Data Mining in Churn Analysis Model for Telecommunication Industry

Khalida binti Oseman¹
Sunarti binti Mohd Shukor¹
Norazrina Abu Haris¹
Faizin bin Abu Bakar¹
¹Operation Support Systems Division
E-mail: khalida@tmrnd.com.my, sunarti@tmrnd.com.my, azrina@tmrnd.com.my, faizin@tmrnd.com.my

ABSTRACT

In today’s competitive world, customer churn remains to be as one of the most pressing concerns for network providers. Through research, it is found that the cost of acquiring new customers is much higher than retaining the existing ones. However, predicting customer churn to tackle this main concern is an extremely difficult undertaking. There are a few approaches to predictive modelling for churn and these models are to be mapped to customers with a certain pattern of behaviours, profiling and product packaging via data mining processes. This paper briefly explains predictive modelling for customer churn based on data mining methods. Next, the utilization of one of its methods to perform predictive model, the Decision Tree Analysis model, is then described and demonstrated in details.

Keywords: Customer Relationship Management (CRM), customer churn, data mining, decision tree, information entropy, information gain

Introduction

The introduction of the wireless telecommunication and the wireless Internet broadband had changed the trend of the telecommunication industry in Malaysia to become highly competitive. Customers opt to use wireless network rather than fixed network due to the mobility freedom. Furthermore, customers have the freedom to choose and switch from one package to another package offered by the service provider, or from one service provider to another service provider. Thus, customer churn has become a major concern that is now harassing network providers in Malaysia. Through research, it is found that the cost of acquiring new customers is much higher than retaining the existing ones (Ali Tamaddoni Jahromi, 2009; Huang Ming, Niu Wenying, Liang Xu, 2009). Therefore, churn predictive study is one of the steps that can be taken to prevent loss in revenue. This study utilizes data mining techniques to build a model for churner prediction.

Data mining is a process of extracting and analysis of patterns, relationships and useful information from massive databases. It usually involves four classes of tasks which are the classification, clustering, regression and association rule learning [Gary Cokins and Ken King]. There are two main types of data mining: verification-oriented (the system verifies the user's hypothesis) and discovery-oriented (the system finds new rules and patterns autonomously) (Hongxia Ma, Min Qin and Jianxia Wang, 2009). The taxonomy of data mining methods can be summarized as Figure 1 (MO Zan, ZHOA Shan, LI Li and LIU Ai-Jun, 2007).
Verification method deals with evaluation of a hypothesis proposed by an external source (i.e. an expert etc.). This method includes the traditional statistical methods such as the goodness-of-fit test, t-test of means, and analysis of variance. These methods are less associated with data mining than their discovery-oriented counterparts because most data mining problems are concerned with selecting a hypothesis (out of a set of hypotheses) rather than testing a known one (MO Zan, ZHOA Shan, LI Li and LIU Ai-Jun, 2007). On the other hand, discovery methods are methods that identify patterns in the data automatically. The discovery method branch consists of prediction methods and description methods. Description-oriented data mining methods focus on understanding the way the underlying data operates while the prediction-oriented methods aim to build a behavioral model that can get newly and unseen samples and is able to predict values of one or more variables related to the sample. The classification methods arrange data into predefined group and can be divided into five (5) branches; neural networks, Naïve Bayesian networks, decision trees, support vector machine and instance based methods.

This paper concentrates on decision tree method which is a tree-shaped structure that represents sets of decisions or prediction of data trends (Lee. S., and Siau.K, 2001). It is suitable to describe sequence of interrelated decisions or prediction of future data trends and has the capability to classify entities into specific classes based on feature of entities (Michael J. Berry and Gordon S. Linoff ,2004)( Oded Z. Maimon and Lior Rokach ,2005). According to Tan, Steinbach and Kumar in their book, each tree consists of three types of nodes: root node, internal node and terminal node/leaf (Pang-Ning Tan, Michael Steinbach and Vipin Kumar, 2005). The topmost node is the root node and it represents all of the rows in the dataset. Nodes with child nodes are the internal nodes while nodes without child node are called the terminal node or leaf.
The Architecture/Process in Data Mining

Data mining methodology consists of ten (10) processes; translate a business problem into a data mining problem, select appropriate data, get to know the data, create a model set, fix the problems with the data, transform data, build models, assess models, deploy models and assess results. However, these steps can be further summarized into five main stages: the initial exploration incorporates with business and data understanding, data preparation, model building, evaluation and deployment. These are the entire steps taken to generate predictions. Figure 2 shows the flow chart of standard processes for data mining.

i. Business and Data Understanding

Here, the objectives and goals are defined by understanding the problems to obtain clearer objectives [Huang Ming, Niu Wenying and Liang Xu, 2009]. Next, the type or criteria of data which is available to us is determined in order to solve our objectives need. Data selected should represent enough quantity of data in a given period of time. Then, the hidden trends in data are discovered by finding the relation of the data selected [Shyam V.Nath and Dr. Ravi S. Behara, 2005].

