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

when the response variable is qualitative or quantitative discrete. Classification trees perform a classification of the observations on the basis of all explanatory variables and supervised by the presence of the response variable. The segmentation process is typically carried out using only one explanatory variable at a time. CTs are based on minimizing impurity, which refers to a measure of variability of the response values of the observations. CTs can result in simple classification rules and can handle the nonlinear and interactive effects of explanatory variables. But their sequential nature and algorithmic complexity can make them depends on the observed data, and even a small change might alter the structure of the tree. It is difficult to take a tree structure designed for one context and generalize it for other contexts.

2.2. The applications of data mining techniques on credit scoring

Rosenberg and Gleit (1994) pointed out that many static and dynamic models have been used to assist decision-making in the area of consumer and commercial credit. The decisions of interest include whether to extend credit, how much credit to extend, when collections on delinquent accounts should be initiated, and what action should be taken. They surveyed the use of discriminant analysis, classification trees, and expert systems for static decisions, and dynamic programming, linear programming, and Markov chains for dynamic decision models.

Hand and Henley (1997) argued that credit scoring is the term used to describe formal statistical methods which are used for classifying applicants for credit into "good" and "bad" risk classes. Such methods have become increasingly important with the dramatic growth in consumer credit in recent years. A wide range of statistical methods has been applied, though the literature available to the public is limited for reasons of commercial confidentiality.

Paolo (2001) showed that Bayesian methods, coupled with Markov Chain Monte Carlo computational techniques, could be successfully employed in the analysis of highly dimensional complex dataset, such as those in credit scoring and benchmarking. Paolo employs conditional independence graphs to localize model specification and inferences, thus allowing a considerable gain in flexibility of modeling and efficiency of the computations.

Lee et al. (2002) explored the performance of credit scoring by integrating the backpropagation neural networks with the traditional discriminant analysis approach. The proposed hybrid approach converges much faster than the conventional neural networks model. Moreover, the credit scoring accuracy increases in terms of the proposed methodology and the hybrid approach outperforms traditional dscriminant analysis and logistic regression.

Baesens et al. (2003) found that, based on eight real-life credit scoring data sets, both the LS-SVM and neural network classifiers yield a very good performance, but also simple classifiers such as logistic regression and linear discriminant analysis perform very well for credit scoring.

3. Classification accuracy among data mining techniques

3.1. Description of the data

Our study took payment data in October, 2005, from an important bank (a cash and credit card issuer) in Taiwan and the targets were credit card holders of the bank. Among the total 25,000 observations, 5529 observations (22.12%) are the cardholders with default payment. This research employed a binary variable – default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature (Lee, Yen, Lin, Tseng, & Ma, 2004; Steenackers & Goovaerts, 1989; Updegrave, 1987) and used the following 23 variables as explanatory variables:

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6–X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005;...;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: −1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; ...; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X12–X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005;...;X17 = amount of bill statement in April, 2005.
- X18–X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005;...;X23 = amount paid in April, 2005.

The data was randomly divided into two groups, one for model training and the other to validate the model. Reviewing the literature (Jain, Duin, & Mao, 2000; Nelson, Runger, & Si, 2003) reveals that error rates were often used as the measurement of classification accuracy of models. However, most records in the data set of credit card customers are non-risky (87.88%); therefore, the error rate is insensitive to classification accuracy of models. For the binary classification problem, area ratio in the lift chart can offer better solution for comparing the performance of different models than the one did by the error rate

(Han & Kamber, 2001; Hand et al.; 2001; Witten & Frank, 1999). Therefore, our study employed area ratio, instead of the error rate, to examine the classification accuracy among the six data mining techniques. In the lift chart, the horizontal axis represents the number of total data. The vertical axis shows the cumulative number of target data. There are three curves (see Fig. 1) – model curve, theoretically best curve, and diagonal baseline curve, in the lift chart (Berry & Linoff, 2000). The greater the area between the model curve and the baseline curve, the better the model. Area ratio is defined as

$$\text{Area ratio} = \frac{\text{area between model curve and baseline curve}}{\text{area between theoretically best curve and baseline curve}} \quad (1)$$

