Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 36 (2009) 1685–1694

Expert Systems with Applications

www.elsevier.com/locate/eswa

Constructing a reassigning credit scoring model

Chun-Ling Chuang [a], Rong-Ho Lin [b,*]

[a] Department of Information Management, Kainan University, No. 1, Kainan Road, Luzhu, Taoyuan, 33857, Taiwan, ROC
[b] Department of Industrial Engineering & Management, National Taipei University of Technology, No. 1, Section 3, Chung-Hsiao East Road, Taipei 106, Taiwan, ROC

Abstract

Credit scoring model development became a very important issue as the credit industry has many competitions and bad debt problems. Therefore, most credit scoring models have been widely studied in the areas of statistics to improve the accuracy of credit scoring models during the past few years. In order to solve the classification and decrease the Type I error of credit scoring model, this paper presents a reassigning credit scoring model (RCSM) involving two stages. The classification stage is constructing an ANN-based credit scoring model, which classifies applicants with accepted (good) or rejected (bad) credits. The reassign stage is trying to reduce the Type I error by reassigning the rejected good credit applicants to the conditional accepted class by using the CBR-based classification technique. To demonstrate the effectiveness of proposed model, RCSM is performed on a credit card dataset obtained from UCI repository. As the results indicated, the proposed model not only proved more accurate credit scoring than other four common used approaches, but also contributes to increase business revenue by decreasing the Type I and Type II error of scoring system.

Keywords: Credit scoring model; MARS; ANNs; CBR; Type I error

1. Introduction

Credit industry has been rapidly expanded over past few years. Due to the intense competition of credit card issued by banks, more and more people can easily apply a credit card without carefully examining of their credit by the banks. This reckless expansion policy has increased the delinquency rate in the banks. According to the statement issued in 2006 by Financial Supervisory Commission (FSC) of Taiwan, the delinquency rate rose from 1.36% in the second quarter of 2004 to 6.88% in that of 2006.

The objective of credit scoring models is to help the banks to find good credit applications who are likely to observe obligation according to their age, credit limit, income and martial condition. Many different credit scoring models have been developed by the banks and researchers in order to solve the classification problems, such as linear discriminant analysis (LDA), logistic regression (LR), multivariate adaptive regression splines (MARS), classification and regression tree (CART), case based reasoning (CBR), and artificial neural networks (ANNs).

LDA, LR and ANNs are generally used as methods to construct credit scoring models. LDA is the earliest one used for the credit scoring model. However, the utilization of LDA has often been criticized due to the assumptions of linear relationship between input and output variables, which seldom holds, and it is sensitive to deviations from the multivariate normality assumption (West, 2000). In addition to LDA, LR is another common alternative to conduct credit scoring tasks. Basically, the LR model is emerged as the technique in predicting dichotomous outcomes and does not require the multivariate normality assumption. However, both LDA and LR are designed for the relationships between variables are linear, which caused them less accurate in credit scoring.

In handling credit scoring tasks, ANNs became a new alternative and proved more accurate than LDA and LR. However, ANNs is also being criticized for its long training

* Corresponding author. Tel.: +886 28811231x3414; fax: +886 27763964.
E-mail address: rhlin@ntut.edu.tw (R.-H. Lin).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.11.067

process in obtaining the optimal network and not easy to identify the relative importance of potential input variables, in addition, it had certain interpretive difficulties. Therefore, ANNs has limitation of applicability in handling general classification and credit scoring problems (Piramuthu, 1999).

In addition to the above-mentioned techniques, MARS, a common used classification technique, is proved to be a good supporting tool for neural networks as the advantages of MARS can overcome the shortcomings of neural networks (Lee & Chen, 2005). Except for being used in classification models, CBR is also being widely applied in credit scoring models. CBR can use past experiences to find a solution, update database to increase the ability of solving complex and unstructured decision making problems (Shin & Han, 1999; Wheeler & Aitken, 2000).