ii. Data Preparation

Data preparation is the most time-consuming phase in data analysis or data mining processes. In data preparation phase, data is collected, integrated and cleaned. Integration of data may require extracting from multiple sources. Once, the data is in a tabular format, it should be fully characterized. Data needs to be cleaned by resolving any ambiguities, errors, and removing redundant and problematic data. Data which clearly do not contribute in the analysis are removed. Finally, the table should be divided, where appropriate, into subsets in order to optimize the performance of the database, simplify the analysis and allow specific queries to be performed easily.
iii. Modeling Phase

In this phase, we are required to develop a model for future prediction and select appropriate modeling technique to suite with the main objectives and evaluate them in the next phase. Once the model has been selected, different parameters are obtained to make improvement on the results.

iv. Evaluation Phase

This phase is the vital stage in CRISP-DM (Cross Industry Standard Process- Data Mining) model. Evaluation on the response time, confidence level, cost, error rate and the usefulness of the model in achieving the objectives and goals previously defined is done.

v. Deployment

This stage deploys the results through the objectives and goals previously defined at the initial process.

Decision Tree in Churn Analysis

Data mining is used in churn analysis to predict whether a particular customer will churn, when churn is expected to happen and reasons for churn. By predicting which customers are likely to churn, telecommunication companies can reduce the rate of churn by offering customers new incentives or packages to stay. By understanding reasons for churn, providers can improve their services and packages offered. Apart from that, it gives the best strategy for them in terms of cost and effort by decreased total cost of retention campaigns and increased the effectiveness campaigns. There are a few decision tree classification algorithms; C5.0, CART, ID3 and etc. (Song Danwa, Han Ning and Liu Dandan, 2009). This paper focuses on the ID3 algorithm. ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan [15]. There are two steps in classification of decision tree; first step, use the training set to establish a decision tree and second step, use the decision tree to make classification to input record (Song Danwa, Han Ning and Liu Dandan,2009).

Methodology for ID3 algorithm

The algorithm of ID3 introduces the concept of information gain for information theory. Consider the information gain as the measurement of the ability of the attribute classification decision to select the attribute of the node (Song Danwa, Han Ning and Liu Dandan,2009). The decision tree node generated by selecting the biggest information gain property and the different values for information gain property are used to set up the branch. The same process applied to establish the next node and branches. Entropy will measure the amount of information in an attribute.

Let set $E$ is a collection contains of samples, category can take $m$ different values correspond to different category $C_i$, $i = 1, 2, 3, ..., m$. If the selected test property is $A$ property, with $A$ has $v$ different values; $a_1, a_2, ..., a_v$, $A$ will divided the set $E$ into $v$ sub-sets; $E_{1}, E_{2}, ..., E_{v}$. Set $E_j$ is the sample size for the subset $E_j$ belongs to $C_i$. When $A$ is used to divide the current sample collection, the required information entropy is calculated as follows;
vi. Calculate the expected \( A \); 
\[
E(A) = \sum_{j=1}^{v} \frac{E_{ij} + \ldots + E_{mj}}{|E|} I(E_{1j}, E_{2j}, \ldots, E_{mj}).
\]
Then compute the information entropy for set \( E \); 
\[
I(E_{1}, E_{2}, \ldots, E_{m}) = -\sum_{j=1}^{v} \frac{|E_{j}|}{|E|} \log_{2} \frac{|E_{j}|}{|E|}.
\]

vii. For a given subset \( E_{j} \), the information entropy is; 
\[
I(E_{1j}, E_{2j}, \ldots, E_{mj}) = -\sum_{i=1}^{m} \frac{E_{ij}}{|E_{j}|} \log_{2} \frac{E_{ij}}{|E_{j}|}.
\]

\[
p_{ij} = \frac{E_{ij}}{|E_{j}|}, \quad p_{ij} \text{ is the probability of the samples of subset } E_{j} \text{ belong to category } C_{i}.
\]

viii. Get the information gain for using property \( A \); 
\[
Gain(A) = I(E_{1}, E_{2}, \ldots, E_{m}) - E(A)
\]

By using all the three above steps, the calculation of the information entropy for each property, selection of the biggest gain property as a test property for a given set \( E \) and generation of the branches and node can be done.

**Experimental Analysis**

Through data processing, the following properties are selected as an analysis sample; they are length of service, area and total of more than 10 minutes customer engaged. Data set divided into churn and not churns. The training sample data described previously is shown in the following table:

