A Customer Churn Analysis Model in E-business Environment

Wu Hengliang, Zhang Weiwei
School of Management Science and Engineering, Shandong Institute of Business and Technology, Shandong, Yantai, 264005, China
e-mail: hengliangwu@163.com, zwwc507@163.com

Abstract

Today, more and more companies in e-business have acknowledged that their business strategies should focus on identifying those customers who are likely to churn as markets become increasingly saturated. In this paper a new method was put forward to analyze and predict customer churn behavior based on data mining techniques, such as decision tree, clustering, neural network, etc. By collecting a large amount of customer churn questionnaire data, this paper established an e-business customer churn prediction and analysis model, analyzed the related factors which influence customer retention of e-business enterprises, and proposed many corresponding loss control measures.

Keywords: Customer churn model, E-business, Data mining

1. Introduction

With the dramatic growth of the Internet, e-business has boomed rapidly. In the past decade, e-business via the Internet has substantially affected the business world and will continue to be important. However, there also exist some obstacles to the success of e-business, such as customer churn. Today, more and more companies have acknowledged that their business strategies should focus on identifying those customers who are likely to churn as markets become increasingly saturated.

Customer churn refers to a behavior that a customer leaves his/her service company for some reason. Losing customers not only leads to opportunity costs because of reduced sales, but also leads to an increased need of attracting new customers. Research has shown that the cost of acquisition of a new customer is estimated to be ranging from $ 300 to $ 600 [1], and it costs roughly 5-6 times as much to sign on a new customer as to retain an existing one, while a 5% increase in customer retention results in a 25-95% increase in profits [2]. Therefore, e-business is not only about transactional activities but retention of customers as well.

To manage customer churn for companies effectively, it is important to build a customer churn prediction and analysis model. In literature, statistical techniques have been used to create the prediction models, such as logistic regression, naive bayesian classifiers, etc. But above methods have been limited in real application, such as the precision and generalization ability. Therefore, the necessity to explore new prediction approaches is strong and urgent.

2. Data mining

In order to establish a more effective and accurate churn model, many data mining methods have been recently considered. The aim of data mining 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. The following describes several major techniques which can achieve the goals of this research, including decision tree, clustering, neural network, etc.

2.1. Decision tree

Decision tree is a kind of data mining technology used for classification and prediction. 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. Decision tree development usually consists of two phases, tree building and tree pruning. The tree-building phase consists of recursively partitioning the training sets according to the values of the attributes. The partitioning process
continues until all, or most of the records in each of the partitions contain identical values. Certain branches may need to be removed because it is possible that they could consist of noisy data. The pruning phase involves selecting and removing the branches that contain the largest estimated error rate [7].

The famous algorithms of decision tree are C4.5, CART, and CHAID [8]. Although the algorithms are different, the purposes of them are alike. There are many advantages of the decision tree such that the rules resulted from the decision tree are readily understandable, the data can be classified with less calculation, and it can analyze the numeric or categorical data.

### 2.2. Clustering

The purpose of cluster analysis is to discover the natural grouping(s) of a set of patterns or objects. Objects can be customers, web documents, web users, or facilities. An operational definition of clustering can be stated as follows: Given a representation of n objects, find K groups based on a measure of similarity such that the similarities between objects in the same group are high while the similarities between objects in different groups are low [4]. Many researchers use clustering technique to segment customers and markets [5, 6]. The K-means clustering algorithm and the self-organizing map (SOM) are the two most popular clustering techniques.

The K-means clustering algorithm chooses K initial cluster centers randomly from N observations and assigns the remaining N-K observations to the nearest cluster based on the Euclidean distance. Then, the center of each cluster is re-calculated based on observations assigned to each cluster. The next step is to reassign observations to the nearest cluster and again to recalculate the center of each cluster, until no observation is reassigned to a new cluster [4].

### 2.3. Neural network

Neural network attempts to simulate biological neural systems which learn by changing the strength of the synaptic connection between neurons upon repeated stimulations by the same impulse [9]. They have been extensively used to solve many real-world problems.

Neural networks can be distinguished into single-layer perception and multilayer perception (MLP). The multilayer perception consists of multiple layers of simple, two taste, sigmoid processing nodes or neurons that interact by using weighted connections. Regarding the previous researches, approximately 95% of reported neural network business applications utilize MLP neural network with the back propagation learning algorithm [10].

