Full Length Research Paper

A comparative study of data mining techniques in predicting consumers’ credit card risk in banks

Ling Kock Sheng¹* and Teh Ying Wah²

¹Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia.
²Department of Information Science, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia.

Accepted 21 June, 2011

This paper investigates the use of batch and incremental classifiers such as logistic regression, neural networks, C5, naïve bayes updateable, IBk (instance-based learner, k nearest neighbour) and raced incremental logit boost to obtain the best classifier to be used for improving the predictive accuracy of consumers’ credit card risk of a bank in Malaysia. Prior to generating all the models for comparison, the initial set of data is also loaded into an ETL (extraction, transformation, loading) system developed to perform feature selection or attribute relevancy analysis using ID3 algorithm, compiling a subset of data with the highest information gain and gain ratio. An extended test is performed to use equal length binning on some attributes to find if it affects the relevancy of each attribute. The selected subset of data of 24 months is used to generate various data mining models using different training and testing sizes and binning sizes. C5 emerged consistently as the technique that have generated the best models with an average predictive accuracy as high as 94.68%. Sample sizes, equal-length binning sizes and training and testing sizes are all shown to have an effect on accuracy in different intensity.

Key words: Data mining techniques, predictive accuracy, incremental learning schemes.

INTRODUCTION

The credit card industry in Malaysia has undergone some major changes in the last decade as competition mounted and credit risks escalated. The central bank has taken measures to tighten credit spending, by imposing taxes for any new credit card issue. The government foresees that such action is needed to prevent any build up of credit bubbles and the subsequent economic implication that will ensue should the banking system collapse.

Banks are still very much attracted to the credit card business as it is one of the most profitable services to engage in though it is competitive. It could command lucrative margins as high as 1.8% interest charges a month. It is therefore important for banks to manage their risks properly, to maximize margins. Risks mitigation process in banks is more often than not based on information that they can find or mine from their historical database about the borrower and their tendency to default in their payment. The impact of improving predictive accuracy of prompt payment of customer is essential in the drive to ensure that the bank remain resilience.

Objectives of the study

This research seeks to discover the commonly used data mining tools and techniques used in banks in Malaysia. This research also employed the commonly used tools and data mining techniques including incremental learning schemes in an attempt to provide an improved classifier, ETL and data mining solution to help the bank in mitigating their risk portfolio.

LITERATURE REVIEW

Depending on the commercial uses of credit scoring, the methodology to construct credit scoring models varies from bank to bank. It may involve firstly, a sample of historical records classified as "good" and "bad" (or as bad loss, bad profit and good risk depending on the number of categories required) depending on their repayment
performance over a given period. Next, data could be obtained from internal or other external sources, namely, from credit bureau reports. Finally, statistical or other quantitative analysis is performed on the data to derive a credit scoring model (Koh et al., 2006).

With the right credit scoring model, the bank can evaluate any new or existing profiles of customers accurately, enabling them to minimize potential risks that might be looming. Such scoring models, together with the information provided by CCRIS (central credit reference information system) from the Central Bank and other information service provider form the basis to which credit rating systems were established. Credit rating systems is used to categorize the risk’s worthiness of a person as high, medium or low. This allows for decision support by accepting, extending or rejecting any credit request.

Data mining tools, techniques and credit risks evaluation

Various data mining techniques were employed by the banking and credit card industry in managing credit risks. Sinha and Zhao (2008) uses several techniques to examine the performance on business problems, mostly involving binary classification of two categories, that is, bankrupt or non-bankrupt, bad credit or good credit and others. In classification, a set of training data is use as input to build a model describing the predetermined set of data classes. Once the predictive accuracy of the model is acceptable, the model can be used to predict future data tuples (Fayyad et al., 1996; Lee, 2008).

