The Analysis of Telecom Customer Churn based on Data Mining

Jing Liu

Business School, Guangdong University of Foreign Studies, Guangzhou 510006, China

Jing__liu@126.com

Abstract

Customer churn is a key issue to the survival and development of the telecom enterprises, therefore, the analysis and prediction of it is of high importance. On this need, Data Mining technology is introduced in this paper, which includes its features and flows. Then the predictive system of telecom customer churn based on Data Mining is discussed and the neural network and C5.0 Model is built. Furthermore, a warning system for telecom customer churn is also introduced. The application indicated that the prediction model works effectively and the telecom customer churn warning system is of great value in reducing customer churn rate and enhancing customer loyalty.

Keywords

Data Mining; Telecom Enterprise; Customer Churn; Prediction.

1. Introduction

With the globalization and liberalization of the global telecommunication business, the competition in telecom industry becomes increasingly fierce. Relevant statistics show that the cost for developing a new customer is four times than that of retaining a frequent customer. For every 10% drop in customer loyalty, the profits of telecom enterprises reduce by 50%. The importance of reducing customer churn can thus be concluded. The application of Data Mining technology is of great value in predicting future behavior of the enterprises, therefore, the paper discusses the possibility of applying Data Mining technology to the analysis and prediction of customer churn and guiding enterprises to make proper customer management decision. The Data Mining aims at reducing customer churn and enhancing customer loyalty, which is of great significance in the sustainable development of the telecom enterprises.

1.1 Introduction of Data mining technology

Data Mining refers to the technology that extracts and mines knowledge or valuable information from mass data. The technology originated in the 1980s and plays an important role in decision-making. With the coming of big-data age, the quantities of data are mounting and the problem of "rich data with poor knowledge" arises, therefore, it is of high importance to extract knowledge and valuable information in massive amounts of data. In this context, the application of data mining technology is gradually being taken seriously.

Data mining technology has two main features: firstly, it can be applied in dealing with massive data set. Multidisciplinary algorithm and strengthening algorithm are used to process massive data, therefore, the Data Mining technology has strong data-handling capacity; secondly, the technology combines the process of "exploration" and "mining", and stresses the importance of acquiring knowledge and valuable information. Data Mining extracts valuable patterns or models by way of exploring, takes algorithm as its tool and focuses on methodology. Back in 1996, SIG put forward a standard process for data mining -- GRISP-DM [1] (cross-industry standard process for data mining). For now, GRISP-DM has a wide range of application in many fields, as is shown in Figure 1.

