Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Variable selection by association rules for customer churn prediction of multimedia on demand

Chih-Fong Tsai *, Mao-Yuan Chen

Department of Information Management, National Central University, Jhongli 32001, Taiwan

ARTICLE INFO

Keywords:
Customer churn prediction
Multimedia on demand
Data mining
Association rules
Neural networks
Decision trees

ABSTRACT

Multimedia on demand (MOD) is an interactive system that provides a number of value-added services in addition to traditional TV services, such as video on demand and interactive online learning. This opens a new marketing and managerial problem for the telecommunication industry to retain valuable MOD customers. Data mining techniques have been widely applied to develop customer churn prediction models, such as neural networks and decision trees in the domain of mobile telecommunication. However, much related work focuses on developing the prediction models per se. Few studies consider the pre-processing step during data mining whose aim is to filter out unrepresentative data or information. This paper presents the important processes of developing MOD customer churn prediction models by data mining techniques. They contain the pre-processing stage for selecting important variables by association rules, which have not been applied before, the model construction stage by neural networks (NN) and decision trees (DT), which are widely adapted in the literature, and four evaluation measures including prediction accuracy, precision, recall, and *F*-measure, all of which have not been considered to examine the model performance. The source data are based on one telecommunication company providing the MOD services in Taiwan, and the experimental results show that using association rules allows the DT and NN models to provide better prediction performances over a chosen validation dataset. In particular, the DT model performs better than the NN model. Moreover, some useful and important rules in the DT model, which show the factors affecting a high proportion of customer churn, are also discussed for the marketing and managerial purpose.

1. Introduction

Advances in computer and network technologies make the wideband network play a very important role in information community. This leads to a huge number of accounts of wideband network in the world and the number of accounts is still increasing.

As the streaming media technology is mature, transferring voice, video and other plentiful multimedia information in real time is widely used through the Internet today. Consequently, “video and voice pastimes” open the stage of living rooms. IPTV (Internet Protocol Television) breaks through the limit of the traditional video service-flow platform, which becomes the future trend of the global telecommunication industry (Papagiannidis, Berry, & Li, 2006).

Chunghwa Telecom,[1] for example, invests huge amount of capital in IPTV to establish interactive multimedia on demand (MOD) systems. MOD conveys digital programs through ADSL to TV and provides a number of value-added services in addition to traditional TV services, such as video on demand and interactive online learning. Therefore, it could be one of the best paragons of the telecommunication industry in video and voice pastimes.

Therefore, the telecommunication industry promoting the MOD services needs to concern some issues of customer relationship management (CRM). CRM mainly focuses on retaining or even cultivating profit potentials of customers (Hawkes, 2000). This leads to the importance of managing customer churn for the telecommunication industry providing the MOD services.

Data mining techniques have been widely used for customer churn prediction in various domains, such as mobile telecommunications (Ahn, Han, & Lee, 2006), the wireless carrier (Chu, Tsai, & Ho, 2007), and newspaper subscription (Coussement & Van den Poel, 2008).

In general, the data mining process is composed of two major steps, which are data pre-processing and classification/clustering. Data pre-processing (or attribute selection) is one of the most important steps in the data mining process. Its aim is to filter out redundant or irrelevant information from the original data. The classification or clustering step is for the task of prediction, estimation, etc. (Yang & Olafsson, 2006).

* Corresponding author. Tel.: +886 3 4227151; fax: +886 3 4254604.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

[1] http://www.cht.com.tw/CHTFinalE/Web/.

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.06.076

To retain valuable MOD customers or accurately predict MOD customer churn, the aim of this paper is to use association rules (Agrawal, Imielinski, & Swami, 1993) in the pre-processing stage to select more representative variables from the original ones to improve the later prediction performance. In addition, the prediction model is based on backpropagation neural networks and C5.0 decision trees, as two commonly used prediction models in customer churn prediction (Buckinx & Van den Poel, 2005).

This is the first study to use association rules for attribute selection in the MOD customer churn prediction problem. Therefore, the research question of this paper is whether using association rules can improve the prediction performance of the neural network and decision-tree models. That is, the model that combines backpropagation neural networks and decision trees, with association rules is compared respectively with the backpropagation neural networks and decision trees without using association rules in terms of their prediction performances.

The paper is organized as follows: Section 2 briefly describes the definition of customer churn and the data mining techniques used in this paper. In addition, related work is reviewed and some limitations of related work are also discussed. Section 3 presents our research methodology in order to answer the research question of this paper including data pre-processing, training, testing, and validation datasets, construction of the prediction models, and the evaluation measures. Section 4 shows the experimental results and discusses some important rules in the decision-tree model for future marketing and/or managerial strategies. Finally, the conclusion and future work are provided in Section 5.

2. Literature review

2.1. Customer churn

Many highly competitive organizations have understood that retaining existing and valuable customers is their core managerial strategy to survive in industry. This leads to the importance of churn management. Customer churn means that customers 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 customers (Hung & Tsai, 2008; Kim, Park, & Jeong, 2004).

Regarding Berry and Linoff (2004), customer churn in the telecommunications industry can be divided into voluntary and involuntary churners. Voluntary churn means that customers make a decision to terminate their service with the provider. On the other hand, involuntary churn means that the company (or service provider) withdraws the customers' service because of abuse of service, non-payment of service, etc.

Therefore, customer churn in this paper is based on customers who have rent the MOD service for a period of time and then they either terminate that service or are withdrawn by the company due to non-payment of service.

2.2. Data mining techniques

The aim of data mining (or sometimes called knowledge discovery) is to extract useful information and unknown knowledge from data. It is a process of discovering various models, summaries, and derived values from a given collection of data. It usually consists of data selection, data cleaning, data transformation and reduction, mining, interpretation and evaluation, and model deployment for decision support (Kantardzic, 2003).