Traditionally, logistic regression and discriminant analysis are the most widely used approaches to create scoring models in the industry (Yang, 2007). There are however ample evidence on the use of other data mining techniques related to credit evaluation such as bankruptcy prediction (Sung et al., 1999; Kim and Mcleod, 1999; Ryu and Yue, 2005), credit risk assessment (Doumpos et al., 2002), and credit evaluation (Sinha and Zhao, 2008). The techniques employed were as such, neural network (Back et al., 1996; Jo and Han, 1997), logistic regression (Desai et.al., 1996; Xiao et al., 2006) and decision tree (Koh et al., 2004).

The use of incremental learning schemes related to credit scoring or credit evaluation is however less profound. Existing application using mostly static models fail to adapt when environment or population changes over the time (Yang, 2007). The problem on updating the scoring model incrementally is not resolved. Yang adopted the incremental kernel method to build the adaptive scoring system.

Predictive accuracy of a classifier

The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data. This can be estimated using one or more data set or test sets (Han and Kamber, 2006).

Looking at the various data mining techniques under this research study, that is, logistics regression, decision tree, neural networks, naïve bayes updateable, IBk, and raced incremental logit boost, each technique tends to perform better than the rest under different circumstances. Decision tree has a better predictive accuracy in a study by Koh et al. (2004) on credit scoring, whereas neural networks has a better predictive accuracy in a study by Zurada and Lional (2005) on bad debt recovery.

The prevalent use of credit cards in recent years have created large stream of data for the bank to analyse. Incremental schemes are continually being explored to handle such development as it becoming more difficult to work with batch classifier as it requires a much larger system resources and effort to work with. The changes in the spending pattern of the consumers prompted by changes in the economic landscape or vice versa have also brought about a faster change to the risk profile of this business.

Most noticeable of system for incremental learning are credit risks evaluation (Li and Xin, 2009), credit scoring using incremental kernel method (Yang, 2007) and network security using IBk and naïve bayes updateable (Gandhi and Srivatsa, 2010).

ETL in data mining

According to Kimball and Caserta (2004), the ETL system or tool is very important in data mining as it consumes 70% of the resources required for data mining. The three core areas of the ETL processes are, firstly, extraction, the activity of extracting data from various sources and collating it to a target database. Transformation, or data pre-processing is the foundation for data analysis and mining (Zhang et al., 2007). Today’s real-world databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size (Han and Kamber, 2006). Any handicapped data passed on for the mining process and then to decision support system will result in reports and output that will be highly inaccurate, thereby, severely affecting business decisions and the business itself. Data has to be first verified before the mining, (Hsu, 2009).

RESEARCH DESIGN

This research design is made up primarily of a multi-methodological approach (Nunamaker et al., 1990, 1991) to the construction of the data mining solution. A quantitative methodology is adopted to systematically approach this investigation using a survey questionnaire to look at the adoption and maturity level in the use of data mining tools and techniques in the banking industry in Malaysia, prototyping methodology and experimentation for the development of ETL and data mining solution. The targeted population of this survey encompasses the credit cards department of
of all the banks in Malaysia. Banks having services in credit cards include nine local anchor banks and seven fully qualified foreign banks.

Samples sizes for the survey

The targeted population of the survey encompasses the credit cards department of all the banks in Malaysia. Banks having services in credit cards include nine local anchor banks and seven fully qualified foreign banks.

System modelling

The ETL system is designed and constructed with features that will allow the users to reconstruct the data with different equal-length binning sizes and the ability to compute information gain and gain ratios according to ID3 algorithm. These features are needed in this research to build the right data sample to locate the right classifier with an improved accuracy level.

Three batch classifiers and three incremental classifiers were used for this comparison. The techniques used are C5.1, neural net and logistic regression, naïve bayes updateable, IBk and raced incremental logit boost. With all data being fully substantiated, 5 attributes were used. They are age, location, sex, accumulated credit amount, credit limit/income level. Each attribute is assigned as “input”, the predictor field whereas the “class info” field is set as the “output” field, a binary classification of two categories – prompt payment or default in payment, the predicted fields for a machine-learning process of the data mining tool. Initially, various partition sizes are used for the entire batch of data for this analysis. The combination of the training and testing sample and partition sizes are set to 50 to 50%, 80 to 20% and 90 to 10% and partition sizes from 2, 10, 25 and 50 depending on the attributes. A randomly selected batch of subsample record of 5000 records of various partition sizes has been used for the training and testing of the models. As for achieving best predictive accuracy, each model is trained and tested for its highest score.

