Model Selection Strategy for Customer Attrition Risk Prediction in Retail Banking

Fan Li [1], Juan Lei [2], Ying Tian [3], Sakuna Punyapatthanakul [4] and Yanbo J. Wang [1, 5]

[1] Information Management Center, China Minsheng Banking Corp., Ltd.
[2] Department of Retail Banking, China Minsheng Banking Corp., Ltd.
[3] Beijing Dongdan Sub-branch, China Minsheng Banking Corp., Ltd.
No. 2, Fuxingmennei Avenue, Xicheng District, Beijing 100031, China

[4] Consumer Segment Management Department, Retail Business Division, KASIKORNBANK
1 Soi Rat Burana 27/1, Rat Burana Road, Bangkok 10140, Thailand

[5] Institute of Finance and Banking, Chinese Academy of Social Sciences
No. 5, Jianguomennei Dajie, Dongcheng District, Beijing 100732, China

{lifan,leijuan2,tianying,wangyanbo}@cmbc.com.cn
sakuna.p@kasikornbank.com

Abstract

Nowadays customer attrition is increasingly serious in commercial banks, particularly, *high-valued* customers in retail banking. Hence, it is encouraged to develop a prediction mechanism and identify such customers who might be at risk of attrition. This prediction mechanism can be considered to be a classifier. In particular, the problem of predicting risk of customer attrition can be prototyped as a *binary* classification task in data mining. In previous studies, a number of techniques have been introduced in (*binary*) classification study, i.e. artificial-based model, Bayesian-based model, case-based model, tree-based model, regression-based model, rule-based model, etc. With regards to a particular application — predicting customer attrition risk for retail banking, this paper presents four principles in (classification) model selection. To support this model selection study, a set of experiments were run, based on a collection of *real* customer data in retail banking. These results and consequent recommendations are given in this paper.

Keywords: Classification Prediction, Commercial Banks, Customer Attrition Risk, Model Selection, Retail Banking.

1 Introduction

With increased competition within the domestic banking industry, customer churn/attrition is increasingly serious in commercial banks, particularly, *high-valued* customers in retail banking. Nowadays, more and more commercial banks start to pay attention to CRM (Customer Relationship Management), especially the investigation of retaining existing customers. In the work (Luck, 2009), the author clearly states that "*retaining customers is more profitable than building new relationships*". Kandampully and Duddy (1999) even attempt to clarify that attracting a new customer is about five times more costly than retaining an existing customer. Hence, "*the retention of existing customers has become a priority for businesses to survive and prosper*" (Luck, 2009). Consequently, accurately identifying those customers who might be at risk of attrition has become an essential problem. For commercial banks in general, it is suggested to produce a prediction mechanism that can be used to classify whether an existing customer will churn in the near future (in the next business/observation period).

The rest of this paper is organised as follows. The following section indicates the link between our problem of study and the data mining classification task and summarises some related works. Section 3 presents the strategy of model selection, which mainly consists of four principles. Experimental results, based on the collected VIP customer data from a *real* retail banking environment are shown in Section 4. During the experiments, a number of classification models were compared and the most suitable one was selected for each of the principles. Finally, conclusions and direction for future work are given at the end of this paper.

2 Related Work

2.1 Classification Models

The customer attrition risk identification problem can be prototyped as a *binary* classification task in data mining — *binary* classification, also referred to as 2-*class* classification, "*learns from both positive and negative data samples, and assigns either a predefined category* (*class-label*) *or the complement of this category to each 'unseen' instance*" (Wang *et al.*, 2011). In past decades, many models/techniques have been proposed in the study of (2-*class*) classification that include: Artificial-Based Classification (ABC), Bayesian-Based Classification (BBC), Case-Based Classification (CBC), Tree-Based Classification (TBC), Regression-Based Classification (REBC), Rule-Based Classification (RUBC), etc.

- **ABC Model** aims to solve the classification problem by using *AI* (*Artificial Intelligence*) techniques. One typical approach is the Artificial Neural Network (ANN). Knowledge on ANN classification can be found in the study (Berson and Smith, 1997, 375-406).

