Data Mining Techniques in Customer Churn Prediction

Chih-Fong Tsai[1,*] and Yu-Hsin Lu[2]

[1]*Department of Information Management, National Central University, Taiwan,* [2]*Department of Accounting and Information Technology, National Chung Cheng University, Taiwan*

Received: August 21, 2009; Accepted: October 5, 2009; Revised: October 15, 2009

Abstract: Customer churn prediction is one of the most important problems in customer relationship management (CRM). Its aim is to retain valuable customers to maximize the profit of a company. To predict whether a customer will be a churner or non-churner, there are a number of data mining techniques applied for churn prediction, such as artificial neural networks, decision trees, and support vector machines. This paper reviews some recent patents along with 21 related studies published from 2000 to 2009 and compares them in terms of the domain dataset used, data pre-processing and prediction techniques considered, etc. Future research issues are discussed.

Keywords: Churn prediction, data mining, customer relationship management.

1. INTRODUCTION

Since enterprises in the competitive market mainly rely on the profits which come from customers, Customer Relationship Management (CRM) always concentrates on confirmed customers that are the most fertile source of data for decision making. That is, highly competitive organizations have understood that retaining existing and valuable customers is their core managerial strategy to survive in their industries. However, to create and retain customers is difficult and costly in terms of marketing. Consequently, this leads to the importance of churn management [1, 2].

As customer churning will likely result in the loss of businesses, churn prediction has received increasing attention in the marketing and management literature over the past time. In addition, it shows that a small change in the retention rate can result in significant impact on businesses [3, 4].

Customer churn can be regarded as customers who are intending to move their custom to a competing service provider. Therefore, many firms need to assess their customers' value in order to retain or even cultivate the profit potential of the customers [5, 6].

In order to effectively manage customer churn for companies, it is important to build a more effective and accurate customer churn prediction model. In the literature, data mining techniques have been used to create the prediction models.

This paper is organized as follows: Section 2 describes the data mining process including data pre-processing, data prediction, and evaluation methods. Next, related towork using data mining techniques for churn prediction is compared. In addition, discussions of this literature review are provided. Finally, the conclusion is provided in Section 3.

*Address correspondence to this author at the Department of Information Management, National Central University, Taiwan; Tel: +886-3-422-7151; Fax: +886-3-4254604; E-mail: cftsai@mgt.ncu.edu.tw

2. LITERATURE REVIEW

2.1. The Data Mining Process

Data mining has emerged over recent years as an extremely powerful approach to extracting meaningful information from large databases and data warehouses. Since the increased computerization of business transactions, improvements in storage and processing capacities of computers, as well as significant advances in knowledge discovery algorithms, those all have contributed to the evolution of the data mining [7]. Some data mining techniques were described by Aharoni *et al.* [8]. The methodology of data mining views the discovery of information from a database as a four-step process [1]. First, the business problem must be identified. After the problem is defined and related data are collected, the next step is to process the collected data by data transformation, data cleaning, etc. for the later mining process. The third step is to apply some specific mining algorithm(s) over the processed data. In this paper, prediction/classification algorithms can be used. Finally, the mining result is evaluated to examine whether the finding is useful for the business problem.

2.1.1. Data Pre-Processing

The data in the real world is always incomplete, noisy, and inconsistent because of not applicable, human or computer error at data entry, errors in data transmission, or from different data sources, etc. Therefore, the major tasks in data pre-processing includes data cleaning, data integration, data transformation, data reduction, and data discretization [9].

Data cleaning is one of the three biggest problems in data warehousing (Kimball, 1996). In data cleaning process, some tasks may be to fill in missing values, identify outliers, smooth out noisy data, correct inconsistent data, and resolve redundancy caused by data integration. Missing and noisy data are resolved by using attribute mean to fill in, or employing a regression function to find a fitted value generally.

1874-4796/10 $100.00+.00

Data integration is to combine data from multiple sources into a coherent store. In integration processing, the redundant data problem always occur, since the same real world entity, attribute values from different sources have different names, or one attribute may be a derived attribute in another table. Therefore, researchers should carefully identify real world entities from multiple data sources by using correlation analysis. Otherwise, careful integration of the data from multiple sources may help to reduce/avoid redundancies and inconsistencies and improve mining speed and quality [9].

Data transformation is usually used to smooth the noisy data, summarize, generalize, or normalize the data scale falls within a small, specified range. In addition, data reduction aims to obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results. Since complex data mining may take a very long time to run on the complete data set, data reduction is usually employed in the data pre-processing stage. Data reduction tasks include data cube aggregation, dimensions reduction (e.g., removing unimportant attributes), data compression, and to reduce data volume by choosing alternative, smaller forms of data representation.