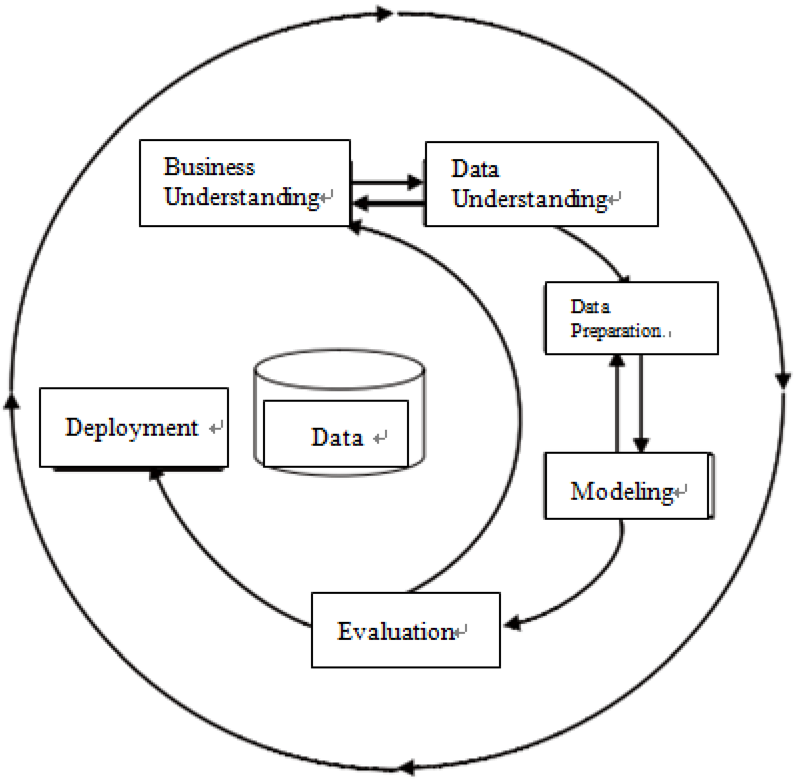

Figure 1. Standard Process of GRISP-DM

GRISP-DM mainly includes six stages: ① Business Understanding Stage: form accurate understanding of business issues through communications and change it into data mining issues; ② Data Understanding: collect, understand and filter data that is needed and conduct quality evaluation of the related data; ③ Data Preparation: the cleaning, aggregation, transformation, decomposition of data, which makes a preparation for the application of Data Mining; ④Modeling: set up models based on correlation method and solve business issues [2]; ⑤ Evaluation: to evaluate the established model; Deployment: to deploy the model into specific application. In general, data mining is a closed-loop optimization, which obtains knowledge through repeated mining and exploration.

2. The analysis of telecom customer churn based on Data Mining

The six stages of Data Mining mentioned above can also be applied in the analysis of telecom customer churn. The Business Understanding Stage corresponds to the understanding of the issue of customer churn and to convert it into Data Mining issues. The churn can be divided into passive churn (cancellation for overdue billing) and active churn (active cancellation). The prediction of telecom customer churn is essentially the mining of the relationship between correlation attributes and off-grid, which aims at converting the prediction of telecom customer churn into the issue of Data Mining; Data Understanding Stage means to collect, understand and filter the related data, and the stage corresponds to the understanding of telecom customer data. As the stages of Data Preparation, Modeling, Evaluation and Deployment are key procedures to the analysis of telecom customer churn based on Data Mining, a concrete analysis will then be conducted for the above stages.

2.1 Data Preparation

For Data Mining, valuable information is excavated from massive data, which sets a higher request for the efficiency of data storage and data analysis. The key of Data Mining [3] is to sort out the related data and select the data subset pertinently based on the domain of application. Furthermore, the accuracy of the data should be ensured in the Data Mining process. In the process of data-collecting, one or more variables may be left out and be proved to be useful in Data Mining, therefore, it is of high importance to ensure the accuracy of data prediction and selection.

Redundant or wrong data are inevitable in the mining process of massive data and can cause problems. For instance, data missing may occur, which can affect the objectivity and authenticity of the data characters and distribution. In addition, the noise data may also appear and affects the accuracy of the extract pattern. All the above shows that the mass of data brings greater difficulty to knowledge discovery, therefore, Data Preparation, which means a series process of data selection, cleaning, attribute transformation, treatment of data missing and decomposition is of high importance.

For the analysis of telecom customer churn based on Data Mining, the sample data will be split into the training set and the testing set before modeling. The former set is applied in model training while the latter is used in model testing and the split ratio of the two is 1:2. Then the split data can be mapped to the training set and the testing set respectively. The whole splitting process is achieved through Partition, as is shown in Figure 2:

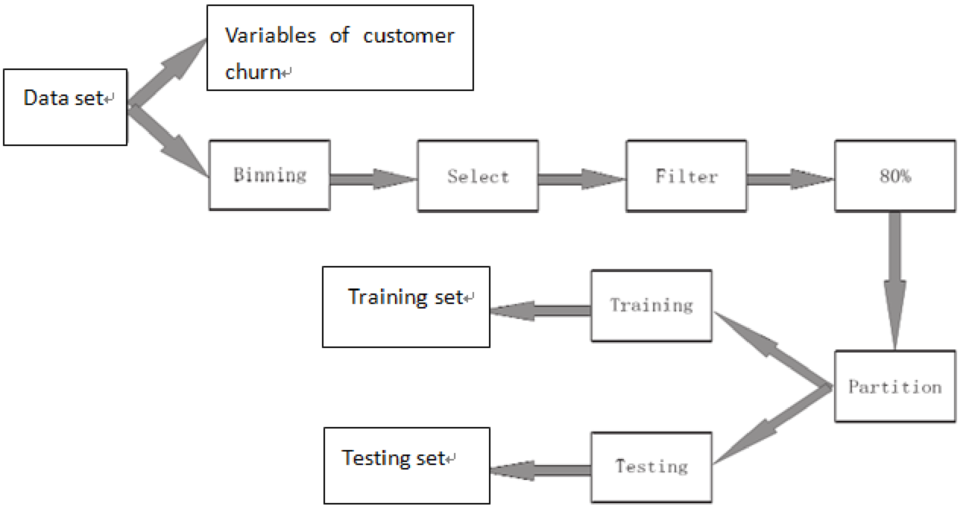

Figure 2. Data Preparation Process

2.2 Modeling

Modeling is the key link of the whole Data Mining process, which is based on data set and the attaining of goals. In the process of modeling, one or more data mining procedures, which may include neural network, Decision Tree and quantity statistics can be selected to achieve the modeling and quantification of the implied relation of the data. Constant trials and optimization are needed to obtain the best fitted model. In the process of establishing the whole model, the training set split from the Data Preparation Stage and the classification algorithm, which includes data regression, neural networks and C5.0, will be applied. It must be noted that each algorithm has its own advantages. Taking neural network and C5.0 as an example: ①neural network algorithm: strong general estimation function, with lower requirements for customers' statistics and mathematical background in the process of training and application, strong applicability [4]; ②C5.0: stable performance in cases of data missing and greater number of fields, high model training efficiency, presents modeling rules directly through Decision Tree. To achieve the optimal modeling effect, this paper establishes the mixed model of neural network and C5.0.

2.2.1 Decision Tree of Training model

The Decision Tree is an intuitive description of C5.0, which can express the model rules intuitively. Each leaf node corresponds to a specific subset of training data; each terminal node of Decision Tree corresponds to a situation in the training data set. That is, any one particular data record has only one possible prediction, and the Training Decision Tree model process is shown in Fig.3.

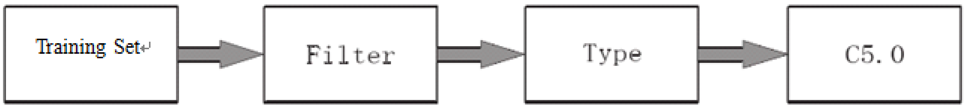

Fig.3. Decision Tree model establishing process

In the whole process of establishing the model, the main function of the Training Set node is to transfer data; the main function of the Filter node is to filter the unneeded fields; the main function of the Type node is to select the input variables and the output variables reasonably. The main function of the C5.0 node is to establish the classification model.

After the prediction model is established, in order to further enhance the predictive effect, there is a need to optimize and adjust the model based on the current predictive effect of the Test Set. Generally speaking, there are two ways to optimize and adjust the model, namely parameter adjustment method and misclassification cost method. For the Decision Tree prediction model, there are two kinds of misclassification costs, and the misclassification matrix is as shown in Table 1:

Table 1 the misclassification matrix of Decision Tree prediction model

Customer status in the sample	Predicting loss	Predicting non-loss
Actual loss	0	1
Actual non-loss	1	0

When the prediction model is accurate, the misclassification cost is 0; however, the cases should be noted when the misclassification cost is 1 and the prediction is wrong, that is, the actual loss of customers is predicted to be non-loss customers. In general, telecom customer churn accounts for a relatively small proportion, and the purpose of prediction is to know the customers that are about to be lost in advance; compared to the case that the non-loss of customers is predicted to be the loss of customers, the misclassification cost of predicting the loss of customer as the non-loss customers is higher. Based on the proportion of loss of customers and non-loss of customers in the sample data, the weighting method of loss of customer is used to gradually increase the misclassification cost of the loss of customers.

2.2.2 C5.0 and the hybrid model of neural network

Take the above trained Decision Tree model as a part of the input, train the neural network model, and a hybrid model is built. Its establishing process has a neural net node more than the Decision Tree model [5]. Based on the Decision Tree prediction model C5.0 when business is not optimal, the neural net model is trained by the neural net node. In this process, take the original data attributes and the Decision Tree model results as the input variables to improve the accuracy of hybrid model prediction.

2.3 Model evaluation

Model evaluation is also the model testing, and it is the key step of the whole Data Mining. In the above process of data preparation and model establishment, it is necessary to ensure that the established prediction model is assessable and achievable, which is to say, the accuracy of the prediction model can be evaluated, and at the same time, the prediction model should be able to be applied in the real environment and the practice of the loss analysis of telecom enterprises. Specific evaluation criteria are shown in Table 2.

Table 2 Model evaluation criteria analysis

Evaluation criteria	Explanation of valuation criteria
Model Lift value	The higher the Lift value, the better performance of the model
Model recall factor	The higher the model recall factor, the better performance of the model; however, it is contrary to the model hit rate
Model hit rate	The higher the model hit rate, the better performance of the model

3. Customer churn warning system

After establishing the telecom customer churn prediction model, combined with the practical application, the customer churn warning system is established. It should be noted that the consumption information of telecom customer is usually stored in the production database of telecom operators, and in order to avoid the impact to the production database, the consumption information cannot be directly invoked. To solve this problem, based on the ETL tool, this paper automatically invokes the daily customer data from the production database to the local database, on the basis of ensuring the stable operation of the production database to ensure the timeliness and effectiveness of the selected customer data in the customer churn warning system [6].