- **BBC Model** aims to solve the classification problem using the *Bayesian* theory. One well known approach is Naïve Bayes (NB). In the work (Wang, 2007, 24-25), the author depicts the general idea of NB classification.
- **CBC Model** aims to solve the classification problem by *lazily* utilising the training examples/ cases. One typical approach is using *k*–Nearest Neighbours (*k*–NN). Knowledge relating to *k*–NN classification has been contributed by Cunningham and Delany (2007).
- **TBC Model**'s approach to solving the classification problem is based on a *greedy* algorithm. The result of classifier construction using TBC is to build a *Decision Tree* (*DT*). C4.5/C5.0 (Quinlan, 1993) is the best known *DT* classification approach.
- **REBC Model** aims to solve the classification problem using the statistical *regression* study. The approach is namely Logistic Regression (LR). Study as related to LR classification can be found in the work (Witten and Frank, 2005, 121-125).
- **RUBC Model** aims to solve the classification problem by generating a set of "*IF–THEN*" patterns/rules, where each rule is expressed in the form of "*attribute*(*s*) ⇒ *category*". The generated set of (human readable and understandable) rules represents the (constructed) classifier and presents to the end users why and how the classification predictions have been made. One typical mechanism in this school is namely RIPPER (Cohen, 1995).

2.2 Previous Work

In the previous studies of customer attrition risk identification, Khan, Jamwal and Sepehri (2010) indicate that TBC DT, REBC LR, and ABC ANN can be almost equally well applied in the Internet Service Provider (ISP) industry. By running experiments on real data collected from the ISP industry, Khan, Jamwal and Sepehri (2010) point out that ABC ANN slightly outperforms the other two models/approaches on "overall accuracy"; and REBC LR slightly outperforms the other two models/ approaches on "churner hit rate" (also referred to as "recall of attrition"). In the work (Li, 2009), the author suggests to use TBC DT in banking customer attrition risk prediction. However Li's work does not show the experimental results in model performance evaluation.

3 Model Selection Strategy

With regards to a particular application — predicting customer attrition risk for retail banking, our study proposes four principles that can be applied strategically in classification model selection.

3.1 Principle of Good Performance

Broadly speaking, we say a classification model demonstrates good performance if the model shows good classification accuracy. Simply speaking, classification accuracy is the fraction of correctly predicted "instance–category" mappings, which is calculated as the number of correctly classified instances divided by the total number of instances to be classified.

Other measures that have been used in classification performance evaluation, especially in *binary* classification (Zheng and Srihari, 2003), include: recall, precision, the F1 measure, micro-averaging, macro-averaging, etc. With respect to the application of Text Categorization (TC), Sebastiani (2005) explains the reason for applying these evaluation measures in *binary* classification rather than using accuracy alone: "*in binary TC application the two categories c and c are usually unbalanced, i.e. one contains for more members than the other*". Therefore, "*building a classifier that has high accuracy is trivial*", i.e. a classifier could directly assign the majority class-label found in the *training* dataset for all *test* instances and reach a high classification accuracy without any (serious/ intelligent) computation to be involved.

In our study, the data collected from a *real* retail banking environment is not (necessarily) class-*balanced* too — usually the number of customers who are going to churn is less than the number of customers who will stay. However, this is not the basis for recommending utilising recall and precision in classification model evaluation/ selection. In a banking context, accurately/correctly identifying a customer who is at risk of attrition is more valuable than finding a customer who will truly stay. Hence, the involvement of "recall of attrition" is strongly recommended in model evaluation and in our study, this measure in some cases stands for the execute-ability of customer retention. The "recall of attrition" is calculated as the number of truly churned customers who have already been correctly predicted divided by the total number of truly churned customers.

