. the kernel support vector machine) have received a lot of attention and have been critiqued for, *inter alia*, their large time complexities. In fact the best-known time complexity for training a kernel based support vector machine is still quadratic (Bordes, Ertekin, Weston, & Bottou, 2005). As a result, when applied to credit scoring substantial computational resources are consumed when training on reasonably sized real world datasets. Accordingly, efforts to develop and apply new classifiers to credit scoring, which are capable of separating nonlinear data while remaining relatively inexpensive computationally, are well placed.

This paper investigates the suitability for credit scoring of a recently developed support vector machine based algorithm that has been proposed by Gu and Han (2013). Their clustered support vector machine has been shown to offer comparable performance to kernel based approaches while remaining cheap in terms of computational time. Furthermore, this study makes some novel adjustments to their implementation and explores the use of radius basis function (RBF) kernels in addition to the linear kernel posited by Gu and Han.

The remainder of this paper is presented as follows. Section 2 outlines a brief review of the literature concerning the field of credit scoring and sets the stage for the proposed CVSM model for credit scoring that is presented in Section 3. The details of the historic clients' loan dataset and modeling method are highlighted in Section 4. Section 5 presents the study results, and Section 6 discusses the findings, presents conclusions, and outlines possible directions for future research.

* Tel.: +1 (246) 417 4302; fax: +1 (246) 438 9167.
E-mail address: terry.harris@cavehill.uwi.edu

[1] This has been applied because credit-scoring data is often not linearly separable.

http://dx.doi.org/10.1016/j.eswa.2014.08.029
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

2. Background and related works

2.1. Overview

Credit scoring has been critical in permitting the exceptional growth in consumer credit over the last decades. Indeed without accurate, automated credit risk assessment tools, lenders could not have expanded their balance sheets effectively over this time. This section presents a brief review of the relevant literature that has emerged in this space.

2.2. What is credit scoring?

Credit scoring can be viewed as a method of measuring the risk attached to a potential customer, by analyzing their data to determine the likelihood that the prospective borrower will default on a loan (Abdou & Pointon, 2011). According to Eisenbeis (1978), Hand and Jacka (1998), and Hand, Sohn, and Kim (2005) credit scoring can also be described as the statistical technique employed to convert data into rules that can be used to guide credit granting decisions. As a result, it represents a critical process in a firm's credit management toolkit. Durand (1941) posited that the procedure includes collecting, analyzing and classifying different credit elements and variables in order to make credit granting decisions. He noted that to classify a firm's customers, the objective of the credit evaluation process, is to reduce current and expected risk of a customer being "bad" for credit. Thus credit scoring is an important technology for banks and other financial institutions as they seek to minimize risk.

2.3. Related works

Over the years, the demand for consumer credit has increased exponentially. According to Steenackers and Goovaerts (1989), this increase in the demand for credit can be attributable to the increased levels of consumption and the reliance on credit to support this activity. In the United States, this rising level of consumerism followed the introduction of the first modern credit card in 1950s, so that by the 1980s over 55% of American households owned a credit card. Crook, Edelman, and Thomas (2007) posited that by this time, in the US, the total amount of outstanding consumer credit was over $700 billion. Comparatively, at the end of June 2013 this figure had risen to a staggering $2800 billion, a 400% increase (BGFRS, 2013).

Henley (1994) noted that the increasing demand for consumer credit has led to the development of many practical the scoring models, which have adopted a wide range of statistical and non-linear methods. Similarly, Mays (2001) posited that a number of various techniques have been used to build credit scoring applications by credit analyst, researchers, and software developers. These techniques have included; discriminant analysis, linear regression, logistic regression, decision trees, neural networks, support vector machines, k-means, etc.

In recent times, the use of more complex non-linear techniques, such as neural networks, and support vector machines, to build credit scoring applications has seen significant increases in the reported accuracy and performance on benchmarking datasets (Baesens et al., 2003). Irwin, Warwick, and Hunt (1995) and Paliwal and Kumar (2009) both provide evidence that advanced statistical techniques yield superior performance when compared to traditional statistical techniques, such as discriminant analysis, probit analysis and logistic regression. Masters (1995) also provided evidence that the use of sophisticated techniques, such as neural networks, was essential because they had the capability to more accurately model credit scoring data that exhibits interactions and curvature. However, as pointed out by Hand (2006) the increased performance of these more advanced techniques could be illusionary and if real, diminished due to shifts in the class distribution over time. The following sub-sections present a brief discussion concerning some of the classical and advanced statistical models used for credit risk assessment.

2.4. Discriminant analysis

In his seminal paper, Fisher (1936) proposed the use of discriminant analysis to differentiate between two or more classes in a dataset. Since that time, Durand (1941) and Altman (1986) have both applied Fisher's (1936) discriminant analysis to credit scoring. Durant used discriminant analysis to assess the creditworthiness of car loan applicants, while Altman used it to explore corporate bankruptcy proposing his popular Z-scores (Altman, 1968). In works published separately by Desai, Crook, and Overstreet (1996), Hand and Henley (1997), Hand, Oliver, and Lunn (1998), Sarlija, Bensic, and Bohacek (2004), and Abdou and Pointon (2009), they showed that discriminant analysis is indeed a valid technique for credit scoring. Hand and Henley (1997) noted that discriminant analysis, a parametric statistical technique, was well suited to credit scoring because it was designed to classify groups and variables into two or more categories or discriminate between two groups. However, Saunders and Allen (1998) noted that with this type of method certain assumptions about the data must be met. These assumptions include, normality, linearity, homoscedasticity, non-multicollinearity, etc. Falbo (1991) and Sarlija et al. (2004) posited that despite these limitations, over the years this technique has been frequently applied to build credit scoring applications, and it remains one of the most popular approaches taken today when classifying customers as creditworthy or un-creditworthy.

Several authors have criticized the use of discriminant analysis in credit scoring. Eisenbeis (1978) point-out a number of the statistical problems in applying discriminant analysis to credit scorecard development. These problems include the following: group definition, classification error prediction, estimating population priors, and the use of linear functions instead of quadratic functions, to mention a few. Nevertheless, Greene (1998) and Abdou (2009) noted that despite these limits, discriminant analysis is one of the most commonly used techniques in credit scoring.

2.4.1. Linear regression

Another popular classical statistical technique applied to credit scoring is linear regression. This method has developed into an essential component of data analysis in general and is concerned with describing the relationship between a dependent variable and one or more independent variables. Thus, customers' historical payments, guarantees, default rates and other factors can be analyzed using linear regression to set up a score for each factor, and compare it with the bank's cut-off (threshold) score. Hence, only if a new customer's score exceeds the bank's cut-off score will credit be granted (Hand & Jacka, 1998).

In its basic form, linear regression used for credit scoring requires the establishment of a threshold score. This threshold credit score is derived from the relationships between the firm's historic clients' features and their associated weights. As can be seen in the linear equation, $Z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$, where the variable n denotes the number of features collected from past and potential clients. These features are represented by the x's, which are multi-dimensional vectors in $\mathbb{R}^m$, where m denotes the number of clients in the historical clients' database. The θ's represent the weights, and the feature variables and their weights used to calculate a credit score, $Z \in \mathbb{R}$, thus when an applicant scores