Many studies have contributed to increasing the accuracy of the classification model with various methods (Ong, Huang, & Tzeng, 2005; Lee & Chen, 2005; Huang, Tzeng, & Ong, 2006). However, most of the previous studies have only concentrated in building a more accurate credit scoring or behavioral scoring model. Even such scoring models are accurate, some misclassification patterns could be emerged, such as Type I error or Type II error. The purpose of this paper is to explore the performance of ANNs-based credit scoring model and reassign the rejected good credit applicants to the preferable accepted class by using the CBR-based classification technique. The increased accepted credit applicants of reassign classification systems in credit card can be significant benefit to decision makers, which contributes to increase business revenue and reduce Type I error of scoring system.

The other part of the paper is organized as follows. Credit scoring literature will be reviewed in Section 2. Section 3 gives a brief outline of commonly used techniques in building credit scoring model. The empirical results of the five built models and the methodology of RCSM are presented in Section 4. Finally, Section 5 addresses the conclusions and summarizes the study results.

2. Literature review

In this section, six common techniques in building credit scoring models will be discussed. The first two models, LDA and LR, are mostly used for classification problems in the area of statistics. The other four models, CART, MARS, ANNs and CBR, are known for their excellent ability of machine learning areas.

LDA was first proposed by Fisher (1936) as a classification technique. It has been reported so far as the most commonly used technique in handling classification problems (Lee, Sung, & Chang, 1999). In the simplest type of LDA, two-group LDA, a linear discriminant function (LDF) that passes through the centroids (geometric centres) of the two groups can be used to discriminate between the two groups. The LDF is represented by

$$\mathrm{LDF} = a + b_1x_1 + b_2x_2 + \cdots + b_px_p \tag{1}$$

where a is a constant, and b_1 to b_p are the regression coefficients for p variables. LDA has been widely applied in a considerable wide range of application areas, such as business investment, bankruptcy prediction, and market segment (Kim, Kim, Kim, Ye, & Lee, 2000; Lee, Jo, & Han, 1997; Trevino & Daniels, 1995). LDA also has been used by Bardos (1998) and Desai, Crook, and Overstreet (1996) in building credit scoring models.

LR is a widely used statistical technique in which the probability of a dichotomous outcome is related to a set of potential independent variables. The LR model does not necessarily require the assumptions of LDA. However, Harrell and Lee (1985) found that LR is as efficient and accurate as LDA even though the assumptions of LDA are satisfied. If there are two groups (e.g. accept and reject), the binary logistic regression (BLR) is used. The probability (p_1) of an object belonging to group 1, and the probability (p_2) of it belonging to group 2, is given by

$$\ln(p_1/p_2) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \tag{2}$$

where (p_1/p_2) is called the odds ratio and $\ln(p_1/p_2)$ the logit transform of p_1; x_n is the nth predictor variable; and b_n is the coefficient of the nth predictor variable. In this equation, the logit transform is being used to relate the probabilities of group membership to a linear function of the predictor variables. LR models have been widely discussed in social research, bankruptcy prediction, and market segmentation (Kay, Warde, & Martens, 2000; Laitinen & Laitinen, 2000; Suh, Noh, & Suh, 1999). LR has also been explored by Laitinen (1999) in building credit scoring models.

CART, a statistical procedure introduced by Breiman, Friedman, Olshen, and Stone (1984), is primarily used as a classification tool, with the objective to classify an object into two or more categories. A CART analysis generally consists of three steps. In a first step an overgrown tree is build, which closely describes the training set. This tree is called the maximal tree and is grown using a binary split-procedure. In a next step the overgrown tree, which shows overfitting, is pruned. During this procedure a series of less complex trees is derived from the maximal tree. In the final step, the tree with the optimal tree size is selected using a cross-validation (CV) procedure

The maximal tree is build using a binary split-procedure, which starts at the tree-root. The tree-root consists of all objects of the training set. At each level, a mother group is considered which is split in two exclusive daughter groups. In the next step, every daughter group becomes a mother group. Every split is described by one value of one descriptor, chosen in such a way that all objects in a daughter group have more similar response variable values. The split for continuous variables is defined by “ $Xi < Aj$” where Xi is the selected explanatory variable and Aj its split value (shown in Fig. 1). CART has been widely discussed in science, weather forecasting, social research, medical