In the extreme situation of 100% "recall of attrition", all customers are going to churn. In this case, all truly churned customers can definitely be identified in the prediction phase, but obviously this is not a good prediction model. Consequently, the "precision of attrition" measure is of central concern in our study. The "precision of attrition" is calculated as the number of truly churned customers who have already been correctly predicted divided by the total number of customers who have been predicted to churn. Actually, this measure represents the cost-level of customer retention — i.e. a high "precision of attrition" means that most of the churn-predicted customers are truly churned customers. This results in the most efficient management of customer retention funds.

In our study, we use the "overall accuracy" and the "recall of attrition" plus "precision of attrition" to measure whether a classification model satisfies the principle of good performance.

3.2 Principle of Efficiency

Simply speaking, efficiency is a measure of time and can be used to demonstrate how fast an information system can be processed. The most straightforward way to evaluate an information system's efficiency is to count the system's running time in seconds. In our study, we count both the "classifier building time" and the "overall running time" to determine whether a classification model satisfies the principle of efficiency.

3.3 Principle of Instance Ranking

In a *real* banking environment, especially in retail banking, the number of existing customers (shown as the collected customer data-instances) can be as large as some ten millions. We can assume that at least 10% of these customers are predicted/going to churn, which means there are more than one million customers who require us to implement customer-care (for customer retention). Obviously, this number is too large to be handled by a commercial bank at one time. Hence, it is necessary to distinguish customers with a high probability to churn from the ordinary ones. Further, it is encouraged to rank all customer-instances in descending order, based on their churn probabilities, so that the bank can easily select a suitable size of predicted (the most probable) churn-customers to implement customer-care. In our study, we use both the "smoothness of probability distribution" and the "slope of probability distribution" to measure whether a classification model satisfies the principle of instance ranking.

3.4 Principle of Rule Generation

In a general context, classification models can be separated into two schools — (*i*) classification without rule generation *vs.* (*ii*) classification with rule generation (and presentation). In the early stage of classification investigation (1960's ~ 1980's), most of the models/ approaches were proposed by the first school. In the past two decades (1990's ~ present), the tree-based and rule-based classification techniques were introduced. These aim to generate human-readable and human-understandable patterns/rules while classifying "unseen" data-instances and present to the end users why and how the classification predictions have been made. In our study, the rules generated to demonstrate why and how a set of customers are at risk of attrition are clearly stated and are essential to the design of an *effective* customer-care program for customer retention.

4 Experimental Results

In this section, we present four groups of evaluations for our proposed model selection strategy — one for each principle, using a set of collected VIP customer data from a *real* retail banking environment. All evaluations were obtained using the WEKA software [1] (Witten and Frank, 2000; Witten and Frank, 2005; Witten, Frank and Hall, 2011). The experiments were run on a 3.00 GHz Pentium(R) Dual-Core CPU with 1.96 GB of RAM running under the x86 Windows Operating System.

4.1 Description of Data

From a *real* commercial bank's EDW (Enterprise Data Warehouse), we collect a set of retail banking VIP customer (attrition) data across nine *continuous* months. The class-labels (also noted as the category-attribute values) of this dataset are "*attrition*" *vs.* "*non-attrition*", which represents the churn status of each (VIP) customer in the seventh month to the ninth month. The features (also noted as the data-attributes) in this dataset are grouped into customer's "*geo-demographical*", "*financial*", "*product*", "*transaction*" and "*loan*" information, which together, clearly depicts each customer in the first month to the sixth month.

After the data cleansing process — data-instances with missing and/or noisy values are eliminated. We then randomly select 2000 "*attrition*" plus 2000 "*non-attrition*" data-instances to create a class-*balanced* dataset. In feature selection, we choose only 12 simple data-attributes, i.e. customer's age, number of products holding, time to stay with the bank, etc., based on suggestions given by the bank's financial managers.

4.2 Description of Models and Approaches

In the WEKA software version 3.6.4, the implementation of ABC ANN approach is namely MultilayerPerceptron; the implementation of BBC NB is namely NaiveBayes; the CBC *k*–NN approach is namely IBk, and in our study *k* was set to be 5; the TBC DT (C4.5/C5.0) is namely J48; the REBC LR is namely Logistic; and the RUBC RIPPER approach is namely JRip. In our experiments, we ran these implemented methods using our prepared data on the WEKA platform.