The following describes the commonly used data mining techniques, which are association rules, neural networks, and decision trees.

2.2.1. Association rules

Association rules are one of the major data mining techniques. They are used to discover multiple independent elements that co-occur frequently and to discover rules that relate to the co-occurred elements in a given dataset (Agrawal et al., 1993).

A well-known application of association rules is market basket analysis. A market basket contains purchasing transactions of customers. That is, it is a collection of items or itemsets, which are purchased by a customer in a single transaction. As the number of customer transactions is usually very large and frequent itemsets are exponential to the number of different items, association rules can be used to examine as many frequent itemsets as possible. Questions like what products tend to be purchased together can be answered. Therefore, the task of using association rules is to reduce a large amount of information to a small and more understandable set of statistically supported statements (Kantardzic, 2003). For example, customers who bought product *A* will also buy product *B* with 81.25% probability. On the other hand, customers who bought product *B* will also buy product *A* with 65% probability.

The general concept of association rules is as follows: let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items in a given dataset *DB*, and *DB* contains a set of transactions, where each transaction *T* is a set of items such that $T \subseteq I$. Let *X* be a set of purchased items. A transaction *T* is said to contain *X* if and only if $X \subseteq T$. An association rule can be represented by the form $X \Rightarrow Y$, where $X \subseteq I, Y \subseteq I$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ holds in the transaction set *DB* with *confidence c* if *c* % of the transactions contain *X* as well as *Y*. The rule $X \Rightarrow Y$ has *support s* in the transaction set if *s* % of the transactions in *DB* contain $X \cup Y$. Confidence means the strength of the rule, and support indicates the frequency of the patterns occurring in the rule. As a result, rules with high confidence and strong support can be referred to as *strong rules* (Kantardzic, 2003).

In Sohn and Kim (2003), association rules are used to analyze mobile customer patterns in order to identify a number of potential customer groups for further marketing strategies.

2.2.2. Neural networks

Neural networks (or artificial neural networks) learn by experience, generalize from previous experiences to new ones, and can make decisions. They have been extensively used to solve many real-world problems (Haykin, 1999). Regarding the study by Wong, Bodnovich, and Selvi (1997), approximately 95% of reported neural network business applications utilize multi-layer perceptron (MLP) neural network with the back propagation learning algorithm.

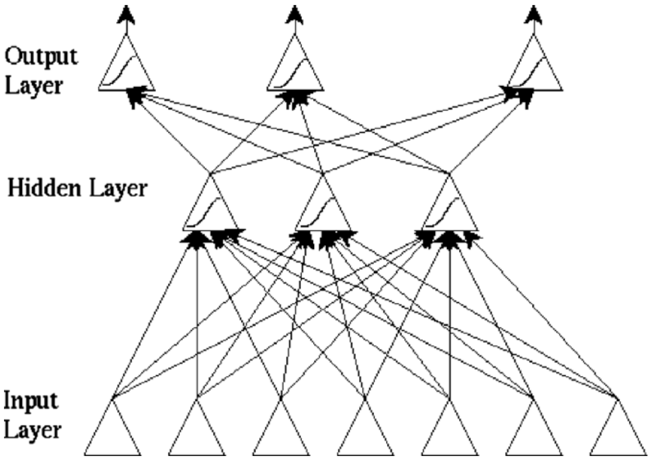


Fig. 1. The three-layer neural network.

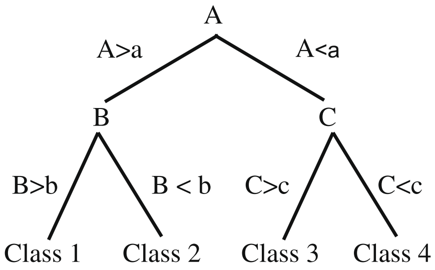


Fig. 2. Example of a decision tree.

Fig. 1 shows an example of a three-layer neural network, which consists of an input layer including a set of sensory nodes as input nodes, one or more hidden layers of computation nodes, and an output layer of computation nodes. The input nodes/neurons are the feature values of an instance, and the output nodes/neurons represent a discriminator between its class and all of the other classes. That is, each output value is a measure of the network's confidence that the class corresponding to the highest output value is returned as the prediction for an instance. Each interconnection has associated with a scalar weight which is adjusted during the training phase.

The backpropagation learning algorithm performs weights tuning to define whether or not hidden unit representation is most effective at minimising the error of misclassification. That is, for each training example, its inputs are fed into the input layer of the network and the predicted outputs are calculated. The differences between each predicted output and the corresponding target output are calculated. This error is then propagated back through the network, and the weights between the two layers are adjusted so that if the training example is presented to the network again, then the error would be less. As a result, the algorithm captures properties of the input instances, which are most relevant to learning the target function (Haykin, 1999).

2.2.3. Decision trees

A decision-tree model classifies an instance by sorting it through the tree to the appropriate leaf node, i.e. each leaf node represents a classification. Each node represents some attribute of the instance, and each branch corresponds to one of the possible values for this attribute (Mitchell, 1997). Fig. 2 shows an example of a decision tree. This model can make a decision for the four-class classification problem based on three different variables (A, B, and C).

2.3. Summary of related work

In Table 1, related work of data mining applications in customer churn prediction and in the telecommunications industry is summarized and compared in terms of their techniques used and problem domains. Note that we do not consider studies focusing on the problem of customer segmentation (or market segmentation) since their goals are to identify potential customer groups that are different from churn prediction.