The test is then extended to have a larger subsample of 120,000 records representing 24 months of data from 5000 customer’s records. Each month of data is added incrementally. The equal length binning size used is also extended to 70 except for attributes with only a binary classification of two categories. The training and testing partitions sizes are set to 90 and 10% respectively.

The last part of the test involve the use of incremental learning scheme to ascertain if each smaller batches or streaming records would help to further improve the predictive accuracy by repeatedly rebuilding the model from each instances of data provided. The subsample size of 120,000 records, equal length binning of 70 and training and testing sizes of 90 and 10% remains. The incremental classifiers used for the test were naïve bayes updateable, IBk and raced incremental logit boost.

RESULTS

Descriptive analysis

The survey results indicated a progressive adoption of data mining techniques to evaluate credit risks in banks with 90% of them indicating some uses of it in various capacities. The most common technique used is confirmed as logistic regression as traditionally employed.

The level of use, about 60%, is predominantly found at the regional headquarters instead of the branch level. Most respondents are however not too sure about the predictive accuracy achieved on the prompt payment of customers.

Information gain and gain ratios

The results from the computation of information gain is as shown in Figure 1 and gain ratios with different equal-length binning sizes does suggest that the higher the partition size, the higher information gain and gain ratios will be.
Data mining techniques compared

The testing and training results for using one batch of the subsample at 5000 records and for using the entire subsample at 120,000 records spanning 24 months of 5000 each are as shown in Figure 2. The results using the incremental schemes are shown further.

Testing results of batch classifiers using a small dataset

C5 emerged as the best classifier with 91.46% accuracy level when tested. All models were trained and tested using a sample of 5000 records with the 90% of the sample used for training and 10% randomly picked to get the test results.

The results also show that equal-length binning sizes of various attributes and testing partition sizes do affect predictive accuracy. The single equal-length test of 10, 10, 2, 50 and 25 for each attribute 1 to 5 (age, location, sex, amount owing and credit limit/income level) respectively for neural network achieved the single highest single predictive accuracy of 92.46% as shown in Table 1.

Testing results of incremental learning schemes using a large dataset

Instance-based IBk has the best predictive accuracy from the three incremental learning models generated. The accuracy rate is at 93.63%. Naïve bayes updateable and raced incremental logit boost could only garner an accuracy level of 90.24 and 90.23%. This means the incremental learning scheme is not performing better that the C5 or J48 batch classifier using this set of data and setting.

LIMITATIONS AND ENHANCEMENTS

One of the obvious limitations encountered in this research is the limited attributes used. There is always a possibility that some other combination of attributes and data subset could provide a better information gain and gain ratios resulting in the overall predictive accuracy of the model to be improved. The limited number of attributes also resulted in a limited dimension in which data
The various equal-length binning sizes appeared to have an effect on predictive accuracy. The higher the number of records or percentages sizes of samples use for testing, the better the training and testing accuracy seem to be.

This observation and relationship could be validated if more records are obtained for testing the models. This could further improve predictive accuracy.

C5 or decision tree seems to have the best predictive accuracy among not only the batch classifier, but also the incremental ones. It will be good if an incremental decision tree technique is made available to see if it will exceed what the current tool is able to improve on.

**Conclusion**

The successful completion of this research has provided new insights into the use of data mining tools and techniques in a bank. Factors such as the sample size, the equal-length partition sizes and the training and testing partition size have had an effect on predictive accuracy.

The objectives put forward was successfully met with the identification of a batch classifier and model that attain a higher level of predictive accuracy beyond that is initially expected.

**REFERENCES**


