ELSEVIER

Expert Systems with Applications 29 (2005) 41–47

Expert Systems with Applications

www.elsevier.com/locate/eswa

Building credit scoring models using genetic programming

Chorng-Shyong Ong[a], Jih-Jeng Huang[a], Gwo-Hshiung Tzeng[b,c,*]

[a]*Department of Information Management, National Taiwan University, Taipei, Taiwan*
[b]*Institute of Management of Technology, National Chiao Tung University, Ta-Hsuch Rd, Hsunchu 300, Hsinchu 1001, Taiwan*
[c]*College of Management, Kainan University, Taoyuan, Taiwan*

Abstract

Credit scoring models have been widely studied in the areas of statistics, machine learning, and artificial intelligence (AI). Many novel approaches such as artificial neural networks (ANNs), rough sets, or decision trees have been proposed to increase the accuracy of credit scoring models. Since an improvement in accuracy of a fraction of a percent might translate into significant savings, a more sophisticated model should be proposed to significantly improving the accuracy of the credit scoring mode. In this paper, genetic programming (GP) is used to build credit scoring models. Two numerical examples will be employed here to compare the error rate to other credit scoring models including the ANN, decision trees, rough sets, and logistic regression. On the basis of the results, we can conclude that GP can provide better performance than other models.

Keywords: Credit scoring; Artificial neural network (ANN); Decision trees; Genetic programming (GP); Rough sets

1. Introduction

Credit scoring models have been widely 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 decision, insuring credit collections, and diminishing possible risk (Lee, Chiu, Lu, & Chen, 2002; West, 2000). Since an improvement in accuracy of a fraction of a percent might translate into significant savings (West, 2000), a more sophisticated model should be proposed to significantly improve the accuracy of the credit scoring model in this paper.

In order to obtain a satisfied credit scoring model, numerous methods have been proposed. Roughly, these methods can be classified to parametric statistical methods (e.g. discriminant analysis and logistic regression), non-parametric statistical methods (e.g. k nearest neighbor and decision trees), and soft-computing approaches (e.g. artificial neural network (ANN) and rough sets). Recently, ANNs are the most popular tool used for credit scoring and has been reported that its accuracy is superior to that of traditional statistical methods in dealing with credit scoring problems, especially in regards to non-linear patterns (Desai, Crook, & Overstreet, 1996, 1997; Mahlhotra & Malhotra, 2003; Jensen, 1992; Piramuthu, 1999). However, on the other hand, ANN has been criticized for its poor performance when incorporating irrelevant attributes or small data sets (Castillo, Marshall, Green, & Kordon, 2003; Feraud & Cleror, 2002; Nath, Rajagopalan, & Ryker, 1997).

In order to build an effective discriminant function, two issues should be considered. First, the relationships among attributes and classes may be linear or non-linear. Second, the irrelevant attributes should be removed in order to increase the accuracy of the classification model. In this paper, GP is employed to automatically and heuristically determine the adequate discriminant functions and the valid attributes simultaneously. In addition, unlike ANNs which are only suited for large data sets, GP can perform well even in small data sets (Nath et al., 1997).

In order to efficiently obtain the discriminant function, the data set is preprocessed by discretization. Two real-world

* Corresponding author. Address: Institute of Management of Technology, National Chiao Tung University, Ta-Hsuch Rd, Hsunchu 300, Hsinchu 1001, Taiwan. Tel.: +886 3571212157505; fax: 886 35753926.
E-mail address: ghtzeng@cc.nctu.edu.tw (G.-H. Tzeng).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.01.003

cases will be used below to compare the accuracy rate to other classification models including the logistic regression model, ANN, decision trees and rough sets. On the basis of the results, we can conclude that GP can provide better performance than other models.

The rest of this paper is organized as follows. Section 2 describes the models for credit scoring. Discretization and genetic programming are proposed in Section 3. Two real-world examples are used to demonstrate the proposed method in Section 4. Discussions are presented in Section 5 and conclusions are in Section 6.

2. Credit scoring models

In this section, we describe three popular models used in building credit scoring models. The first model is logistic regression, which is mostly used for classification problems in the area of statistics. The second model is ANN, which is known for its excellent ability of learning non-linear relationships in a system. The third model is rough sets, which is one kind of induction based algorithms, and has been widely used in classification problems since 1990s.