<table>
<thead>
<tr>
<th>Customer No</th>
<th>Length of Service &gt; 20 years</th>
<th>Area</th>
<th>Total of &gt; 10 Minutes</th>
<th>Churn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>False</td>
<td>Rural</td>
<td>FALSE</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>False</td>
<td>Sub-urban</td>
<td>TRUE</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>False</td>
<td>Rural</td>
<td>FALSE</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>False</td>
<td>Sub-urban</td>
<td>FALSE</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>False</td>
<td>Sub-urban</td>
<td>TRUE</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>True</td>
<td>Urban</td>
<td>TRUE</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>False</td>
<td>Rural</td>
<td>FALSE</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>True</td>
<td>Urban</td>
<td>TRUE</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>True</td>
<td>Sub-urban</td>
<td>TRUE</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>False</td>
<td>Rural</td>
<td>FALSE</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>True</td>
<td>Rural</td>
<td>FALSE</td>
<td>Yes</td>
</tr>
<tr>
<td>12</td>
<td>False</td>
<td>Sub-urban</td>
<td>TRUE</td>
<td>No</td>
</tr>
<tr>
<td>13</td>
<td>False</td>
<td>Sub-urban</td>
<td>FALSE</td>
<td>No</td>
</tr>
<tr>
<td>14</td>
<td>False</td>
<td>Rural</td>
<td>FALSE</td>
<td>Yes</td>
</tr>
<tr>
<td>15</td>
<td>False</td>
<td>Urban</td>
<td>TRUE</td>
<td>No</td>
</tr>
<tr>
<td>16</td>
<td>False</td>
<td>Sub-urban</td>
<td>TRUE</td>
<td>No</td>
</tr>
<tr>
<td>17</td>
<td>False</td>
<td>Sub-urban</td>
<td>TRUE</td>
<td>No</td>
</tr>
<tr>
<td>18</td>
<td>False</td>
<td>Sub-urban</td>
<td>FALSE</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Below are the calculations to formulate classification and construct decision tree by using ID3 algorithm:

ix. Information entropy calculation of the given sample set.
   From table 1, we notice that the set data belong to churn \((u_1)\) is 11 and not churn \((u_2)\) is 19. Then the probability of the samples of subset churn is \(p(u_1) = \frac{11}{30}\) and \(p(u_2) = \frac{19}{30}\). So the information entropy of given sample set is
   \[
   I(11,19) = - \frac{11}{30} \log_2 \frac{11}{30} - \frac{19}{30} \log_2 \frac{19}{30} = 0.948078230.
   \]

x. Calculate each property’s information entropy:
   First we calculate for the area property;
   ▪ for the rural area, 6 positive examples and 4 negative examples,
     \[
     I(6,4) = - \frac{6}{10} \log_2 \frac{6}{10} - \frac{4}{10} \log_2 \frac{4}{10} = 0.52877124
     \]
   ▪ for sub-urban area, 5 positive examples and 8 negative examples,
     \[
     I(5,8) = - \frac{5}{13} \log_2 \frac{5}{13} - \frac{8}{13} \log_2 \frac{8}{13} = 0.9612366
     \]
   ▪ for urban area, 0 positive examples and 7 negative examples,
     \[
     I(0,7) = 0
     \]
   So the information entropy of area property is
   \[
   E(area) = \frac{10}{30} I(6,4) + \frac{13}{30} I(5,8) + \frac{7}{30} I(0,7) = 0.592792941
   \]

Next, calculate the information entropy for length of service, area and total of > 10 minutes properties. The information entropy obtained for each property is;
\[
E(length\ of\ service > 20\ year) = 0.67039231\] and the information entropy for
\[
E(total\ of\ ≥10\ minutes) = 0.3122507.
\]
xi. Get the information gain for each property:

- Gain(area) = $I(11,19) - E(area) = 0.3552853$
- Gain (length of service > 20 years) = $I(11,19) - E(length of service > 20 years) = 0.2776859$
- Gain (total of > 10 minutes) = $I(11,19) - E(total of > 10 minutes) = 0.3122507$

It is clear that the property for area has the largest information gain, which gives the largest amount of information. So the area property should be selected as a classification property.

xii. Set up the decision tree.

Select the area property as a root node and this property have three value and heads to three branches. Then the training data sample is divided into three sub-sets to generate a decision tree which contains of three child nodes, shows in Figure 3:

![Figure 3: “Area” property as root node classification decision tree](image)

To obtain the other branches, repeat the procedure above. Finally, the whole decision tree is constructed (shown in Figure 4).

![Figure 4: Classification Decision Tree](image)

Through data classification and construction of decision tree, the highest churn customer factor can be discovered. From the decision tree analysis, the first classification attribute that contribute to churn is the
area of the subscribers, which is the main contribution factors. Furthermore, the area of the subscribers is related to the lengths of services and total of minutes for customer churn. Based on Figure 4 we can deduce that; if area is rural and length of service more than 20 years, subscribers are not likely to churn to other providers, and if area is sub-urban and the total of minutes that customer engages in line less than 10 minutes, the subscribers are likely to churn.

Conclusion

This paper introduced how to use classification rule decision tree of data mining to analyze customer churn factor in telecommunication industry. A specific training sample set is used to conduct an experiment of customer churn factor using decision tree. Based on the corresponding results as well as on decision tree analysis, rules factor of customer churn can be presented. Area is the main factor for customer to churn, apart from two minor causes for customer to churn. It is easily to understand the rule information by using the decision tree.

References


2. Huang Ming, Niu Wenying, Liang Xu(2009), *An Improved Decision Tree Classification Algorithm based on ID3 and the Application in Score Analysis*, Chinese Control and Decision Conference, 1876-1879.


6. MO Zan, ZHOA Shan, LI Li, LIU Ai-Jun. (2007).* A Predictive Model of Churn in Telecommunications Based on Data Mining*. IEEE International Conference on Control and Automation, Guangzhou, China