Based on the telecom customer churn prediction model established in the above, the report server generates the alarm report, and the user can access the information of the alarm report through the mobile terminal and so on. In addition, the communications industry's MAS server can also be used every day to push the alarm report in the form of MMS to the telecom front-line Customer Manager, in order to know the telecom front-line marketing. The telecom customer churn warning system structure is shown in Figure 4.

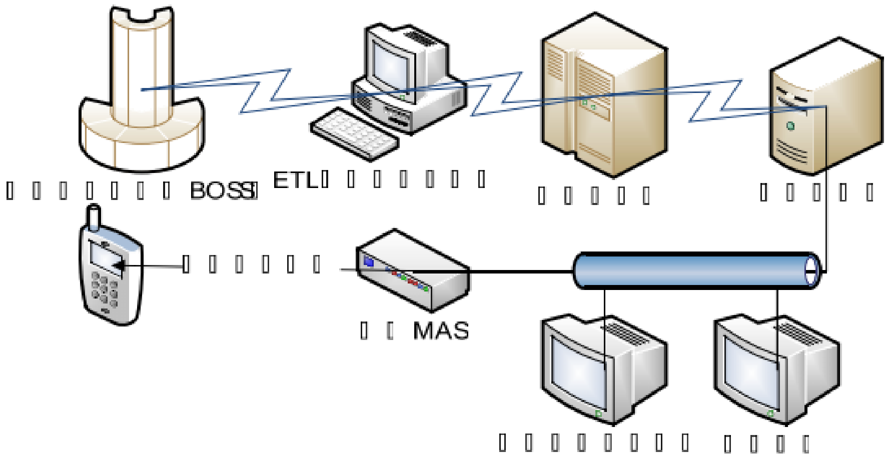

Figure 4. Telecom customers churn warning system structure

The prediction model of telecom customer churn and warning system proposed in this paper were put into use in a communication company in September 2015, and the evaluation results of C5.0 and hybrid model of neural network are shown in Table 3. After a long time data analysis, it can be seen that the predictive effect of the neural network and C5.0 hybrid model is good, and the Lift value, recall rate, and hit rate can effectively meet the requirements. Since the introduction of the customer churn warning system, the customer retention rate and adhesiveness have been gradually improved, and it has an important guidance on the business operation for the front-line manager. The front-line manager can accurately grasp the customer situation through the customer churn warning system and ensure the pertinence and reliability marketing work, while the policy-makers can promote the healthy development of market through the development of marketing policies and market environment regulation, which have important significance for the customers' retention of the entire enterprise.

Table 3 Model evaluation results

Evaluation Project	Correct number of prediction	Total number of prediction	Lift value	Recall factor	Hit rate
Hybrid model	426	483	48.53	60.15	85.70

4. Conclusion

In summary, in the era of large data, Data Mining technology is more and more widely used; for the analysis and prediction of telecom customer churn, the application of Data Mining technology is of vital importance. This paper presents a prediction model and warning system of telecom customer churn based on Data Mining technology. The actual application of a telecommunication company shows that telecom customer churn prediction model has a good predictive effect. The lift value, margin rate and hit rate can meet the demand effectively. The warning system has positive significance in the following aspects: improving customer adhesiveness, reducing customer churn rate, and guiding the work of first-line manager, etc.

References

[1] Wang Zhong, Yang Chao, Liao Zuowen. The Application of Data Mining in Telecom Customers Churn [J]. Software Guide, 2009,11:198-200.

[2] Luo Ye. The Study and Application of a Prediction Model of Telecom Customer Churn Based on Data Mining Technology [D]. Suzhou University,2008.

[3] Xu Yang. Research and Application of Loss of Customers in Telecommunication Based on Multi-Agent Data Mining Technology [D]. Central South University, 2009.

[4] Li Ai-qun, Qian Han, Wang Ruzhuan, Deng Song. Telecommunication Carriers Customer Churn Analysis Based on Distributed Hybrid Data Mining [J]. Computer Technology and Development, 2010,10:43-46.

[5] Wang Zhong.The Application of Data Mining Clementine in Telecom Customers Churn Problems [J]. Information Security and Technology, 2010,07:89-93.

[6] Liang Lu, Wang Biao, Wang Jianhui, Liu Dongning. Analysis of telecom customer churn based on fine-grained association rule mining [J].Caai Transactions on Intelligent Systems, 2015, 03: 407-413.