In addition, the neural network contains one or more several intermediary layers between the input and output layers. Such intermediary layers are called hidden layers and nodes embedded in these layers are called hidden nodes. Based on prior research results, MLP is a relatively accurate neural network model [11, 12].

### 3. Customer churn model

The goal of this paper is to establish an effective customer churn model for e-business companies to retain customers based on the above data mining techniques. The establishing process includes four major steps: data collection, data pre-processing, customer classification, and model construction. A detailed description of establishing the model is listed in Figure 1.
3.1. Data collection

To establish the churn model, the first phase is to collect all kinds of relevant data, such as the basic information of customers, transaction records, and other data about customer perception, satisfaction, expectation, transfer cost, etc. We know that huge amounts of customer data have been recorded automatically in the e-business transactions, which become the basis of customer churn analysis. But sometimes it even needs questionnaire investigation to obtain some data.

3.2. 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. At the same time, the data collected above often include a large number of attributes, but only a small fraction will be used. Therefore, to improve the process performance and reduce the complexity of churn model, the attributes irrelevant to the model must be picked out. Here, this pre-processing can be done by using the relative importance analysis of neural network.

3.3. Customer classification

In some cases, customer classification is important for those companies engaged in e-business in response to the growing complexity in business markets. After all, it is not necessary to retain each customer, because the ultimate aim of customer churn prediction is to maximize the profit of companies. Therefore, before establishing, we should classify the customers depending on their value to find the high value and high churn probability customers so as to improve performance and efficiency of the churn model and find out really effective retention strategies. In this study, we mainly use clustering analysis technique to finish the customer classification.

3.4. Model construction

When data are ready, the final step is to choose appropriate data mining tools and techniques to establish churn model. In this study, the churn model is mainly established by two techniques, decision tree and neural network. In addition, a tool, called SPSS Clementine, is used with its C5.0 and Neural Net algorithms. Among them, C5.0 not only can produce the decision tree but also can generate its corresponding rule set. In a sense, rule set describes the information of its decision tree in a simplified or refined way to predict whether a customer will churn or not.

4. Empirical study

4.1. Questionnaire survey
A survey was developed to collect customer data from a group of customers who had internet shopping experiences in a certain e-business company of China. A total of 500 questionnaires were distributed, with 476 usable responses returned for a response rate of 95.2%. To establish the model as we proposed above, at first, we set several data processing rules as the following:

- If a customer’s consumption within 6 months was less than 2,000 yuan, we defined his/her current value was low, or it was high.
- If a customer had not increased his/her consumption in the past two years, we defined his/her potential value was low, or it was high.
- If a customer didn’t want to shop in the company website anymore, we thought he/she had lost, or he/she was active.
- If a customer thought an attribute (e.g. safety reliability) was important for the website, but he/she didn't think the website did well on it, we then defined his/her attitude to the website was negative, or it was affirmative.
- If the amount of attributes in which customer had affirmative attitude to a problem (e.g. trust problem) on the website was more than half of the total attributes, we then defined his/her attitude to the problem was negative, or it was affirmative.

Based on the above rules, we designed a lot of customer churn parameters, as shown in Table 1.

<table>
<thead>
<tr>
<th>Item</th>
<th>Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer churn</td>
<td>Churn</td>
<td>Symbol of churn or not</td>
</tr>
<tr>
<td>Customer value</td>
<td>Current_value</td>
<td>Current value of customer</td>
</tr>
<tr>
<td></td>
<td>Potential_value</td>
<td>Potential value of customer</td>
</tr>
<tr>
<td>Customer information</td>
<td>Sex</td>
<td>Customer’s sex</td>
</tr>
<tr>
<td></td>
<td>Marriage</td>
<td>Be married or not</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>Customer’s age</td>
</tr>
<tr>
<td></td>
<td>Child</td>
<td>Symbol of having children or not</td>
</tr>
<tr>
<td></td>
<td>R_c_mark</td>
<td>Symbol of rural or city</td>
</tr>
<tr>
<td></td>
<td>Qualification</td>
<td>Customer’s qualification</td>
</tr>
<tr>
<td></td>
<td>Occupation</td>
<td>Customer’s occupation</td>
</tr>
<tr>
<td></td>
<td>Revenue</td>
<td>Customer’s revenue</td>
</tr>
<tr>
<td>Customer attitude (affirmative or negative)</td>
<td>Trust</td>
<td>Attitude to credit problems</td>
</tr>
<tr>
<td></td>
<td>Perception</td>
<td>Attitude to perception problems</td>
</tr>
<tr>
<td></td>
<td>Satisfaction</td>
<td>Attitude to satisfaction problems</td>
</tr>
<tr>
<td></td>
<td>Expect</td>
<td>Attitude to expectation problems</td>
</tr>
<tr>
<td></td>
<td>Transfer_cost</td>
<td>Attitude to transfer cost problems</td>
</tr>
</tbody>
</table>

