Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 2473–2480
Expert Systems with Applications
www.elsevier.com/locate/eswa

The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients

I-Cheng Yeh [a,*], Che-hui Lien [b]

[a] Department of Information Management, Chung-Hua University, Hsin Chu 30067, Taiwan, ROC
[b] Department of Management, Thompson Rivers University, Kamloops, BC, Canada

Abstract

This research aimed at the case of customers' default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown, this study presented the novel "Sorting Smoothing Method" to estimate the real probability of default. With the real probability of default as the response variable (Y), and the predictive probability of default as the independent variable (X), the simple linear regression result ($Y = A + BX$) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability of default.

Keywords: Banking; Neural network; Probability; Data mining

1. Introduction

In recent years, the credit card issuers in Taiwan faced the cash and credit card debt crisis and the delinquency is expected to peak in the third quarter of 2006 (Chou, 2006). In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit card for consumption and accumulated heavy credit and cash–card debts. The crisis caused the blow to consumer finance confidence and it is a big challenge for both banks and cardholders.

In a well-developed financial system, crisis management is on the downstream and risk prediction is on the upstream. The major purpose of risk prediction is to use financial information, such as business financial statement, customer transaction and repayment records, etc., to predict business performance or individual customers' credit risk and to reduce the damage and uncertainty.

Many statistical methods, including discriminant analysis, logistic regression, Bayes classifier, and nearest neighbor, have been used to develop models of risk prediction (Hand & Henley, 1997). With the evolution of artificial intelligence and machine learning, artificial neural networks and classification trees were also employed to forecast credit risk (Koh & Chan, 2002; Thomas, 2000). Credit risk here means the probability of a delay in the repayment of the credit granted (Paolo, 2001).

From the perspective of risk control, estimating the probability of default will be more meaningful than classifying customers into the binary results – risky and non-risky. Therefore, whether or not the estimated probability of default produced from data mining methods can represent the "real" probability of default is an important problem. To forecast probability of default is a challenge facing practitioners and researchers, and it needs more study (Baesens, Setiono, Mues, & Vanthienen, 2003; Baesens et al., 2003; Desai, Crook, & Overstreet, 1996; Hand &

* Corresponding author.
E-mail address: icyeh@chu.edu.tw (I.-C. Yeh).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.12.020

Henley, 1997; Jagielska & Jaworski, 1996; Lee, Chiu, Lu, & Chen, 2002; Rosenberg & Gleit, 1994; Thomas, 2000).

Because the real probability of default is unknown, this study proposed the novel "Sorting Smoothing Method" to deduce the real default probability and offered the solutions to the following two questions:

(1) Is there any difference of classification accuracy among the six data mining techniques?
(2) Could the estimated probability of default produced from data mining methods represent the real probability of default?

In the next section, we review the six data mining techniques (discriminant analysis, logistic regression, Bayes classifier, nearest neighbor, artificial neural networks, and classification trees) and their applications on credit scoring. Then, using the real cardholders' credit risk data in Taiwan, we compare the classification accuracy among them. Section 4 is dedicated to the predictive performance of probability of default among them. Finally, Section 5 contains some concluding remarks.

2. Literature review

2.1. Data mining techniques

In the era of information explosion, individual companies will produce and collect huge volume of data everyday. Discovering useful knowledge from the database and transforming information into actionable results is a major challenge facing companies. Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules (Berry & Linoff, 2000). Right now, data mining is an indispensable tool in decision support system and plays a key role in market segmentation, customer services, fraud detection, credit and behavior scoring, and benchmarking (Paolo, 2001; Thomas, 2000).

The pros and cons of the six data mining techniques employed in our study are reviewed as follows (Han & Kamber, 2001; Hand, Mannila, & Smyth, 2001; Paolo, 2003; Witten & Frank, 1999).

2.1.1. K-nearest neighbor classifiers (KNN)

K-nearest neighbor (KNN) classifiers are based on learning by analogy. When given an unknown sample, a KNN classifier searches the pattern space for the KNN that are closest to the unknown sample. Closeness is defined in terms of distance. The unknown sample is assigned the most common class among its KNN. The major advantage of this approach is that it is not required to establish predictive model before classification. The disadvantages are that KNN does not produce a simple classification probability formula and its predictive accuracy is highly affected by the measure of distance and the cardinality k of the neighborhood.

2.1.2. Logistic regression (LR)

Logistic regression can be considered a special case of linear regression models. However, the binary response variable violates normality assumptions of general regression models. A logistic regression model specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables. The major advantage of this approach is that it can produce a simple probabilistic formula of classification. The weaknesses are that LR cannot properly deal with the problems of non-linear and interactive effects of explanatory variables.

2.1.3. Discriminant analysis (DA)

Discriminant analysis, also known as Fisher's rule, is another technique applied to the binary result of response variable. DA is an alternative to logistic regression and is based on the assumptions that, for each given class of response variable, the explanatory variables are distributed as a multivariate normal distribution with a common variance–covariance matrix. The objective of Fisher's rule is to maximize the distance between different groups and to minimize the distance within each group. The pros and cons of DA are similar to those of LR.

2.1.4. Naïve Bayesian classifier (NB)

The naïve Bayesian classifier is based on Bayes theory and assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. Bayesian classifiers are useful in that they provide a theoretical justification for other classifiers that do not explicitly use Bayes theorem. The major weakness of NB is that the predictive accuracy is highly correlated with the assumption of class conditional independence. This assumption simplifies computation. In practice, however, dependences can exist between variables.

2.1.5. Artificial neural networks (ANNs)

Artificial neural networks use non-linear mathematical equations to successively develop meaningful relationships between input and output variables through a learning process. We applied back propagation networks to classify data. A back propagation neural network uses a feed-forward topology and supervised learning. The structure of back propagation networks is typically composed of an input layer, one or more hidden layers, and an output layer, each consisting of several neurons. ANNs can easily handle the non-linear and interactive effects of explanatory variables. The major drawback of ANNs is – they cannot result in a simple probabilistic formula of classification.

2.1.6. Classification trees (CTs)

In a classification tree structure, each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes. The top-most node in a tree is the root node. CTs are applied