3.2. Results

The lift charts of the six data mining techniques are shown below (Figs. 2–7). From Table 1, in the training data, based on error rates, K-nearest neighbor classifiers and classification trees have the lowest error rate (=0.18). For the area ratio, K-nearest neighbor classifiers, with the highest area ratio (=0.68), performs better than other methods. But in the validation data, artificial neural networks achieve the best performance with the highest area ratio (=0.54) and the relatively low error rate (=0.17). Because the validation data is the effective data set used to measure the generalization classification accuracy of models, therefore, we can conclude that artificial neural networks is the best model among the six methods.

In this credit card scoring case, most cardholders are classified into creditable customers (87.88%). Therefore, error rates are not the appropriate criteria to evaluate the performance of models. For example, in validation data,

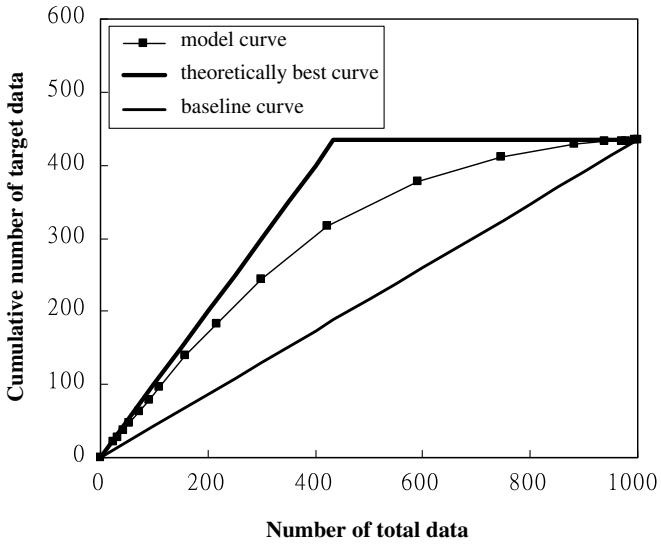

Fig. 1. Lift chart.

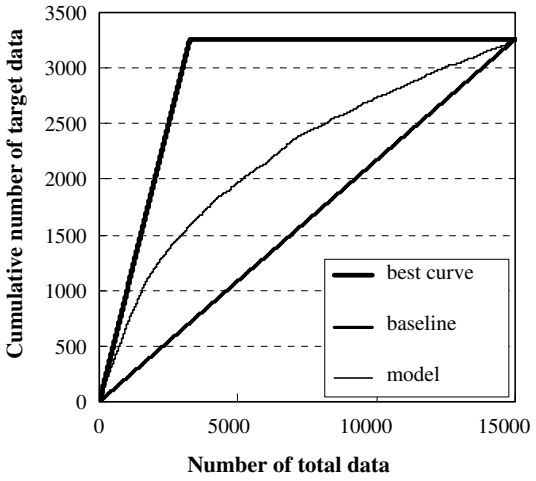

Fig. 2. Lift chart of K-nearest neighbor classifiers.

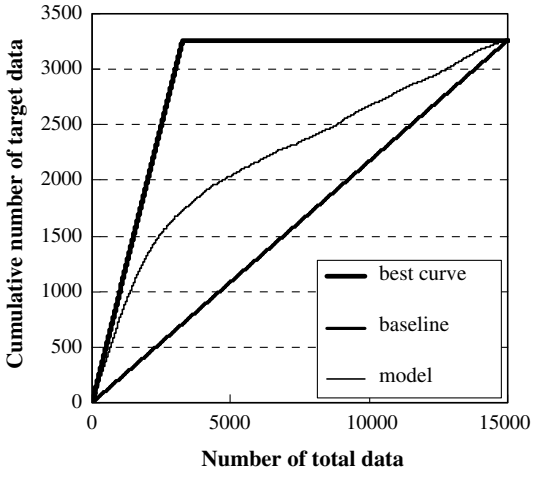

Fig. 3. Lift chart of logistic regression.