Regarding the comparison of related work, there are some implications:

- Neural networks and decision trees have been considered and have shown their applicability in customer churn prediction in the telecommunication industry.
- The target of these case studies for telecommunication customer churn prediction is mostly mobile-based customers. However, there is no an empirical study of examining the performance of data mining techniques for predicting MOD customer churn.
- Many related studies focus on the step of developing prediction models and model evaluation in the data mining process, but few consider the data pre-processing step using some data mining technique. That is, using some related technique, e.g. association rules to pre-process data may be able to improve the final performance of prediction models.
- As data pre-processing is an important step in the data mining process to extract useful and representative features from the original data, association rules have not been used in the data reduction step in the literature.

Therefore, the aim of this paper is to examine whether association rules can be adapted in the data pre-processing stage to reduce a large amount of information to a small and more understandable data variables in order to improve the prediction performance of using neural networks and decision trees as the prediction models. In addition, the combination of the data reduction and model development steps using data mining techniques is investigated for the problem of MOD customer churn prediction.

3. Research methodology

3.1. The dataset

3.1.1. Overview

The case company is one of the telecommunication companies that provides MOD services in Taiwan. The dataset is composed of 37,882 MOD customer data from 2005/10/01 to 2006/12/21 (about 15 months in total). In addition, there are 22 different variables in the dataset, which are listed in Table 2.

3.1.2. Data pre-processing

The aim of this data pre-processing stage is to consider data cleaning for missing values and noisy data and data transformation if any.

Table 1
Summary of related work.

Work	Domain	Techniques
Luo et al. (2007)	Personal handy-phone system service (PHSS)	Neural networks and decision trees
Estévez et al. (2006)	Subscription fraud prevention in telecommunications	Fuzzy logic and neural networks
Hung et al. (2006)	Wireless telecommunications	Neural networks and decision trees
Qi et al. (2006)	Mobile telecommunication	Decision trees
Buckinx and Van den Poel (2005)	Retailing	Neural networks
Weiss (2005)	Fraud detection and network fault isolation in telecommunications	Neural networks and decision trees
Chiang et al. (2003)	Network banking	Association rules
Neelakanta and Preechayasomboon (2002)	Subchannel allocation in telecommunication subscriber access lines	Neural networks
Wei and Chiu (2002)	Mobile telecommunication	Decision trees
Berson et al. (2000)	Mobile telecommunication	Decision trees

Table 2
List of the 22 variables.

Variables	Description
• MODID	To identify registered MOD customers
• BILLID	To identify the branches providing the MOD services
• STARTDATE	The date to rent the MOD services
• ENDDATE	The date to terminate the MOD services
• STATUS	To record either churn or not
• BIRTHDAY	Customer's birthday
• ACCEPTEDM	Customers who either accept product promotion or not
• MODUSERTYPE	E.g., personal, school, or company usage (eight different purposes)
• MODISPFLAG	To identify either dedicated MOD or not
• MODPAYTYPE	The method of paying the MOD services (four types of payment)
• IDNO	national social security no. for individuals, unified business no. for companies, etc.
• PROM	Whether the services are recommended by the employee or not
• APPLYCLASS	E.g., military, official accommodation, retire consultant, etc. (31 different classes)
• CUSTOMERTYPE	Three classes including (1) apartment house, (2) business, and neither (1) and (2)
• USPEED	The speed of uploading data
• DSPEED	The speed of downloading data
• DISCOUNTCODE	To record the discount method
• BTYPE	Broadband services for either ADSL or FTTB
• LINELENGTH	The broadband distance from the MOD server to the customer
• OPENTIMES	The times of logging on MOD
• BASICPAY	The basic fee for the MOD services
• MODPAY	Extra fees for customer's preferences

- MODID and IDNO: as they both can be used to identify unique customers and they are duplicated, IDNO is not considered in this experiment.
- BIRTHDAY: this variable can be used to transform into AGE information. Thus, BIRTHDAY is deleted, but AGE is included as a new variable in this dataset.
- Similarly, STARTDATE and ENDDATE can derive a new variable, USEDAY, to represent the duration of renting the MOD services. Therefore, STARTDATE and ENDDATE are deleted.
- Finally, some data that are inconsistent, e.g. BILLID does not provide the MOD services-are excluded.

Therefore, the processed dataset contains 37,672 data and 18 different variables for our experiment.

3.1.3. Training, testing, and validation datasets

As there is no exact answer about the proportion of the training and testing data to construct the optimal prediction model, we examine eight different proportions for training and testing the model, which are shown in Table 3.

Note that these training and testing data are randomly selected. Table 4 shows the result of using the Z test and chi-squared test in order to ensure that the proportion of customer churn is the same as the population.

Table 3
The proportion of the training and testing data.

Dataset	Training proportion (%)	Testing proportion (%)
1	50	50
2	55	45
3	60	40
4	65	35
5	70	30
6	75	25
7	80	20
8	85	15

Table 4
The Z test and chi-squared test of the eight datasets.

Dataset	Training/testing	No. of samples	Z test	Chi-squared test
1	Training set	18,836	−0.94615	0.76690
	Testing set	18,836	0.94615	0.76690
2	Training set	20,720	0.30717	0.08083
	Testing set	16,952	−0.33959	0.09880
3	Training set	22,604	0.73054	0.45720
	Testing set	15,068	−0.89477	0.68587
4	Training set	24,487	0.23019	0.04539
	Testing set	13,185	−0.31370	0.08430
5	Training set	26,371	−0.62189	0.33131
	Testing set	11,301	0.94998	0.77313
6	Training set	28,254	0.53907	0.24895
	Testing set	9418	−0.93370	0.74685
7	Training set	30,138	−0.02067	0.00037
	Testing set	7534	0.04134	0.00146
8	Training set	32,022	−0.21155	0.03834
	Testing set	5650	0.50363	0.21729

In addition to the testing dataset for model evaluation, we also consider a validation dataset that is composed of 11,722 data from 2006/12/22 to 2007/04/30 (about 4 months) to further evaluate the prediction models. This new unknown dataset can be regarded as the 'real-world' case to validate the prediction model.

