European Journal of Operational Research 222 (2012) 168–178

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data

Soner Akkoç *

Department of Banking and Finance, School of Applied Sciences, Dumlupınar University, Central Campus, 43100 Kütahya, Turkey

A R T I C L E I N F O

Article history:
Received 11 March 2011
Accepted 19 April 2012
Available online 27 April 2012

Keywords:
OR in banking
Credit scoring
Neuro fuzzy
ANFIS
Artificial neural networks

A B S T R A C T

The number of Non-Performing Loans has increased in recent years, paralleling the current financial crisis, thus increasing the importance of credit scoring models. This study proposes a three stage hybrid Adaptive Neuro Fuzzy Inference System credit scoring model, which is based on statistical techniques and Neuro Fuzzy. The proposed model's performance was compared with conventional and commonly utilized models. The credit scoring models are tested using a 10-fold cross-validation process with the credit card data of an international bank operating in Turkey. Results demonstrate that the proposed model consistently performs better than the Linear Discriminant Analysis, Logistic Regression Analysis, and Artificial Neural Network (ANN) approaches, in terms of average correct classification rate and estimated misclassification cost. As with ANN, the proposed model has learning ability; unlike ANN, the model does not stay in a black box. In the proposed model, the interpretation of independent variables may provide valuable information for bankers and consumers, especially in the explanation of why credit applications are rejected.

1. Introduction

Recently, the number of Non-Performing Loans (NPLs) has rapidly increased, due to both the effects of global crisis and the appetite for increased risk, in parallel with banks granting more loans without enough evaluation. The global financial crisis has also affected the Turkish banking sector, and as of December 2009, the ratio of NPL to Gross Loans increased to 5.3% from 3.6%. When analyzed the consumer loans of the Turkish banking sector, can be seen in the following ratios; NPL in housing, 2.1%, in consumer loans, 4.4%, in vehicle loans, 10.7% and in credit card loans, 10.8% (BRSA, 2009). This situation has increased the credit risk of the financial sector and has brought about discussions relating to the credit scoring models' efficiency.

Nowadays, credit scoring models using statistical techniques, Operational Research, and Artificial Intelligence (AI) technologies are being developed (Thomas, 2000). Credit scoring models help credit institutions evaluate credit applications with respect to consumers' characteristics such as age, income, and marital status (Chen and Huang, 2003). The objective of credit scoring models is to sort the applications: those that have a high probability of performing financial obligations are assigned to a "good credit" group, and those that have a low probability of performing financial obligations are assigned to a "bad credit" group. Therefore, credit scoring models are basically classification problems (Hand, 1981; Hsieh, 2004; Lee et al., 2006). If these assignments are made accurately, more creditworthy applicants are granted credit, thereby increasing profits; non-creditworthy applicants are denied credit, thus decreasing losses (West, 2000).

In parallel with the growing credit volume of the financial sector, many different credit scoring models have been developed by banks and researchers in order to evaluate credit applications, including Linear Discriminant Analysis (LDA), Logistic Regression Analysis (LRA), Multivariate Adaptive Regression Splines (MARS), Classification and Regression Tree (CART), Artificial Neural Network (ANN), Support Vector Machines (SVM) and Genetic Algorithm (GA) (Abdou et al., 2008; Abdou, 2009; Angelini et al., 2008; Bellotti and Crook, 2009; Chen and Huang, 2003; Chuang and Lin, 2009; Cinko, 2006; Desai et al., 1996; Hsieh and Hung, 2010; Hsieh, 2004, 2005; Huang et al., 2006, 2007; Lee and Chen, 2005; Kim and Sohn, 2010; Lee et al., 2002, 2006; Lee, 2007; Li et al., 2006; Luo et al., 2009; Malhotra and Malhotra, 2003; Nanni and Lumini, 2009; Ong et al., 2005; Paleologo et al., 2010; Sustersic et al., 2009; Tong et al., 2012; Tsai and Wu, 2008; Tsai et al., 2009; West, 2000; West et al., 2005; Yu et al., 2008). Both LDA and LRA have been widely used in credit scoring (Chuang and Lin, 2009; Crook et al., 2007; Desai et al., 1996; Lee et al., 2006; Thomas, 2000; West, 2000). However; LDA makes some assumptions: a linear relationship among the independent variables and a normal

* Tel.: +90 274 2652031x4631; fax: +90 274 2652155.
E-mail addresses: akkocsoner@hotmail.com, sakkoc@dpu.edu.tr

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2012.04.009

distribution of the variables. LDA is criticized because it is unable to provide justification for these assumptions (Thomas, 2000; West, 2000). LRA is used for making prediction on a data set with dichotomous outcomes. In contrast to LDA, LRA does not require the normality assumption. But both models assume that there is a linear relationship among variables, so both of these models may not have enough predictive accuracy in credit scoring (Lancher et al., 1995; Lee and Chen, 2005; Thomas, 2000; West, 2000).

When we look at the last two decades, ANN comes out as an important alternative in financial prediction studies, and draws attention from many researchers with its high prediction accuracy. ANN depends mainly upon transferring the processes of human brain to the computer environment. Unlike statistical techniques, ANN does not require any assumptions, and in research about credit scoring, ANN performs better than both LDA and LRA (Abdou et al., 2008; Chen and Huang, 2003; Desai et al., 1996; Lee and Chen, 2005; Lee et al., 2002; Malhotra and Malhotra, 2003; Sustersic et al., 2009; Tsai et al., 2009; West, 2000). However, ANN is also criticized; (i) for its long training process in developing the optimal network's architecture, (ii) for its inability to identify the relative importance of potential input variables, and (iii) because the model acts as a black box without logic or rule-based explanations for the input–output approximation; in other words, for its inability to explain the underlying principle for the decision to reject applications (Chen and Huang, 2003; Piramuthu, 1999; Trippi and Turban, 1996; West, 2000).