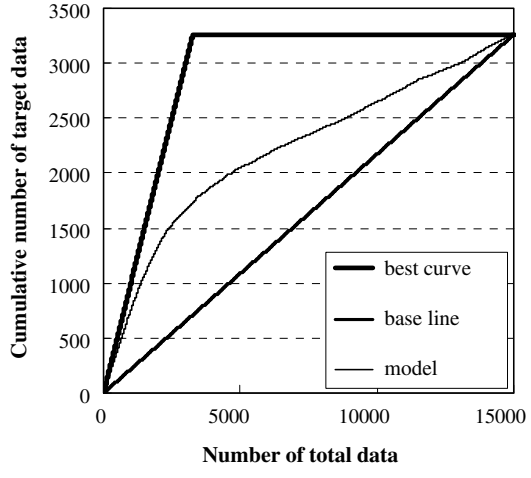

Fig. 4. Lift chart of discriminant analysis.

K-nearest neighbor classifiers have the lowest error rate, but based on area ratio, they do not perform better than naïve Bayesian classifier, artificial neural networks, and

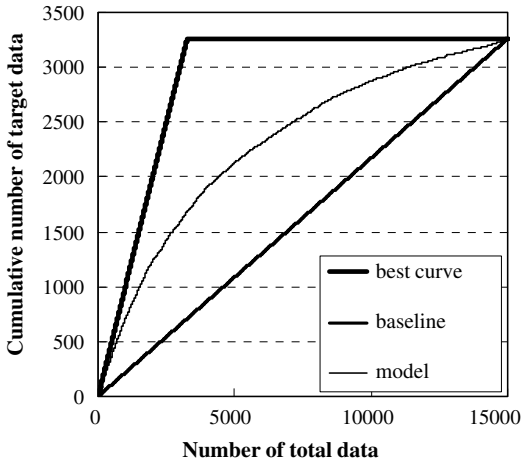

Fig. 5. Lift chart of naïve Bayesian classifier.

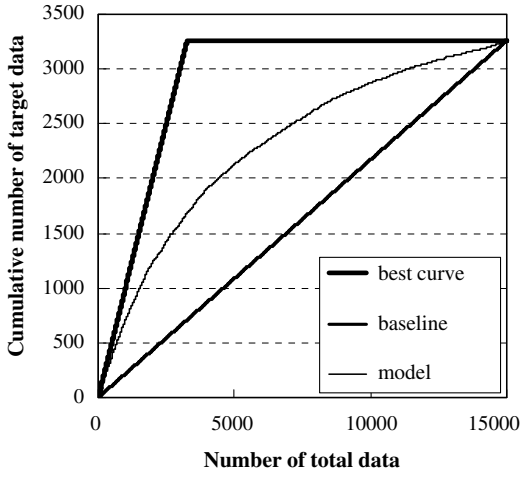

Fig. 6. Lift chart of artificial neural networks.

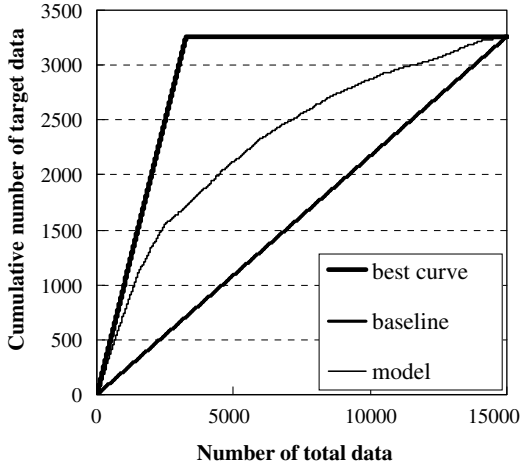

Fig. 7. Lift chart of classification trees.

classification trees. When using the area ratio in the validation data, the classification result shows the performance of the six data mining methods is ranked as: artificial neural networks, classification trees, naïve Bayesian classifier, K-nearest neighbor classifiers, logistic regression, and discriminant analysis, respectively.