4.3 Description of Results

First of all, the six classification models/approaches (as introduced above) are evaluated by "overall accuracy", "recall of attrition" and "precision of attrition" (see Table 1). The experimental results were obtained using the Ten-fold Cross Validation (TCV) setting. From the evaluation, the BBC NB model/approach (as highlighted in Table 1) only shows 61.8% "overall accuracy". Although it recognises 94.3% of the customers who are truly going to churn, the cost of obtaining this "recall of attrition" is very high (the "precision of attrition" is only 57.2%). In this case, we suggest abandoning BBC NB. Results from other models/approaches are valued at the same level.

Models Approaches	Recall of Attrition	Precision of Attrition	Overall Accuracy
ABC ANN	78.9%	87.8%	83.950%
BBC NB	94.3%	57.2%	61.800%
CBC *k*–NN	79.2%	83.1%	81.525%
TBC DT	80.3%	87.6%	84.425%
REBC LR	79.3%	79.3%	79.325%
RUBC RIPPER	79.9%	88.3%	84.675%

Table 1: Experimental results for the principle of good performance

Secondly, the "classifier building time" and the "overall running time" for the six classification models/approaches were counted in seconds and the results are shown and compared as follows (see Table 2). In general, the "classifier building time" is less than 0.7 seconds and the "overall running time" is less than 8 seconds, except ABC ANN (as highlighted in Table 2). It can be evaluated that the run-time efficiency of ABC ANN is at least 15 times higher than the efficiency of other

[1] The well known WEKA software, a Data Mining and Machine Learning Software in Java, may be obtained from http://www.cs.waikato.ac.nz/~ml/weka/.

models/approaches (calculated as 12.55 √ 0.7 ≈ 17.9 and 122 √ 8 ≈ 15.6). Note that in our experiments, the prepared dataset involves only 13 data-attributes (including the category-attribute) and contains only 4000 data-instances. If it were to handle a very large data collection in our study, the run-time efficiency of ABC ANN may not demonstrate equal endurance. Hence, it is suggested to abandon the ABC ANN model/approach.

Models Approaches	Classifier Building Time (in Sec.)	Overall Running Time (in Sec.)
ABC ANN	12.55	122
BBC NB	0.02	1
CBC *k*–NN	0.0001	7
TBC DT	0.2	2
REBC LR	0.14	2
RUBC RIPPER	0.61	7

Table 2: Experimental results for the principle of efficiency

The results of the third evaluation (for the principle of instance ranking) are shown as follows (see Table 3, 4 and Figures 1 ~ 6). From the experiments, the six produced classifiers (with regards to the issue of *anti-overfitting*), based on the six introduced classification models/ approaches were employed to assign a score of churn/ attrition probability (between 0 and 1) to each of the 4000 originally given customers/data-instances. Note that this operation can be done easily in WEKA by simply selecting the "Output predictions" box after clicking the "More options… (Classifier evaluation options)" button under the "Weka Explorer — Classify" subtitle.

We draw the customer attrition probability distributions generated by each of the six approaches/classifiers into graphs (see Figures 1~6). From these figures, we see that the Figures 5, 2 and 1 show better graph smoothness ("smoothness of probability distribution") than the Figures 4, 6 and 3. In fact, the probability (score) values of ABC ANN, BBC NB and REBC LR are *continuous*, whereas the probability values of CBC *k*–NN, TBC DT and RUBC RIPPER are *discrete*. It can be argued that the usability of *discrete* valued probability distribution is weak, e.g. there are only 9 *discrete* values through the customer attrition probability distribution of CBC *k*–NN and RUBC RIPPER (see Table 4), so that it is difficult to catch the top 400 and later on the next 500 customers who are the most probable to churn, like the result provided by REBC LR as follows (see Table 3).