Neuro Fuzzy (NF) systems are relatively new hybrid AI technologies developed by using ANN and Fuzzy Logic (FL) simultaneously; there is little research in applying them to credit scoring models. The purpose of this study is to investigate the ability of the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) credit scoring model, which is based on statistical techniques and NF, by using the credit card data of an international bank operating in Turkey. The performance of the proposed model is also compared with LDA, LRA, and ANN. The main contribution of this paper is that the three stage hybrid ANFIS credit scoring model is a competitive modeling approach for credit card evaluation. In addition to these, we try to interpret with 3D graphs how the proposed model makes credit decisions, which may provide valuable information for bankers and consumers.

The rest of the paper is organized as follows. We will review the literature of credit scoring in Section 2. Section 3 gives a brief outline of LDA, LRA, ANN and NF in building credit scoring model. Section 4 presents the data and the empirical results of the credit scoring models. Finally, concluding remarks are given in Section 5.

2. Literature review

Because of the vast volume of existing literature focusing on credit scoring, we will review only credit scoring studies using commonly used statistical techniques, ANN and NF, in this section. Durand (1941) first actualized credit scoring by LDA, by searching the differences between good and bad credit groups. Since then, statistical techniques, primarily LDA and LRA, have been used in financial prediction studies (Altman, 1968; Martin, 1977; Meyer and Pifer, 1970; Sinkey, 1975; West, 1985).

Since the 1990s, ANN has been also used in modeling credit scoring. Desai et al. (1996) developed credit scoring models with ANN on a data set of 1962 credit consumers, obtained from three different credit unions. Among the subjected models (ANN, LDA and LRA), the performance of ANN was the best, especially in predicting bad credit. Malhotra and Malhotra (2003) got similar results on a set of 1078 data, obtained from twelve different credit unions. West (2000) compared the performance of five different ANN models on credit scoring with LDA, LRA, k nearest neighbor, Kernel Destiny Estimation and CART. West (2000) stated that different ANN models can be used successfully in credit scoring and that LRA can be an alternative to ANN.

Lee et al. (2002) proposed a hybrid credit scoring model which integrates ANN with LDA. The performance of this proposed model has been found more successful than that of LDA, LRA, or ANN separately. In a study using Egypt's personal loan data, Abdou et al. (2008) found that ANN is more successful than LDA, LRA and Probit Analysis. Cinko (2006) obtained successful results using ANN and CART on credit card data. Angelini et al. (2008) achieved a 7% average error rate with ANN on a credit data set consisting of small and medium sized enterprises (SMEs) obtained from an Italian bank. Sustersic et al. (2009) found that credit scoring models developed with ANN are more successful than LRA on consumer credit data, after reducing independent variables with GA and Principal Component Analysis. Chen and Huang (2003) used GA in the process of transferring three rejected credit applications to the conditional acceptance group, and found that the ANN model is more successful than LDA and CART. Lee et al. (2006) have found the models which were developed using CART and MARS on credit card data were more successful than those using LDA, LRA and ANN. Lee and Chen (2005) compared the performance of the LDA, LRA, ANN, MARS, and MARS-ANN models on a data set of mortgage loans obtained from a local Taiwan bank. The best performance was obtained with the ANN model that used the variables found to be more important by MARS. Chuang and Lin (2009) obtained 76%, 76.5%, 77.5%, 79.5%, and 82.5% prediction performance from the LDA, LRA, CART, ANN, and MARS-ANN models respectively, on the German credit data. In the last part of the study, when the data that were transferred to the bad credit group were reevaluated with Case Based Reasoning (CBR), the accuracy rate of the model went up to 86%. Tsai et al. (2009) found that the Data Envelopment Analysis-LDA and ANN models were more successful than LDA and LRA on Taiwanese consumer credit data.

In recent years, ensemble classifiers have been proposed to improve the performance of credit scoring models. The main idea of ensemble classifiers is to combine a number of classifiers into one multiple classifier (Nanni and Lumini, 2006). West et al. (2005) ascertained that ensemble classifier-ANN models reduced the error rate of single classifiers by 3% or 5%. Yu et al. (2008) found that while ANN and SVM are more successful than LRA among single classifiers, the best performance was obtained from ensemble classifiers-ANN. Similarly, in the study done by Nanni and Lumini (2009), ANN was determined to be the best among single classifiers, but the best performance was generally obtained from the random subspace ensemble of classifier with the Levenberg–Marquardt neural net model. In the study of Tsai and Wu (2008), ensemble classifiers-ANN performed better in only one of the three datasets. Hsieh and Hung (2010) developed ensemble classifier credit scoring models after they separated German credit data into good, bad, and borderline groups with Cluster Analysis. Finlay (2011) compared the performance of multiple classifiers and found that Error Trimmed Boosting outperformed all other multiple classifiers on UK credit data.

Most of the reviewed studies focus on whether or not banks should grant credit to consumers who apply to them. On the other hand, the models, which help to make decisions on how to evaluate some demands of the existing customers like raising credit limits, are called behavioral scoring models (Thomas, 2000). Hsieh (2004) has developed behavioral scoring models on the credit card data with self-organizing map ANN. In this research, bank consumers are classified into three major profitable groups, and the results of this study can be used in developing marketing strategies. In another study, Hsieh (2005) drew the conclusion that cluster analysis raised the performance of credit scoring models based on ANN.