Table 1
Classification accuracy

Method	Error rate		Area ratio	
	Training	Validation	Training	Validation
K-nearest neighbor	0.18	0.16	0.68	0.45
Logistic regression	0.20	0.18	0.41	0.44
Discriminant analysis	0.29	0.26	0.40	0.43
Naïve Bayesian	0.21	0.21	0.47	0.53
Neural networks	0.19	0.17	0.55	0.54
Classification trees	0.18	0.17	0.48	0.536

4. Predictive accuracy of probability of default

To estimate the real probability of default, the novel approach, called Sorting Smoothing Method (SSM), was proposed in this study. Firstly, according to the predictive probability, order the validation data from the minimum to the maximum. Secondly, use the SSM to estimate the real probability of default as follows:

$$P_i = \frac{Y_{i-n} + Y_{i-n+1} + \cdots + Y_{i-1} + Y_i + Y_{i+1} + \cdots + Y_{i+n-1} + Y_{i+n}}{2n+1} \quad (2)$$

where P_i = estimated real probability of default in the ith order of validation data; Y_i = binary variable with real default risk in the ith order of validation data; $Y_i = 1$ stands for "happened"; $Y_i = 0$ stands for "not happened"; n = numbers of data for smoothing.

With the estimated real probability of default (seen as real default probability), the following procedure could

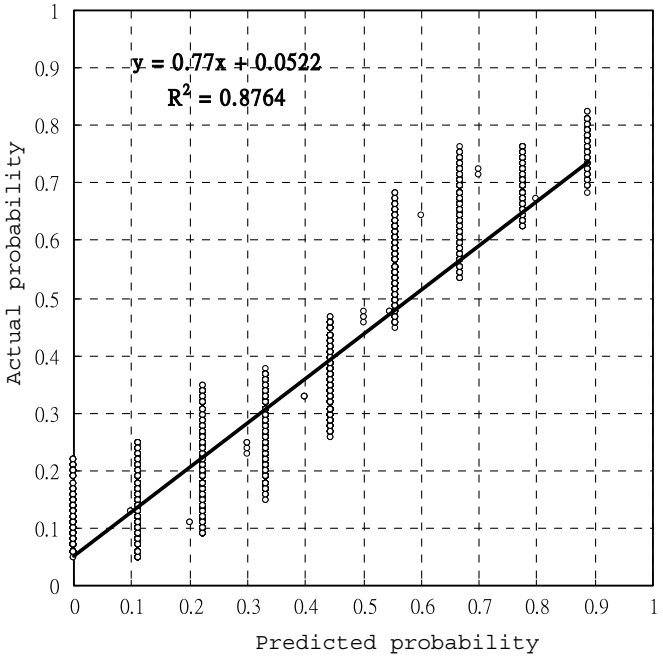

Fig. 8. Scatter plot diagram of K-nearest neighbor classifiers.