Models Approaches	Number of Instances (based on the attribution probability score)					Value Desc. of Probability Distribution
	≥ 0.9	≥ 0.8	≥ 0.7	≥ 0.6	≥ 0.5	
ABC ANN	1475	1552	1739	1793	1883	Continuous
BBC NB	2076	3002	3189	3253	3301	Continuous
CBC *k*–NN	1018	1543	1543	1895	1896	Discrete
TBC DT	1472	1641	1777	1783	1783	Discrete
REBC LR	407	578	915	1486	2007	Continuous
RUBC RIPPER	1213	1576	1600	1807	1807	Discrete

Table 3: Experimental results for the principle of instance ranking

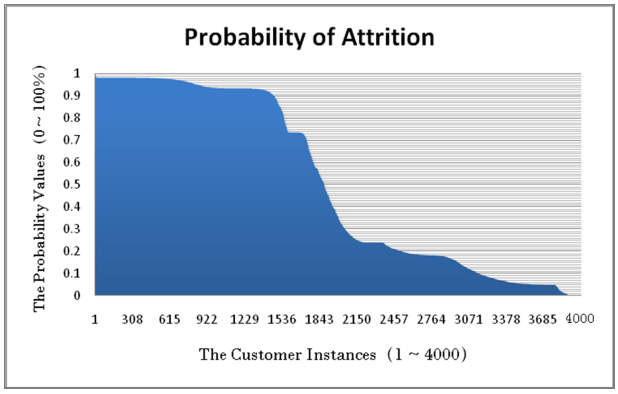

Figure 1: The probability distribution graph of customer attrition by ABC ANN

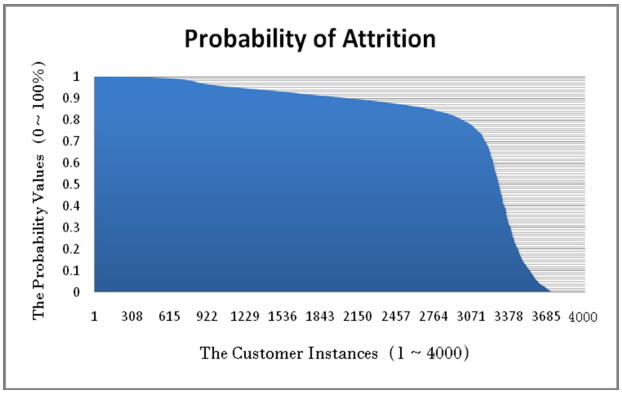

Figure 2: The probability distribution graph of customer attrition by BBC NB

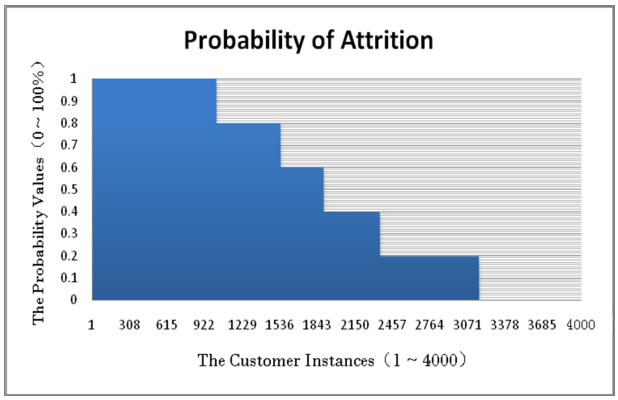

Figure 3: The probability distribution graph of customer attrition by CBC *k*–NN

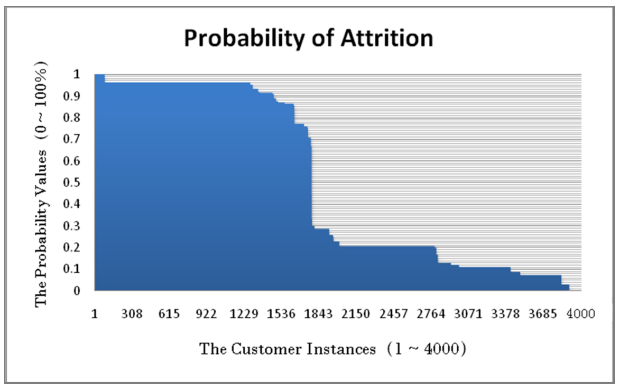

Figure 4: The probability distribution graph of customer attrition by TBC DT

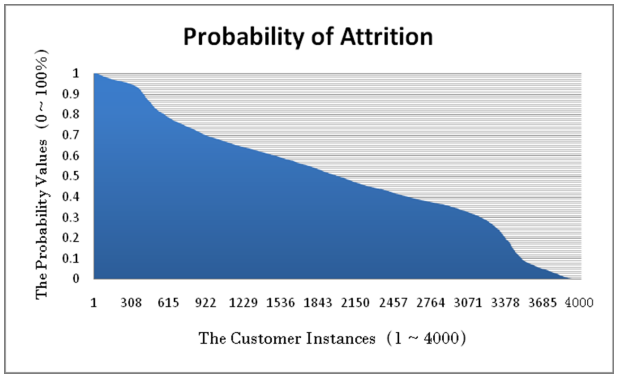

Figure 5: The probability distribution graph of customer attrition by REBC LR

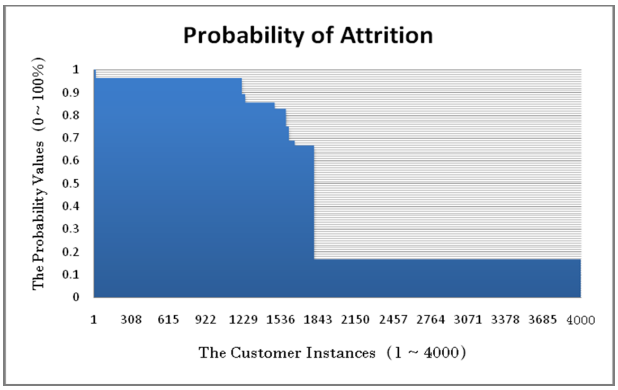

Figure 6: The probability distribution graph of customer attrition by RUBC RIPPER

The # of Discrete Values	CBC *k*–NN		RUBC RIPPER	
	The Discrete Values (in ↓ order)	Accumulated Number of Customers	The Discrete Values (in ↓ order)	Accumulated Number of Customers
1	1	1018	1	13
2	0.8	1543	0.963	1213
3	0.6	1895	0.893	1241
4	0.5	1896	0.855	1483
5	0.4	2355	0.828	1576
6	0.3	2356	0.75	1600
7	0.2	3166	0.688	1648
8	0.1	3167	0.667	1807
9	0	4000	0.167	4000

Table 4: Detailed description of the customer attrition probability distribution for CBC *k*–NN & RUBC RIPPER