3.2. Attribute selection by association rules

Using association rules, different confidence and support values are examined. In this paper, we only report the best results that are based on the confidence value = 1 and the support value = 0.03. These rules are ranked as their importance by

$$\text{Importance } A \Rightarrow B = \log\left(\frac{P(A/B)}{P(A)}\right) \qquad (1)$$

In particular, the rule that has the importance value less than 0.8 is excluded. As a result, the extracted rules that relate to the important attributes are listed in Table 5. Note that STATUS = F means customer churn.

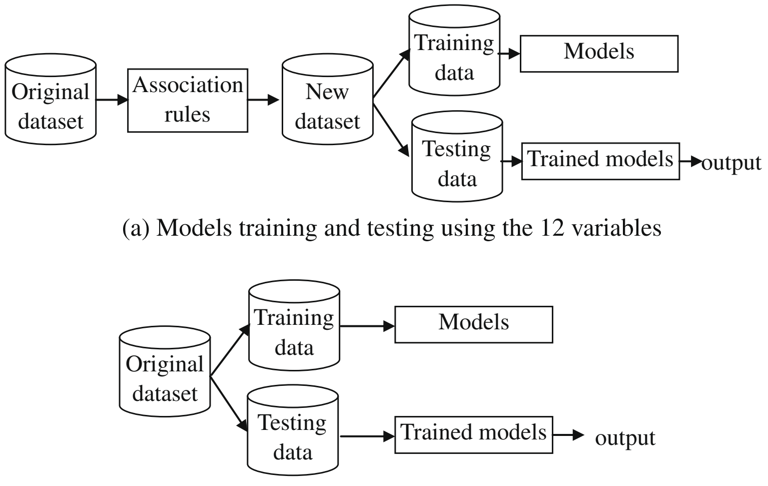
Therefore, the new dataset containing the 12 different variables out of 22 is used to compare the original dataset containing the 22 different variables by neural networks and decision trees in terms of their prediction performance.

3.3. Prediction models construction

To construct the prediction models, it contains the training and testing stage. Fig. 3a and b show the model construction process with and without considering association rules in the data pre-processing stage, respectively.

Table 5
The important association rules.

Importance	Rule
0.869	USEDAY < 116, OPENRATE = 0.025594591484375 – 0.138675599775 ⇒ STATUS = F
0.856	DISCOUNTCODE = MD29, USEDAY = 116–207 ⇒ STATUS = F
0.841	DISCOUNTCODE = SV44, EMP = 0 ⇒ STATUS = F
0.838	DISCOUNTCODE = SV37, AGE = 35–40 ⇒ STATUS = F
0.836	DISCOUNTCODE = SV37, ACCEPTEDM = 3 ⇒ STATUS = F
0.834	DISCOUNTCODE = SV37, SEX = 2 ⇒ STATUS = F
0.829	OPENRATE > = 0.43024975835, USPEED = 640K ⇒ STATUS = F
0.829	OPENRATE > = 0.43024975835, DSPEED = 8M ⇒ STATUS = F
0.826	DISCOUNTCODE = SV44, LINELENGTH < 1006 ⇒ STATUS = F
0.825	DISCOUNTCODE = SV37, BASICPAY < 71 ⇒ STATUS = F
0.821	DISCOUNTCODE = SV37, FIXLINEPAYTYPE = 4 ⇒ STATUS = F



(a) Models training and testing using the 12 variables

(b) Models training and testing using the original 22 variables

Fig. 3. Training and testing the prediction models.

Table 6
Confusion matrix.

		Predicted	
		Class 1	Class 2
Actual	Class 1	A	C
	Class 2	D	B

Table 7
Prediction performance by the testing dataset (22 variables).

Dataset	Model	Status	Precision (%)	Recall (%)	Accuracy (%)	*F*-measure
1	DT	F	94.70	77.47	96.21	0.852
		U	96.41	99.29		
	NN	F	86.42	79.35	95.33	0.827
		U	96.66	97.95		
2	DT	F	91.70	76.84	95.71	0.836
		U	96.26	98.84		
	NN	F	86.27	78.58	95.17	0.822
		U	96.50	97.92		
3	DT	F	92.32	78.17	96.01	0.847
		U	96.51	98.93		
	NN	F	86.70	79.30	95.37	0.828
		U	96.66	98.01		
4	DT	F	93.23	81.41	96.51	0.869
		U	96.98	99.02		
	NN	F	86.07	79.01	95.19	0.824
		U	96.56	97.88		
5	DT	F	92.37	78.31	95.88	0.848
		U	96.37	98.89		
	NN	F	87.38	77.40	95.05	0.821
		U	96.20	98.08		
6	DT	F	89.10	86.80	96.67	0.879
		U	97.86	98.27		
	NN	F	86.64	81.18	95.61	0.838
		U	96.97	97.96		
7	DT	F	89.05	77.52	95.41	0.829
		U	96.31	98.40		
	NN	F	84.59	74.65	94.41	0.793
		U	95.84	97.72		
8	DT	F	91.28	82.62	96.32	0.867
		U	97.08	98.65		
	NN	F	86.77	79.71	95.27	0.831
		U	96.59	97.93		
Avg.	DT	F	91.72	79.52	96.09	0.853
		U	96.72	86.5		
	NN	F	86.36	78.65	95.18	0.823
		U	96.5	97.93		

Table 8
Prediction performance by the validation dataset (22 variables).