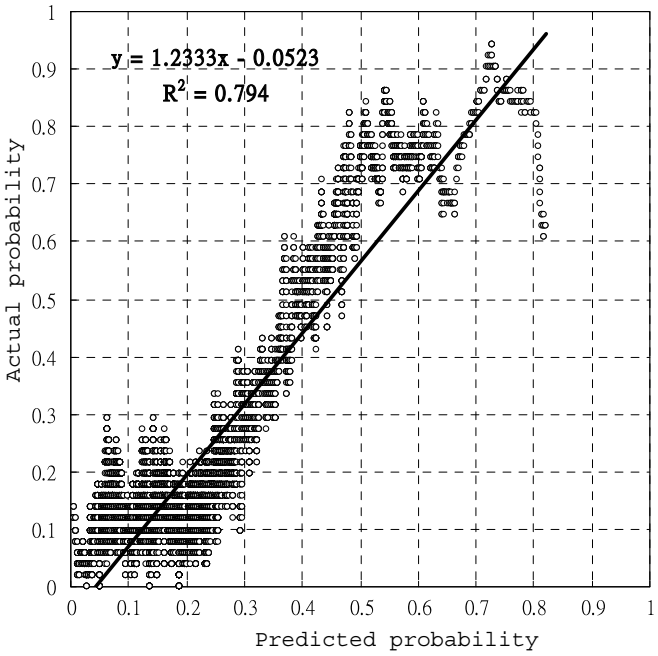

Fig. 9. Scatter plot diagram of logistic regression.

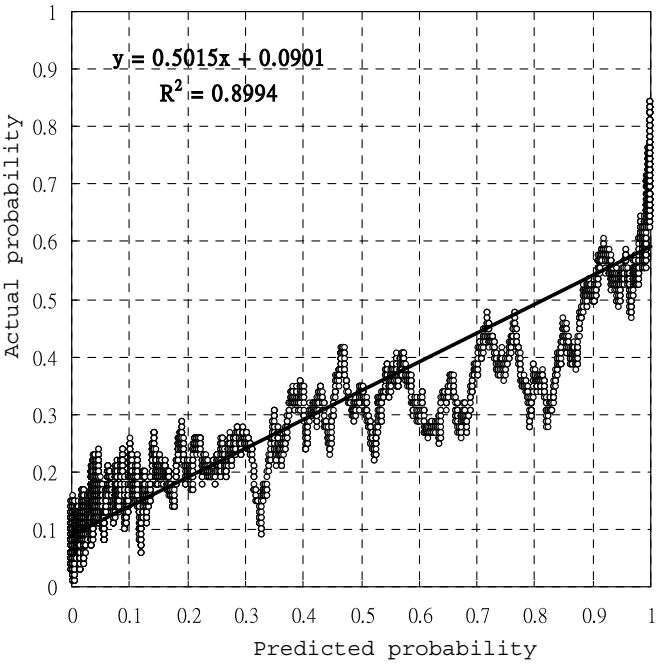

Fig. 11. Scatter plot diagram of naïve Bayesian classifier.

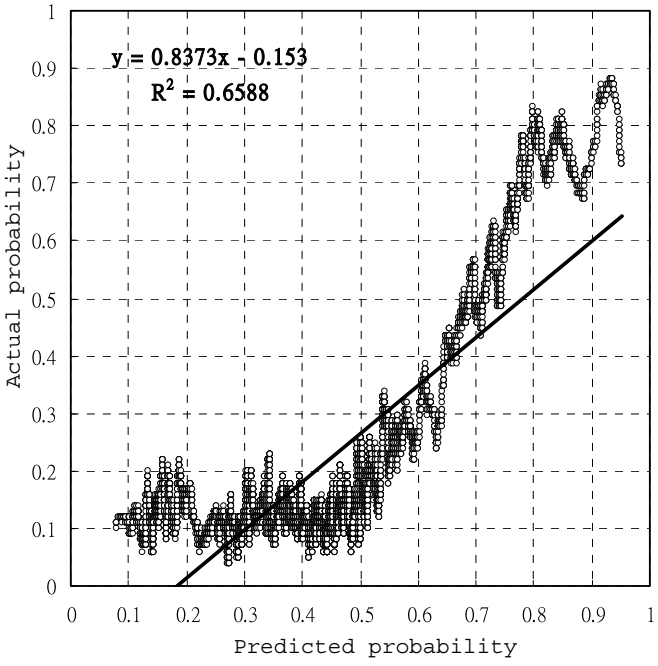

Fig. 10. Scatter plot diagram of discriminant analysis.