Finally, data discretization divides the range of a continuous attribute into intervals since some classification algorithms only accept categorical attributes. After data preprocessing, data analysis and mining can be proceed efficiently and effectively.

2.1.2. Data Prediction/Classification

After data are pre-processed, knowledge discovery algorithms can be applied to the processed data. The type of algorithms used, depends on the nature of the problem. If the problem can be viewed as a problem of classification or prediction, and a complete set of training data is available, then the problem is well structured. Supervised learning algorithms like multilayer neural networks, regression, or decision trees can be used to learn the relationship between variables and correct decisions. The followings describe these three well-known algorithms in churn prediction.

Neural networks are the popular and widely used algorithm in data mining. It attempts to simulate biological neural systems which learns by changing the strength of the synaptic connection between neurons upon repeated stimulation by the same impulse [10]. Neural networks can be distinguished into single-layer perceptron and multilayer perceptron (MLP). The multilayer perceptron consists of multiple layers of simple, two taste, sigmoid processing nodes or neurons that interact by using weighted connections. The MLP network may contain several intermediary layers between input and output layers. Such intermediary layers are called as hidden layers and composed of a number of nodes embedded in these layers, which are called as hidden nodes. Based on prior research results [11], multilayer perception is a relatively accurate neural network model. Data mining techniques and methods were described by Qian *et al.* with the description of some neural networks [12].

A decision tree is constructed by many nodes and branches on different stages and various conditions. It is a very popular and powerful tool for many prediction and classification problems, since it can produce a number of decision rules. Several algorithms of decision trees have been developed, such as C4.5 and C5.0. Among them, classification and regression trees (CART) developed by Breiman *et al*. [13] is a non-parametric statistical method to construct a decision tree to solve classification and regression problems.

Logistic regression is a widely used statistical modeling technique to compare with other data mining algorithms, in which the probability of a dichotomous outcome is related to a set of potential independent variables and it is used to forecast the value of two class labels or sequence variables [14, 15]. Although it is one of the traditional statistical techniques, the logistic regression model does not necessarily require the assumptions of discriminant analysis, but it is as efficient and accurate as discriminant analysis.

2.1.3. Evaluation Methods

To evaluate the performance of churn prediction models, the average prediction accuracy and Type I and Type II errors are usually examined in related literature. Table **1** shows a confusion matrix used for obtaining the performance measures.

Table 1. Confusion Matrix

		Actual	
		Non-Churners	**Churners**
Predict	Non-churners	a	b (II)
	churners	c (I)	d

The rate of prediction accuracy is based on $\frac{a+d}{a+b+c+d}$.

Type I error means the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In customer churn prediction, Type I error means the event was occurred when the model predicts the non-churners group as the churners group. Type II error represents the error of rejecting a null hypothesis when it is the true state of nature. In customer churn prediction, the Type II error means the event was occurred when the model predicts the churners group as the non-churners group.

In addition, in order to enhance the reliability of the evaluation result *N*-fold cross-validation is usually used. It is based on dividing *N* equal parts of a given dataset, in which *N*-1/*N* of the dataset performs model training, and the rest for model testing. Therefore, every subset will be trained and tested *N* times, and the average prediction performance can be obtained consequently.

2.2. Comparisons of Related Work

Table **2** provides the comparisons of recent related work using data mining techniques for churn prediction. In particular, the domain datasets and pre-processing methods used, classification techniques developed, and evaluation methods considered are compared.

Table 2. Comparisons of Related Work

Authors	Domain Datasets	Pre-Processing Methods	Prediction Techniques	Evaluation Methods	
				Type I/II Error	Cross Validation
Glady *et al*. (2009) [16]	Belgian financial service	X	Neural Network (MLP); Decision tree; Logistic regression	X	X
Pendharkar (2009) [17]	Cellular wireless network services	X	Genetic Algorithm + NN; Z-Score classification model	X	X
Xie *et al.* (2009) [18]	Chinese banking	Feature selection	Neural Network; Decision tree; Support vector machine	X	X
Tsai and Lu (2009) [19]	American Telecom Companies	X	Hybrid Neural Network;	V	V
Burez and Van den Poel (2008) [20]	Pay-TV company	X	Random forests Survival analysis	X	X
Coussement and Van den Poel (2008) [21]	Belgian newspaper publishing company	X	Support vector machines; Random forests Logistic regression	X	V
Coussement and Van den Poel [22]	Belgian newspaper publishing company	Dimension reduction	Logistic regression	X	V
Burez and Van den Poel (2007) [23]	Pay-TV company	X	Logistic regression; Markov chains; Random forests	X	X
Chu *et al.* (2007) [24]	Taiwan telecom company	Dimension reduction	Decision tree; Growing Hierarchical Self-organizing map	X	X
Luo *et al.* (2007) [25]	Personal Handy- phone System Service	X	Neural Network; Decision tree	X	V
Hung *et al.* (2006) [26]	Wireless telecom services	Dimension reduction	Decision tree; Neural Network; K-means	X	V
Nie *et al.* (2006) [27]	Charge Email	Feature selection	Decision tree	X	X
Buckinx and Van den Poel (2005) [28]	Retail industry	X	Logistic regression; Neural Network; Random forests	X	X
Larivie`re and Van den Poel (2005) [4]	Belgian financial services	X	Random forests; Regression forests; Linear regression; Logistic regression	X	X
Van den Poel and Larivie`re (2004) [3]	European financial services	Input value for missing data	Hazard models; Survival analysis	X	X
Kim and Yoon (2004) [29]	Korea mobile carriers	X	Binomial logit model	X	X
Au *et al.* (2003) [30]	Credit card and PBX databases	Attribute transformation	Decision tree; Neural Network	X	X

