DATA 606 project: **Predicting Default of Credit Card Clients**

Paper 2: Literature Review

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 statements, customer transactions, and repayment records, etc., to predict business performance or individual customers’ credit risk and to reduce the damage and uncertainty. From the perspective of risk control, estimating the probability of default will be more meaningful than classifying customers into the binary results – risky and nonrisky.

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 the 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 & 0957-4174/$ - see front matter 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.eswa.2007.12.020 \* Corresponding author. E-mail address: icyeh@chu.edu.tw (I.-C. Yeh). www.elsevier.com/locate/eswa Available online at www.sciencedirect.com Expert Systems with Applications 36 (2009) 2473–2480 Expert Systems with Applications Henley, 1997; Jagielska & Jaworski, 1996; Lee, Chiu, Lu, & Chen, 2002; Rosenberg & Gleit, 1994; Thomas, 2000).

In the era of information explosion, individual companies will produce and collect a huge volume of data every day. 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 to discover meaningful patterns and rules (Berry & Linoff, 2000). Right now, data mining is an indispensable tool in decision support systems 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).

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).

There are two major findings from this literature. 1. Is there any difference in 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?

**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.