From the "slope of probability distribution", we see that the Figures 1, 2, 4 and 6 show a flatter curve slope than the Figures 3 and 5 for the first 1500 customers. The flatter curve slope represents the difficulty of identifying significant churn-predicted customers from ordinary customers. From Table 3, we see that ABC ANN, BBC NB, TBC DT and RUBC RIPPER catch respectively 1475, 2076, 1472 and 1213 customers that are predicted with higher than 90% probability to churn. Note that we only have 2000 truly churned customers in total.

By adopting both the "smoothness of probability distribution" and the "slope of probability distribution", the only model/approach that satisfies the principle of instance ranking is REBC LR. It is advisable to abandon other models/approaches (as highlighted in Table 3).

Finally, we look into the principle of rule generation. A set of experiments were run on our prepared dataset. The experimental results are shown as follows (see Table 5), where only TBC DT and RUBC RIPPER are able to generate classification rules and present to the end users why and how the customer attrition predictions have been made. Based on the TBC DT approach, a tree classifier was constructed that contains 82 leaf nodes (classification rules). Figure 7 partially shows the tree classifier. Moreover, the RUBC RIPPER approach generates 9 classification rules; Figure 8 partially lists the rule classifier. The quality of these generated rules, in terms of the extent to which they correlate with a *priori* knowledge, can be confirmed by experienced financial managers.