2.1. Logistic regression

Logistic regression model is one of the most popular statistical tools for classification problems. Logistic regression model, unlike other statistical tools (e.g. discriminant analysis or ordinary linear regression), can suit various kinds of distribution functions such as Gamble, Poisson, normal, etc. (Press & Wilson, 1978) and is more suitable for the credit scoring problems. In additional, in order to increase its accuracy and flexibility several methods have been proposed to extend the traditional binary logistic regression model, including multinomial logistic regression model (Agresti, 1990; Aldrich & Nelson, 1984; DeMaris, 1992; Knoke & Burke, 1980; Liao, 1994) and logistic regression model for ordered categories (McCullagh, 1980). Therefore, the generalized logistic regression model is the general form of binary logistic regression model and multinomial logistic regression model.

Let a p-dimensional explanatory variables $\boldsymbol{x}' = (x_1, x_2, \ldots, x_p)$ and Y be the response variable with categories $1,2,\ldots,r$. Then the multinomial logistic regression model be given by the equation

$$\text{logistic}(\pi) = \ln\left[\frac{P(Y=j|\boldsymbol{x})}{P(Y=k|\boldsymbol{x})}\right] = \boldsymbol{x}'\beta_j, \quad 0 \le j \le r, \quad j \ne k \tag{1}$$

where β_j is a $(p+1)$ vector of the regression coefficients for the jth variable.

Let the last response level be the reference level and then the response probabilities $\pi_1, \pi_2, \ldots, \pi_r$ can be calculated by the equations

$$\pi_r \equiv P(Y=r|\boldsymbol{x}) = \frac{e^{\boldsymbol{x}'\beta_r}}{\sum_{l=1}^{r} e^{\boldsymbol{x}'\beta_l}} = \frac{e^{\boldsymbol{x}'\beta_r}}{e^{\boldsymbol{x}'\beta_r} + \sum_{l=1}^{r-1} e^{\boldsymbol{x}'\beta_l}} = \frac{1}{1 + \sum_{l=1}^{r-1} e^{\boldsymbol{x}'\beta}} \tag{2}$$

$$\pi_j \equiv P(Y=j|\boldsymbol{x}) = \pi_r e^{\boldsymbol{x}'\beta_j}, \quad 1 \le j \le r-1 \tag{3}$$

where l is a response level, and

$$l = l(\beta_j, 1 \le j \le r, j \ne k) = \sum_{i=1}^{n} \ln(P(Y=y_i|\boldsymbol{x}_i)), \quad l \in [1,2,\ldots,r] \tag{4}$$

is the ln likelihood for the multinomial logistic regression model and $\{(y_i, \boldsymbol{x_i}), 1 \le i \le n\}$ denotes the sample of n objects. When the category is equal to two, the multinomial logistic regression model reduces to a binary logistic regression model.

Although logistic regression model can perform well in many applications, when the relationships of the system are non-linear, the accuracy of logistic regression decreases and ANN has been proposed to deal with this problem.

2.2. Artificial neural network

Artificial neural networks were developed to mimic the neurophysiology of the human brain to be a type of flexible non-linear regression, discriminant, and clustering models The architecture of ANN can usually be represented as a three-layer system, named input, hidden, and output layers. The input layer first processes the input features to the hidden layer. The hidden layer then calculates the adequate weights by using the transfer function such as hyperbolic tangent, softmax, or logistic function before sending to the output layer.

Combining many computing neurons into a highly interconnected system, we can detect the complex non-linear relationship in the data. The simple three-layer perceptron, which is most used in credit scoring problems, can be depicted as shown in Fig. 1.

Recently, ANN has been widely used in credit scoring problems, and it has been reported that its accuracy is superior to the traditional statistical methods such as discriminant analysis and logistic regression (Desai et al., 1996, 1997; Jensen, 1992; Mahlhotra & Malhotra, 2003; Piramuthu, 1999). However, as mentioned previously, ANN has been criticized for its poor performance when existing irrelevant attributes or small data sets. Although many methods have been proposed to deal with the problem of variable selection (Feraud & Cleror, 2002; Nath et al., 1997), it is time waste and makes the model more complicated. In addition, other scholars are criticized