Dataset	Model	Status	Precision (%)	Recall (%)	Accuracy (%)	*F*-measure
1	DT	F	49.92	55.55	75.83	0.526
		U	85.34	82.28		
	NN	F	20.67	30.41	55.05	0.246
		U	73.97	62.89		
2	DT	F	83.55	69.31	89.30	0.758
		U	90.74	95.66		
	NN	F	29.76	39.92	62.77	0.341
		U	78.57	70.04		
3	DT	F	74.45	56.26	84.79	0.641
		U	87.09	93.86		
	NN	F	18.67	33.73	48.57	0.240
		U	71.66	53.28		
4	DT	F	69.23	55.06	83.25	0.613
		U	86.58	92.22		
	NN	F	21.46	32.85	54.79	0.260
		U	74.31	61.77		
5	DT	F	83.70	69.38	89.35	0.759
		U	90.77	95.70		
	NN	F	22.93	39.50	53.37	0.290
		U	75.02	57.78		
6	DT	F	85.66	75.21	90.98	0.801
		U	92.41	96.00		
	NN	F	18.45	39.04	43.65	0.251
		U	69.95	45.12		
7	DT	F	79.70	49.96	84.86	0.614
		U	85.78	95.95		
	NN	F	26.20	43.56	56.78	0.327
		U	77.26	60.98		
8	DT	F	88.72	60.64	88.65	0.720
		U	88.63	97.55		
	NN	F	25.91	43.25	56.48	0.324
		U	77.08	60.68		
Avg.	DT	F	76.87	55.05	85.88	0.679
		U	88.42	93.65		
	NN	F	23.01	37.78	53.93	0.285
		U	74.73	59.07		

The prediction models are based on neural networks (NN) and decision trees (DT). For the NN model, it is trained by the back-propagation learning algorithm because of its broad applicability to many business problem domains (Smith & Gupta, 2000). Moreover, as constructing the NN model needs to setup some parameters, such as the training epoch and, numbers of the hidden layer's nodes, which are problem dependent, they are set by the default value[2] to avoid the over-fitting problem (Haykin, 1999). On the other hand, for the DT model, the C5.0 algorithm is chosen as they are widely used in related work (e.g. Hung, Yen, & Wang, 2006; Luo, Shao, & Liu, 2007).

3.4. Evaluation methods

In this paper, we consider prediction accuracy, precision, recall, and *F*-measure as the evaluation methods to examine the performance of the prediction models. By using a confusion matrix shown in Table 6, these evaluation measures can be obtained as follows:

- Accuracy = $\frac{A+B}{A+B+C+D}$
- Precision for Class 1 $= \frac{A}{A+C}$; Precision for Class 2 $= \frac{B}{B+D}$
- Recall for Class 1 $= \frac{A}{A+D}$; Recall for Class 2 $= \frac{B}{B+C}$
- *F*-measure = $\frac{2\times Precision\times Recall}{Precision+Recall}$

4. Experimental results

4.1. Prediction performance by the 22 variables

Table 7 shows the prediction performance of decision trees (DT) and neural networks (NN) based on the 22 variables.

The results show that the sixth dataset containing 75% training data and 25% testing data allows the DT and NN models to provide the best prediction performance (which are 96.67% and 95.61% for accuracy and 0.87932 and 0.8382295 for *F*-measure, respectively). Furthermore, Table 8 shows the performance by considering the validation set to evaluate the NN and DT models.

Using the validation dataset can further examine the reliability of the prediction models. It is interesting that the DT model still can provide relatively good prediction performances (i.e. 90.98% and 0.8009791 for accuracy and *F*-measure, respectively). However, the NN model does not perform very well if compared with the result of NN in Table 7. This means that the NN model could not perform stably for predicting new unknown data over the MOD customer churn problem.

4.2. Prediction performance by the 12 variables

After selecting the 12 representative variables by association rules, the DT and NN models are evaluated in terms of their prediction performances. Tables 9 and 10 show the results of using the testing and validation datasets, respectively.

[2] In this paper, we used the Microsoft SQL Server 2005 software for training and testing the DT and NN models.

Table 9
Prediction performance by the testing dataset (12 variables).

Dataset	Model	Status	Precision (%)	Recall (%)	Accuracy (%)	*F*-measure
1	DT	F	87.49	78.03	95.33	0.825
		U	96.46	98.17		
	NN	F	83.78	76.07	94.55	0.797
		U	96.13	97.58		
2	DT	F	91.82	76.76	95.72	0.836
		U	96.24	98.87		
	NN	F	84.09	76.18	94.56	0.799
		U	96.11	97.61		
3	DT	F	86.05	77.65	95.08	0.816
		U	96.40	97.94		
	NN	F	83.95	75.72	94.54	0.796
		U	96.09	97.63		
4	DT	F	89.35	78.21	95.57	0.834
		U	96.46	98.45		
	NN	F	83.69	77.36	94.63	0.804
		U	96.29	97.50		
5	DT	F	91.14	78.91	95.79	0.846
		U	96.46	98.68		
	NN	F	85.09	74.80	94.39	0.796
		U	95.76	97.75		
6	DT	F	87.80	81.94	95.88	0.848
		U	97.09	98.15		
	NN	F	83.08	74.13	94.27	0.783
		U	95.86	97.54		
7	DT	F	90.50	77.52	95.61	0.835
		U	96.32	98.64		
	NN	F	84.15	75.12	94.40	0.794
		U	95.91	97.63		
8	DT	F	89.34	80.44	95.75	0.847
		U	96.72	98.36		
	NN	F	82.48	76.67	94.23	0.795
		U	96.07	97.22		
Avg.	DT	F	89.19	78.68	95.59	0.836
		U	96.27	87.16		
	NN	F	83.79	75.76	94.45	0.796
		U	96.02	97.56		

In order to compare the model performance shown in Tables 7–10, Figs. 4 and 5 show the best prediction performance of the models using the 22 variables and the 12 variables over the testing and validation datasets, respectively. Note that the number in the bracket means the *i*th dataset out of the eight different proportions of the training and testing sets. In addition, the number followed by the model means the number of variables for model training and testing.