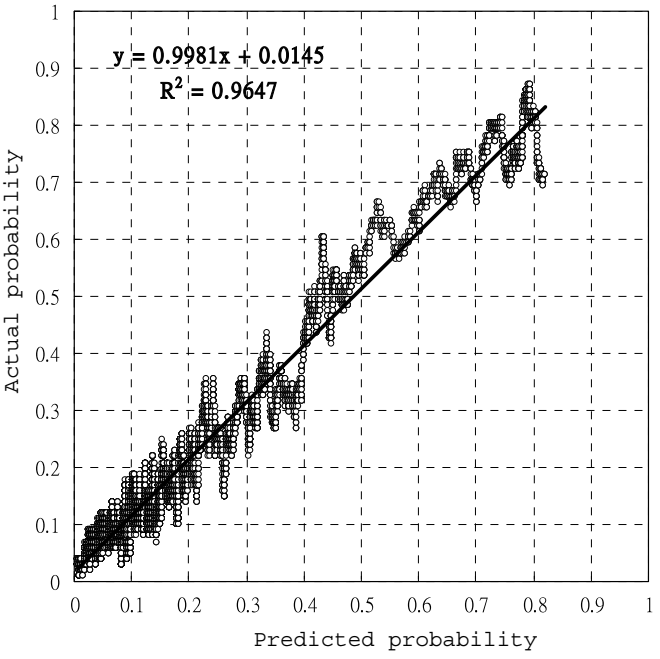

Fig. 12. Scatter plot diagram of artificial neural networks.

be used to explore whether or not the predictive default probability represents the real probability of default:

(1) Scatter plot diagram: the horizontal axis represents the predictive default probability; the vertical axis stands for the estimated real probability of default.
(2) Linear regression: the linear regression line ($Y = A + BX$) is produced from the scatter plot diagram, and the coefficient of determination (R^2) is calculated. If the R^2 is close to one, intercept (A) to zero, and regression coefficient (B) to one, then we can conclude that the predictive default probability produced from data mining methods can represents the real default probability.

In this study, the $n = 50$ was chosen and SSM was employed to estimate the real default probability. The scatter plot diagram, the regression line, and R^2, produced

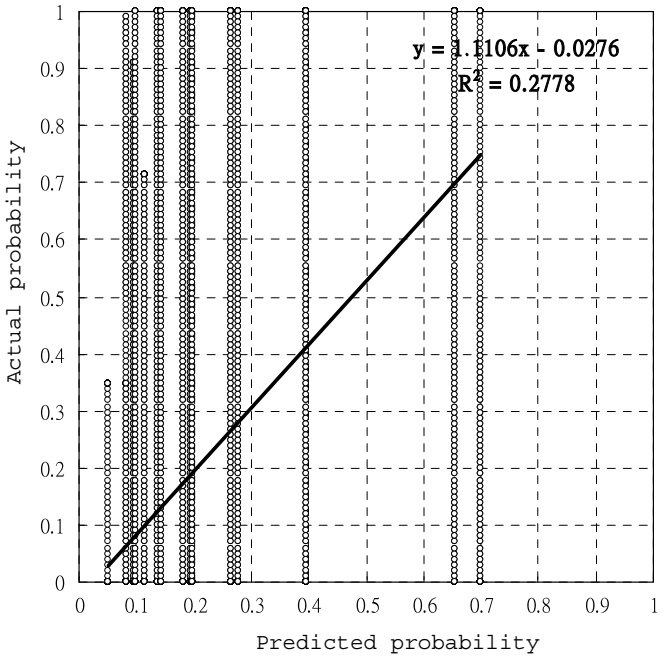

Fig. 13. Scatter plot diagram of classification trees.