4.2. Establishing model for sample data

When finishing pre-processing to the sample data, a customer data mining database was established. Based on the above parameters, the authors segmented the 476 customers into four groups using clustering analysis with the k-means algorithm, including high current value and high potential value customers, low current value and high potential value customers, high current value and low potential value customers, low current value and low potential value customers. Among them, the former two were what the e-business companies should be most concerned about, this paper mainly focused on the first group (hereinafter referred to as high-value customers), which proportion accounted for 36% in total sample data.
According to the age of customers, high-value customer group was further subdivided into the following two groups: one was called high-value young customers, whose age were less than thirty, another was called high-value old customers, whose age were more than thirty, their proportions were 65% and 35%. After that, by using the C5.0 algorithm two customer churn models were established for the above groups respectively, as shown in Figure 2 and Figure 3. With the aid of the models, it became easier for e-business companies to judge whether a high-value customer might be lost, if yes, something should be done to retain him/her, e.g., putting more resources or improving the quality of service, etc.

Figure 2. Decision tree rules of high-value young customers

Figure 3. Decision tree rules of high-value old customers

In addition, we analyzed the relative importance of five categories of problems for high-value young and old customer groups with the Neural Net algorithm in SPSS Clementine, their calculation results were shown in Figure 4 and Figure 5. The results indicated that, whether young or old customers, they all thought trust was the most important problem in e-business, particularly to old customers.

Moreover, there was a significant difference between the two groups, young customers cared a lot about perception of e-business but old customers paid more attention to the transfer cost in e-business. This told us young customers preferred things those could be obviously felt, while old customers were more willing to form a kind of strong emotional bond.
4.3. Analysis of influence factors

Depending on the above findings, if companies wanted to retain their high-value customers, promoting the confidence of customers in e-business transactions would be the most important problem to be solved, but how and where to begin? There were five major factors about trust in the questionnaire: safety and reliability of website, sustaining and stable business, practical product advertisement, strength of company, cost performance of products. The survey in high-value customers showed that almost 100% customers thought the safety and reliability of e-business website was very important, and the importance ratios of other factors from high to low were as follow: 74% (cost performance of products), 65% (sustaining and stable business), 51% (practical product advertisement), 43% (strength of company).

Besides, the study also revealed that, the importance ratio of product delivery and return services in all perception problems was the highest, as well as technical support and after-sale services in all customer satisfaction problems, specialization of products or services in all expectation problems, and the irreplaceability of products or services in all transfer cost problems, etc.

Finally, the relative importance weight of each factor should be recalculated and ordered combined with the practical situation of each company, and priority should be given to solving those more important problems so as to reduce its customer churn rate more effectively.

5. Conclusions

In today’s competitive market, many e-business companies are becoming to a full realization of the importance of the customer-oriented business strategy for sustaining their competitive edge and maintaining a stable profit level. However, to create and retain customers is difficult and costly in term of marketing. Therefore, churn prediction and management is becoming more and more important for companies in the competitive market to predict possible churners and take proactive actions to retain valuable customers and profit. And because of this, to build an effective customer churn model, which provides a certain level of accuracy, has become a research problem for both academics and practitioners in recent years.

Our research mainly concentrates on studying the reasons and factors of customer churn in e-business, developing more effective customer data processing and analysis methods. In this paper, a new churn analysis method was presented, and a kind of churn model based on data mining techniques was also established and tested on the basis of customer survey data. We expect our research to help e-business companies through the confusion of customer churn, and provide some useful clues to researchers in this field.
6. Acknowledgment

This work was supported by the National Natural Science Foundation of China (No.70971124), Shandong Soft Science Research Program (No.2010RKGB1147), Shandong Province Higher Educational Humanities and Social Sciences Research Project (No.J10WG64) and Shandong Province Housing and Urban-Rural Construction Science Research Program (No.2011RK041).

7. References