Models Approaches	Rule Generation	Number of Rules
ABC ANN	×	—
BBC NB	×	—
CBC *k*–NN	×	—
TBC DT	√	82
REBC LR	×	—
RUBC RIPPER	√	9

Table 5: Experimental results for the principle of rule generation

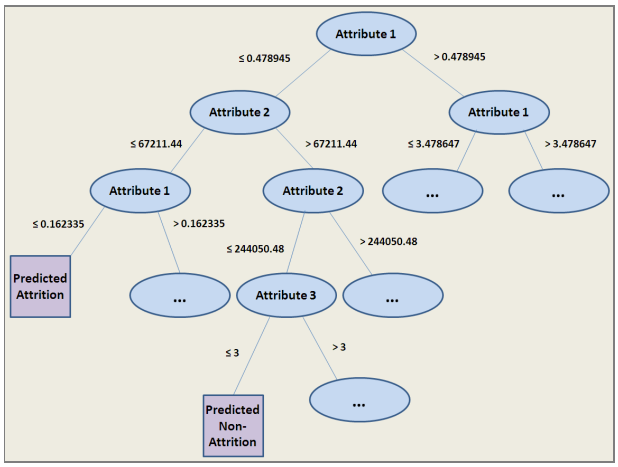

Figure 7: Rule generation by TBC DT
(The J48 decision tree classifier is shown partially)

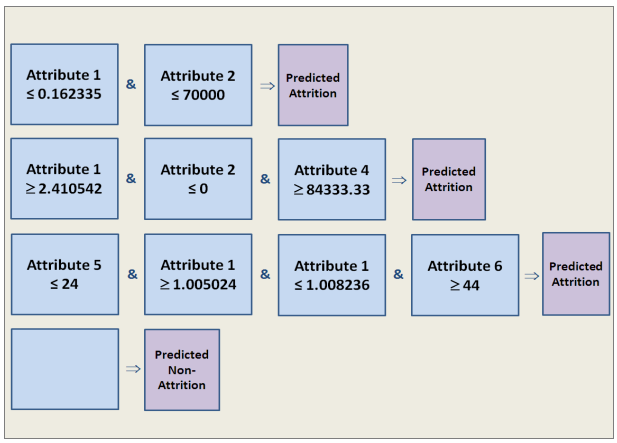

Figure 8: Rule generation by RUBC RIPPER
(The JRip rule classifier is shown partially)

4.4 Summary

Four groups of experiments were run — one for each of the proposed model selection principles. From the experimental results, it can be summarised that the ABC ANN, BBC NB and CBC *k*–NN models/approaches should be abandoned due to the following reasons:

- **ABC ANN** does not demonstrate acceptable performance in terms of "overall accuracy" and "precision of attrition".
- **BBC NB** does not show an acceptable level of efficiency in terms of "classifier building time" and "overall running time".
- **CBC *k*–NN** satisfies neither the principle of instance ranking nor the principle of rule generation.

With ability to satisfy both the principles of performance and efficiency, we suggest to adopt REBC LR since this classification model/approach satisfies the principle of instance ranking well. Furthermore, we recommend also adopting the TBC DT and RUBC RIPPER models/ approaches since both demonstrate equal capacity to satisfy the principle of rule generation.

In a banking context, it is preferable to provide explanation of why and how a customer is predicted to be at risk to churn when adopting the REBC LR classifier. The Voice Of Customer (VOC) analysis (by questionnaire) that investigates the reason of customer attrition, is definitely suggested. The result of VOC analysis, as a substitute of the generated rule list/set, can be used to design the customer-care program for customer retention.

On the other hand, if the TBC DT or RUBC RIPPER classifier were adopted, it would be encouraged to develop a better churn/attrition probability scoring mechanism, which identifies the top-most probable *f*-% customers to churn (in the near future). Here *f* can be any number between 0 and 100. Again, the VOC analysis should be utilised, as it ranks all or a fraction of the churn-predicted customers in descending order, based on their churn probability.