The comparative results show that the best DT and NN models trained by the original 22 variables slightly perform better than the ones trained by the 12 variables based on the testing dataset. However, the DT and NN models trained by the 12 variables provide better prediction performances based on the validation dataset, which is more robust to face real-world problems.

Moreover, DT-12 can provide the best performance by using the 55% and 45% training and testing data (i.e. dataset 2) without using larger training data (like dataset 6 for 75% and 25% training and testing data) to construct the model that is more efficient during training.

On the other hand, it is necessary to compare the DT and NN models for the performance of accurately predicting the customer churn group. That is, the prediction performance for 'STATUS' = F should be examined. Therefore, Figs. 6 and 7 further examine the average precision and recall performances of using the DT and NN models based on the 22 and 12 variables to predict the customer churn group over the testing and validation datasets, respectively.

Regarding Figs. 6 and 7, from the testing dataset to the validation dataset the prediction performances of the DT and NN models using the 22 variables degrade more than the models using the 12 variables.

In summary, as the validation dataset is more difficult for accurate prediction, which could be thought of as real-world cases, the DT and NN models followed by association rules can perform better than the single usage of the DT and NN models. In addition, the experimental results show that the DT models perform better than the NN models. For accurately predicting the churn group, the DT-12 and NN-12 models also perform better than the DT-22 and NN-22 ones.

4.3. The decision rules

As the DT models using the 12 variables perform the best, a number of decision rules can be obtained for customer churn prediction. The original DT model contains nine levels of hierarchy and 90 nodes, which is a very complex tree structure. In order to make the tree structure easy to read and use, i.e. to filter out unnecessary rules without affecting the prediction performance, tree pruning is considered.

There are two important parameters for decision-tree pruning:

- Minimum support. This focuses on the rules containing minimum cases (i.e. data samples) can be excluded. After a number of examinations, we set the minimum support

Table 10
Prediction performance by the validation dataset (12 variables).

Dataset	Model	Status	Precision (%)	Recall (%)	Accuracy (%)	*F*-measure
1	DT	F	50.51	52.44	76.13	0.515
		U	84.69	83.66		
	NN	F	43.44	39.60	72.99	0.414
		U	81.32	83.61		
2	DT	F	93.77	78.22	93.49	0.853
		U	93.42	98.35		
	NN	F	50.16	43.35	75.94	0.465
		U	82.73	86.31		
3	DT	F	87.13	37.59	83.60	0.525
		U	83.19	98.23		
	NN	F	58.93	46.32	79.26	0.519
		U	84.02	89.73		
4	DT	F	93.21	36.88	84.12	0.529
		U	83.17	99.15		
	NN	F	50.22	40.91	75.96	0.451
		U	82.26	87.10		
5	DT	F	75.80	54.28	84.79	0.633
		U	86.67	94.49		
	NN	F	44.03	49.01	72.67	0.464
		U	83.18	80.19		
6	DT	F	93.69	70.90	91.83	0.807
		U	91.41	98.48		
	NN	F	51.95	38.12	76.57	0.440
		U	81.86	88.79		
7	DT	F	68.82	50.42	82.53	0.582
		U	85.47	92.74		
	NN	F	52.58	34.55	76.69	0.417
		U	81.23	90.09		
8	DT	F	58.95	49.61	79.51	0.539
		U	84.75	89.02		
	NN	F	38.52	39.99	70.12	0.392
		U	80.69	79.71		
Avg.	DT	F	77.74	53.8	84.5	0.623
		U	86.6	94.27		
	NN	F	48.73	41.48	75.03	0.445
		U	82.16	85.69		

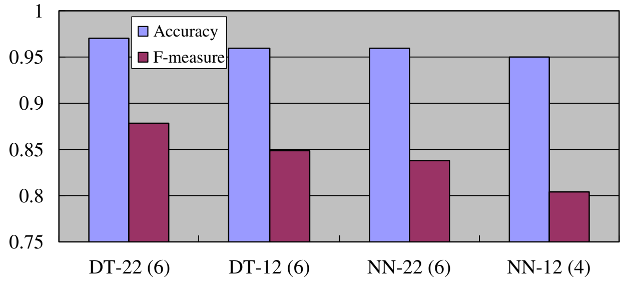


Fig. 4. Comparison of DT and NN using the 22 and 12 variables over the testing dataset.

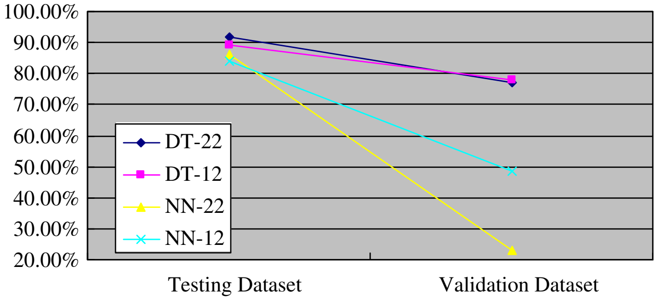


Fig. 6. Precision of DT and NN using the 22 and 12 variables over the testing dataset.