(Table 2) Contd….

Authors	Domain Datasets	Pre-Processing Methods	Prediction Techniques	Evaluation Methods	
				Type I/II Error	Cross Validation
Chiang *et al*. (2003) [31]	Network banking	Attribute transformation	Association Rule	X	X
Wei and Chiu (2002) [32]	Taiwan mobile company	X	Decision tree	X	V
Mozer *et al*. (2000) [33]	Wireless Telecom Industry	X	Neural Network; Logistic regression	X	V
Smith *et al*. (2000) [7]	Insurance company	Dimension reduction	Neural Network; Logistic regression; Decision tree	X	X

2.3. Discussions

Regarding Table **2**, there are various domain problems for customer churn prediction. Specifically, the telecommunication industry is the one major domain field approaching churn prediction. This is because, there are a huge number of accounts of wideband network in the world and the number of accounts is still increasing.

For data pre-processing, it can be seen that the main pre-processing task is feature selection and dimensionality reduction. It is the fact that the number of features (or variables) captured in a dataset is relatively large and not all these features are informative or can provide high discrimination power. However, not all of the related work considers pre-processing their selected dataset. The chosen dataset without any pre-processing is not likely to obtain the best prediction result.

On the other hand, the prediction techniques are mainly based on neural networks, decision trees, and some statistical methods. Very few consider other machine learning techniques, such as genetic algorithms, support vector machines, etc. This implies that there is still a space to improve the prediction performance by more advanced and sophisticated prediction techniques.

Specifically, some recent studies focus on developing hybrid prediction models (e.g. [17, 19]), which are based on combining multiple machine learning techniques. In general, multiple techniques are serially combined, in which the first component can be used to pre-process the given dataset for dimensionality reduction, outlier detection, or other tasks, and the second component is based on some classification techniques trained by the output of the first component for prediction.

Finally, for the evaluation methods, that are very critical to make the final conclusion of the employed data mining method, it is surprising that none of the related work examines the Type I/II errors. It is very important to assess how many churners are incorrectly classified by the prediction model into the non-churners group. Related work only considers the prediction accuracy rate for model validation. Moreover, quite a few studies perform cross validation. That is, many of them use some fixed numbers of training and testing datasets for experiments. This is likely to produce unreliable conclusion for the performance of the prediction models.

3. CURRENT & FUTURE DEVELOPMENTS

As customer churn prediction is a very important problem in customer relationship management, which aims at retaining valuable customers, the review of related work shows that several technical issues of churn prediction need to be further explored. First, for the data pre-processing step, it is unknown that which feature selection method performs the best by selecting the most representative features to make prediction models provide the highest rate of accuracy. That is, there are a number of feature selection methods, which can be applied for churn prediction, such as principal component analysis, genetic algorithms, decision trees, stepwise, etc. [34]. Besides feature selection, outlier detection and removal is another important pre-processing task, which aims at filtering out bad/noisy data that can degrade the prediction performances [35].

Second, some advanced machine learning techniques can be constructed to provide better prediction performances, for example, classifier ensembles (or multiple classifiers), hybrid classifiers, stacked generalization, etc.In short, they combine a number of different classifiers rather than single ones as used in the literature [10, 36].

CONFLICT OF INTEREST

None.

REFERENCES

[1] Berry MJA, Linoff G. Data mining techniques: For marketing, sales, and customer support. John Wiley & Sons: USA 2004.

[2] Ngai EWT, Xiu L, Chau DCK. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Syst Appl 2009; 36(2): 2592-2602.

[3] Van den Poel D, Larivie`re B. Customer attrition analysis for financial services using proportional hazard models. Eur J Oper Res 2004; 157(1): 196-217.