5 Conclusions

Today, customer attrition, especially for *high-valued* customers in retail banking, has become more and more serious in commercial banks. Hence, customer retention, and consequently predictions of customer attrition have become a priority issue for the survival and prosperity of commercial banks. In this paper, we investigated the customer attrition prediction problem based on a collection of customer (attrition) data from a *real* retail banking environment. By prototyping our study into a data mining *binary* classification problem, we listed a number of classification models/approaches that can be selected. We further proposed four model selection principles, and a set of experiments were run based on our prepared dataset. The experimental results show that although none of the models are perfect, the REBC LR, TBC DT and RUBC RIPPER models/approaches are recommended for adoption for customer attrition prediction in retail banking. Further research is suggested to produce an improved classification model/approach that satisfies all our proposed model selection principles simultaneously.

6 Acknowledgements

The authors would like to thank Dr. Jiongyu Li from the China Minsheng Banking Corp., Ltd., Pipit Aneaknithi and Lei Xiao from KASIKORNBANK (Thailand), Jiangtao Lai from IWT Solutions Ltd. (KXEN Exclusive Distributor in China and Hong Kong), Haixia Pan from the College of Software at Beihang University, and Karen Zhang from Work Place Safety and Insurance Board (Ontario, Canada) for their support with respect to the work described here.

7 References

Berson, A. and Smith, S.J. (1997): *Data warehousing, data mining, and OLAP*. New York, NY, McGraw-Hill Companies, Inc.

Cohen, W.W. (1995): Fast effective rule induction. *Proc. of the 12th International Conference on Machine Learning*, Tahoe City, CA, 115-123, Morgan Kaufmann Publishers.

Cunningham, P. and Delany, S.J. (2007): *k*–nearest neighbour classifiers. Technical report (UCD-CSI- 2007-4). University Colledge Dublin, Ireland.

Khan, A.A., Jamwal, S. and Sepehri, M.M. (2010): Applying data mining to customer churn prediction in an internet service provider. *International Journal of Computer Applications* **9**(7): 8-14.

Li, X. (2009): ID3 applying to loss of bank clients. *Computer Technology and Development* **19**(3): 158-167.

Luck, D. (2009): The importance of data within contemporary CRM. In the book *Data Mining Applications for Empowering Knowledge Societies*. 96-109. Rahman, H. (ed). Hershey, PA, IGI Global.

Kandampully, J. and Duddy, R. (1999): Relationship marketing: a concept beyond primary relationship. *Marketing Intelligence and Planning* **17**(7): 315-323.

Quinlan, J.R. (1993): *C4.5: programs for machine learning*. San Francisco, CA, Morgan Kaufmann Publishers.

Sebastiani, F. (2005): Text categorization. In the book *Text Mining and Its Applications to Intelligence, CRM and Knowledge Management (Advances in Management Information)*. 109-129. Zanasi, A. (ed). Southampton, UK, WIT Press.

Wang, W., Wang, Y.J., Xin, Q., Bañares-Alcántara, R., Coenen, F. and Cui, Z. (2011): A comparative study of associative classifiers in mesenchymal stem cell differentiation analysis. In the book *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains*. 223-243. Kumar, A.V.S. (ed). Hershey, PA, IGI Global.

Wang, Y.J. (2007): Language-independent pre-processing of large documentbases for text classification. Ph.D. thesis. University of Liverpool, UK.

Witten, I.H. and Frank, E. (2000): *Data mining: practical machine learning tools and techniques with java implementations*. San Francisco, CA, Morgan Kaufmann Publishers.

Witten, I.H. and Frank, E. (2005): *Data mining: practical machine learning tools and techniques (second edition)*. San Francisco, CA, Morgan Kaufmann Publishers.

Witten, I.H., Frank, E. and Hall, M.A. (2011): *Data mining: practical machine learning tools and techniques (third edition)*. Burlington, MA, Morgan Kaufmann Publishers.

Zheng, Z. and Srihari, R. (2003): Optimally combining positive and negative features for text categorization. *Proc. of the 2003 ICML Workshop on Learning from Imbalanced Data Sets II*, Washington DC.