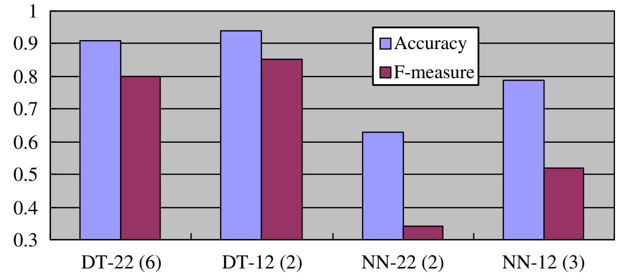


Fig. 5. Comparison of DT and NN using the 22 and 12 variables over the validation dataset.

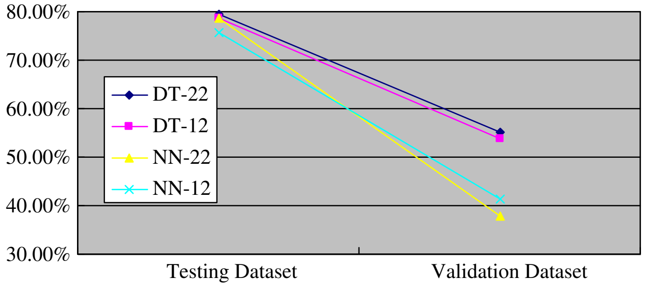


Fig. 7. Recall of DT and NN using the 22 and 12 variables over the testing dataset.

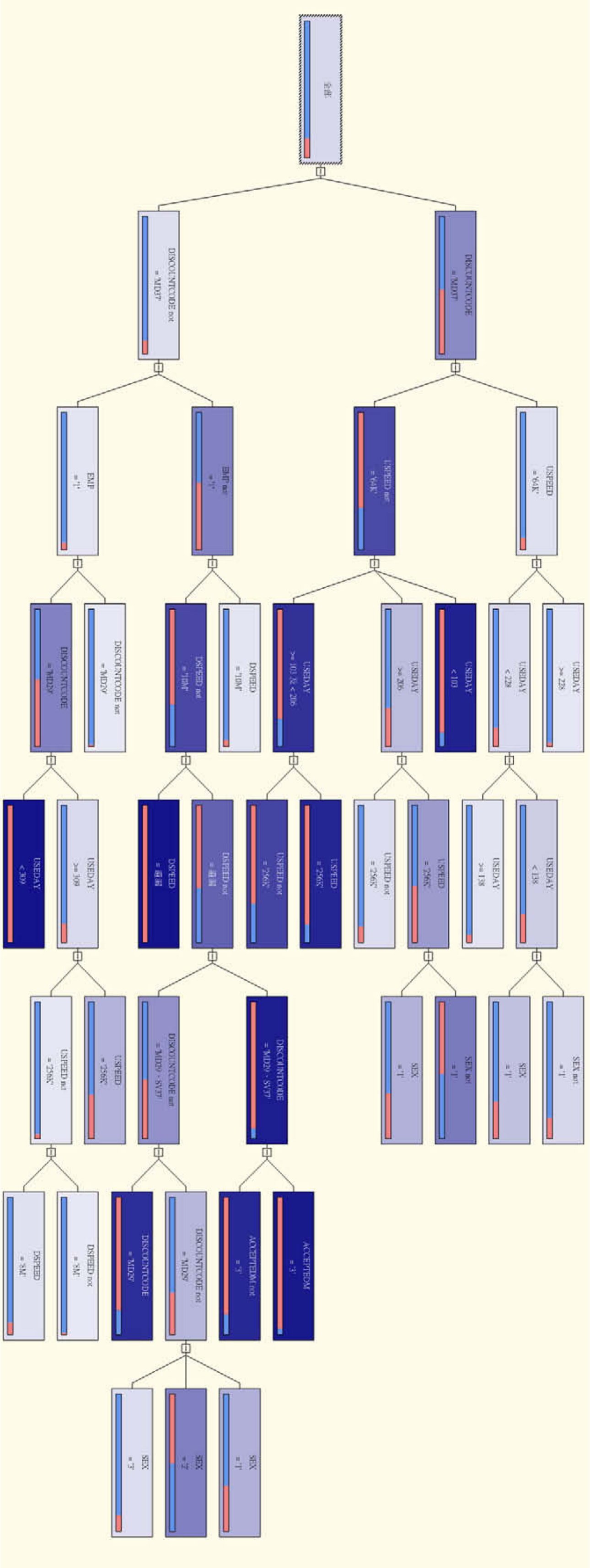


Fig. 8. The decision tree structure.

for 100, i.e. to delete the rule that contains less than 100 cases. The result does not make the prediction performance different.

- Score method. There are two methods to create the tree branches, which are the entropy and Bayesian methods. We found that using the entropy method can provide the best prediction performance.

After tree pruning, the new DT model, which still provides the same prediction performance, contains only 8 levels of hierarchy and 43 nodes. Fig. 8 shows the structure of the new DT model.