Table 2
Summary of linear regression between real probability and predictive probability of default

Method	Regression Coefficient	Regression Intercept	Regression R^2
K-nearest neighbor	0.770	0.0522	0.876
Logistic regression	1.233	−0.0523	0.794
Discriminant Analysis	0.837	−0.1530	0.659
Naïve Bayesian	0.502	0.0901	0.899
Neural networks	0.998	0.0145	0.965
Classification trees	1.111	−0.0276	0.278

from the six data mining techniques are shown from Figs. 8 to 13 and summarized in Table 2. From the result of R^2, the predictive default probability produced from artificial neural networks has the highest explanatory ability ($R^2 = 0.9647$) for real probability of default. In linear regression, only artificial neural networks bear the value of regression coefficient (B) close to one, and the value of intercept (A) close to zero.

5. Conclusion

This paper examines the six major classification techniques in data mining and compares the performance of classification and predictive accuracy among them. The novel Sorting Smoothing Method, for the first time, is presented to estimate the real probability of default.

In the classification accuracy among the six data mining techniques, the results show that there are little differences in error rates among the six methods. However, there are relatively big differences in area ratio among the six techniques. Obviously, area ratio is more sensitive and is an appropriate criterion to measure the classification accuracy of models. Artificial neural networks perform classification more accurately than the other five methods.

In the predictive accuracy of probability of default, artificial neural networks also show the best performance based on R^2 (0.9647, close to 1), regression intercept (0.0145, close to 0), and regression coefficient (0.9971, close to 1). The predictive default probability produced by ANN is the only one that could be used to represent real probability of default. From the perspective of risk control, estimating the probability of default is more meaningful than classifying clients into binary results – risky and non-risky. Therefore, artificial neural networks should be employed to score clients instead of other data mining techniques, such as logistic regression.

References

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science, 49*(3), 312–329.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

Berry, M., & Linoff, G. (2000). *Mastering data mining: The art and science of customer relationship management*. New York: John Wiley & Sons, Inc.

Chou, M. (2006). Cash and credit card crisis in Taiwan. *Business Weekly*, 24–27.

Desai, V. S., Crook, J. N., & Overstreet, G. A. A. (1996). Comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*(1), 24–37.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. San Fransisco: Morgan Kaufmann.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society, Series A – Statistics in Society, 160*(3), 523–541.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Data mining: Practical machine learning tools and techniques*. Cambridge: MIT Press.

Jagielska, I., & Jaworski, J. (1996). Neural network for predicting the performance of credit card accounts. *Computational Economics, 9*(1), 77–82.

Jain, A., Duin, P., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 4–37.

Koh, H. C., & Chan, K. L. G. (2002). Data mining and customer relationship marketing in the banking industry. *Singapore Management Review, 24*(2), 1–27.

Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), 245–254.

Lee, Y. S., Yen, S. J., Lin, C. H., Tseng, Y. N., Ma, L. Y. (2004). A data mining approach to constructing probability of default scoring model. In *Proceedings of 10th conference on information management and implementation* (pp. 1799–1813).

Nelson, B., Runger, G., & Si, J. (2003). An error rate comparison of classification methods with continuous explanatory variables. *IIE Transactions, 35*, 557–566.

Paolo, G. (2001). Bayesian data mining, with application to benchmarking and credit scoring. *Applied Stochastic Models in Business and Society, 17*, 69–81.

Paolo, G. (2003). *Applied data mining: Statistical methods for business and industry*. New York: John Wiley & Sons, Inc.

Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: A survey. *Operations Research, 42*(4), 589–613.

Steenackers, A., & Goovaerts, M. J. (1989). A credit scoring model for personal loans. *Insurance Mathematics Economic, 2*, 31–34.

Thomas, L. C. (2000). A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*, 149–172.

Updegrave, W. L. (1987). How lender size you up. *Money*, 23–40.

Witten, I. H., & Frank, E. (1999). *Data mining: Practical machine learning tools and techniques with java implementations*. San Fransisco: Morgan Kaufman.