[4] Larivie`re B, Van den Poel D. Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Syst Appl 2005; 29(2): 472-484.

[5] Kim M, Park M, Jeong D. The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile

telecommunications services. Telecommun Policy 2004; 28(2): 145-159.

[6] Hung C, Tsai CF. Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. Expert Syst Appl 2008; 34(1): 780-787.

[7] Smith KA, Willis RJ, Books M. An analysis of customer retention and insurance claim patterns using data mining: A case study. J Oper Res Soc 2000; 51: 532-541.

[8] Aharoni, A., Portugaly, E.I., Oren, I.: US20090138304A1 (**2009**).

[9] Han J, Kamber M. Data mining: Concepts and techniques. 2nd ed. Morgan Kaufman: USA 2006.

[10] West D, Dellana S, Qian J. Neural network ensemble strategies for financial decision applications. Comp Oper Res 2005; 32(10): 2543-2559.

[11] Zhang G, Patuwo B, Hu M. Forecasting with artifiical neural networks: the state of the art. Int J Forecast 1998; 14(1): 35-62.

[12] Yang, Q., Gupta, Y., Wilson, K., Sedukhin, I.: EP1520237A2 (**2005**).

[13] Breiman L, Friedman JH, Olshen RA, Stone PJ. Classification and regression trees. Wadsworth International Group 1984.

[14] Cox DR, Snell EJ. Analysis of binary data. 2nd ed. Chapman and Hall: UK 1989.

[15] Hosmer DW, Lemeshow S. Applied logistic regression. Wiley: USA 1989.

[16] Glady N, Baesens B, Croux C. Modeling churn using customer lifetime value. Eur J Oper Res 2009; 197(1): 402-411.

[17] Pendharkar PC. Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services. Expert Syst Appl 2009; 36(3): 6714-6720.

[18] Xie Y, Li X, Ngai EWT, Ying W. Customer churn prediction using improved balanced random forests. Expert Syst Appl 2008; 36(3): 5445-5449.

[19] Tsai C-F, Lu Y-H. Customer churn prediction by hybrid neural networks. Expert Syst Appl 2009; 36(10): 12574-12553.

[20] Burez J, Van den Poel D. Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department. Expert Syst Appl 2008; 35(1-2): 497-514.

[21] Coussement K, Van den Poel D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Syst Appl 2008; 34(1): 313-327.

[22] Coussement K, Van den Poel D. Integrating the voice of customers through call center emails into a decision support system for churn prediction. Inform Manage 2008; 45(3): 164-174.

[23] Burez J, Van den Poel, D. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Syst Appl 2007*;* 32(2): 277-288.

[24] Chu B-H, Tsai M-S, Ho C-S. Toward a hybrid data mining model for customer retention. Knowl-Based Syst 2007*;* 20(8): 703-718.

[25] Luo B, Shao P, Liu D. Evaluation of three discrete methods on customer churn model based on neural network and decision tree in PHSS. *Proceeding of the 1st Inter Symp Data, Privacy and E-Commerce* Washington, DC: USA 2007; pp. 95-97.

[26] Hung S-Y, Yen DC, Wang H-Y. Applying data mining to telecom churn management. Expert Syst Appl 2006; 31(3): 515-524.

[27] Nie G, Zhang L, Li X, Shi Y. The analysis on the customers churn of charge email based on data mining-take one internet company for example. *Proceedings of IEEE Inter Conf Data Mining,* Washington, DC, USA 2006; pp. 843-847.

[28] Buckinx W, Van den Poel D. Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. Eur J Oper Res 2005; 164(1): 252-268.

[29] Kim H-S, Yoon C-H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommun Policy 2004; 28(9-10): 751-765.

[30] Au W-H, Chan KCC, Yao X. A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans Evol Comput 2003; 7(6): 532-545.

[31] Chiang D-A, Wang Y-F, Lee S-L, Lin C-J. Goal-oriented sequential pattern for network banking churn analysis. Expert Syst Appl 2003; 25(3): 293-302.

[32] Wei C-P, Chiu IT. Turning telecommunications call details to churn prediction: A data mining approach. Expert Syst Appl 2002; 23(2): 103-112.

[33] Mozer MC, Wolniewicz R, Grimes DB, Johnson E, Kaushansky H. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Trans Neural Netw 2000; 11(3): 690-696.

[34] Tsai C-F. Feature selection in bankruptcy prediction. Knowl-Based Syst 2009; 22(2): 120-127.

[35] Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. Mach Learn 2000*;* 38: 257-286.

[36] Wolpert DH. Stacked generalization. Neural Netw 1992; 5(2): 241-259.