Since the MOD customer churn prediction model has been constructed, we can find out the pattern of customer churn for marketing and/or managerial strategies. As we can see in Fig. 8 there are dark and gray nodes in the DT model. That is, the dark nodes mean higher proportions of customer churn. For example:

- The 3rd node in the 4th level of hierarchy: ''DISCOUNTCODE = 'MD37' and USPEED not ='64K' and USEDAY < 103''. There are 585 cases, in which 522 cases are churners. Therefore, we should find out whether the discount content of MD37 is problematic, which leads to 88.95% churners who do not use 64K for uploading speed and the rental duration is less than 103 days.
- The 1st node in the 5th level of hierarchy: ''EMP ='1' and DISCOUNTCODE ='MD29' and USEDAY < 309''. There are 586 cases, in which 583 cases are churners. The proportion of customer churn is 99.15%. The discount code'MD29' means that the MOD customers need to sign for 1-year contract in the case company. However, after the employee recommendation, why did the customers terminated the MOD services without using for 1 year. Therefore, the service provider needs to figure out this issue.
- The 5th node in the 6th level of hierarchy: ''EMP not ='1' and DSPEED not ='10M' and DSPEED not ='10M' and DISCOUNTCODE ='MD29, SV37' ''. There are 316 cases, in which 293 cases are churners. The proportion of customer churn is 92.16%. Although'SV37' is a promotion case, the company needs to understand why the customers who are not recommended by any employee and whose downloading speeds are not 10M terminate the MOD services.

Besides reviewing the factors affecting higher proportions of customer churn, we can forecast new MOD customers who may be the churners in the near future based on the prediction model. After identifying the future possible churners as the prediction output of the model, the company can then adopt some strategy to retain the potential churners.

5. Conclusion

It is very important for the telecommunication industry providing the MOD services to concern about one important issue of customer relationship management, i.e. churn prediction. Applying some important and well-known data mining techniques, such as association rules, neural networks, and decision trees, allows us to understand the importance of the pre-processing step during data mining for developing a better churn prediction model. In addition, to deeply understand the effectiveness of a prediction model, we also need to consider the performance of accurately predicting the customer churn group in addition to average prediction accuracy.

The experimental results show that the DT and NN models followed by association rules considered during the pre-processing step can provide reasonable better prediction performances over the validation dataset. In particular, the DT model performs better than the NN model. The DT model provides a number of predicted rules that allow the case company to think how to retain their MOD customers as well as to predict new customers who will be the churners.

In addition, the experimental results also provide some implications. For instance, some important variables that are selected by association rules, such as the type of discount and the times of logging on MOD, are the factors of customer churn. Therefore, the movie watching preferences of MOD customers, for example, can be considered based on cluster analysis. By grouping similar customers into several different clusters for different marketing strategies could enhance the probability of retaining customers.

For another future work, although this paper has developed four different customer churn prediction models and found that the DT model can provide a better performance, this model cannot predict who will terminate the MOD services next month or when. Therefore, time series and trend analysis should be considered in order to further improve the prediction performance.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993) Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD conference on management of data* (pp. 207–216).

Ahn, J.-H., Han, S.-P., & Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy, 30*(10–11), 552–568.

Berry, M. J. A., & Linoff, G. (2004). *Data mining techniques: For marketing, sales, and customer support*. John Wiley & Sons.

Berson, A., Simith, S., & Thearling, K. (2000). *Building data mining applications for CRM*. McGraw-Hill.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*(1), 252–268.

Chiang, D.-A., Wang, Y.-F., Lee, S.-L., & Lin, C.-J. (2003). Goal-oriented sequential pattern for network banking churn analysis. *Expert Systems with Applications, 25*(3), 293–302.

Chu, B.-H., Tsai, M.-S., & Ho, C.-S. (2007). Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems, 20*(8), 703–718.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327.

Estévez, P. A., Held, C. M., & Perez, C. A. (2006). Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. *Expert Systems with Applications, 31*(2), 337–344.

Hawkes, V. A. (2000). The heart of the matter: The challenge of customer lifetime value. *CRM Forum Resources*, 1–10.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). Prentice Hall.

Hung, C., & Tsai, C.-F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. *Expert Systems with Applications, 34*(1), 780–787.

Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*(3), 515–524.

Kantardzic, M. (2003). *Data mining – Concepts, models, methods, and algorithms*. John Wiley & Sons.

Kim, M., Park, M., & Jeong, D. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunications services. *Telecommunications Policy, 28*, 145–159.

Luo, B., Shao, P., & Liu, D. (2007) Evaluation of three discrete methods on customer churn model based on neural network and decision tree in PHSS. In *The first international symposium on data, privacy, and e-commerce* (pp. 95–97).

Mitchell, T. (1997). *Machine learning*. McGraw Hill.

Neelakanta, P. S., & Preechayasomboon, A. (2002). Development of a neuroinference engine for ADSL modem applications in telecommunications using an ANN with fast computational ability. *Neurocomputing, 48*(1–4), 423–441.

Papagiannidis, S., Berry, J., & Li, F. (2006). Well beyond streaming video: IPv6 and the next generation television. *Technological Forecasting and Social Change, 73*(5), 510–523.

Qi, J., Zhang, Y., Zhang, Y., & Shi, S. (2006). TreeLogit model for customer churn prediction. In *IEEE Asia-Pacific conference on services computing* (pp. 70–75).

Smith, K. A., & Gupta, J. N. D. (2000). Neural networks in business: Techniques and applications for the operations research. *Computers and Operations Research, 27*, 1023–1044.

Sohn, S. Y., & Kim, Y. (2003). Searching customer patterns of mobile service using clustering and quantitative association rule. *Expert Systems with Applications, 34*(2), 1070–1077.

Wei, C.-P., & Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications, 23*(2), 103–112.

Weiss, G. M. (2005). Data mining in telecommunications. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook*. Springer.

Wong, B. K., Bodnovich, T. A., & Selvi, Y. (1997). Neural network applications in business: A review and analysis of the literature (1988–1995). *Decision Support Systems, 19*, 301–320.

Yang, J., & Olafsson, S. (2006). Optimization-based feature selection with adaptive instance sampling. *Computers and Operations Research, 33*(11), 3088–3106.