ATLANTIS PRESS
Advances in Intelligent Systems Research, volume 130
6th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2016)

Analysis and Application of Broadband Off-grid User Prediction Model Based on Data Mining

Juan Zhang[1, 2, a*], Xiaoyong Bian[1, 2, b]

[1]College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, China

[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, China

[a]1083181152@qq.com, [b]648183442@qq.com

Keywords: Off-grid users; Data mining; Random forest; Prediction model

Abstract. With the fast development of the communication industry in China, the users of a telecom carrier may transfer their business into another better telecom carrier. Therefore, such potential loss may make the telecom carrier more and more challenging. How to apply data mining technique in the prediction of broadband off-grid users and further the suitable decision to make is an increasingly popular problem. In this paper, a batch of consumer behavior data, i.e., call records and Internet data, are first extracted, transformed and integrated, which are utilized to generate user feature information; then a novel data mining method based on random forest is proposed to build a robust off-grid user prediction model in telecom enterprise and compared with decision tree and support vector machine. The experiments on the real user data of telecom show that the proposed model can efficiently predict most of potential off-grid users in a shorter time. At the same time, it also provides more accurate marketing strategies timely.

Introduction

After the restructuring of telecom operators in China, the competitive of the market is more intense. Incremental market potential is getting smaller and smaller, the development goals of the major operators are increasingly concentrated in the stock market [1]. In the face of huge amount of broadband data, data mining technology can still find the knowledge quickly and accurately, use existing data to predict future activities and provide valuable information for business decision-making to increase profits. R is a universal language which include statistics, forecast analysis and data visualization, can be used to analyze data with the scale of GB and TB and suit data collection, summary, transformation, exploring, modeling and visualization. Customer off-grid has been an important research topic in the world's major telecom operators. In this paper, the CRISP-DM method is used to construct the off-grid user forecast model based on Random forest, compared with the decision tree model and the support vector machine model. The model is implemented by R.

Data Preprocessing

Data Selection. The data set used for model analysis (that is target set) is extracted from the original database according to the user's needs. And this process usually involves random sampling from a large database. Although data mining is generally used for large databases, the usual data analysis requires only a few thousand or tens of thousands of records [2]. Data selection includes selection of target variables, input variables and modeling data. According to business experience and health of attributes, customer attributes, accounting attributes, traffic behavior attributes, online behavior attributes, etc., more than 100 attributes are selected finally. The number of input variables, output ones and modeling data are listed as follows.

Output variables: customer off-grid identification (off-grid customer = TRUE, non-off-line customer = FALSE).

Input Variables: important attributes that affects customer off-grid.

Modeling data: 80,000 user data in September 2015.

Data Missing. In the field of data mining research, data missing is a common problem. It will increase the complexity of the analysis and result in bias [3]. The Logit model is used to describe the distribution of missing variables.

In 2004, Barzi mentioned that in the literature, when the number of missing values is too high (for example, the missing rate is> 60%), the data loses the available value completely. This paper deleted the attributes whose missing rate> 50%. For important attributes which are low proportion missing, use mean or multiple fill method to process [4].

Relevance. Due to too many attributes in the data set, there will be a certain degree of linear correlation between these attributes. In this paper, the principal component analysis (PCA) method is used to merge some highly relevant attributes into a smaller number of attributes. The relationship between variables can be described by the correlation coefficient in PCA. The correlation of the variables shown in Fig. 1.

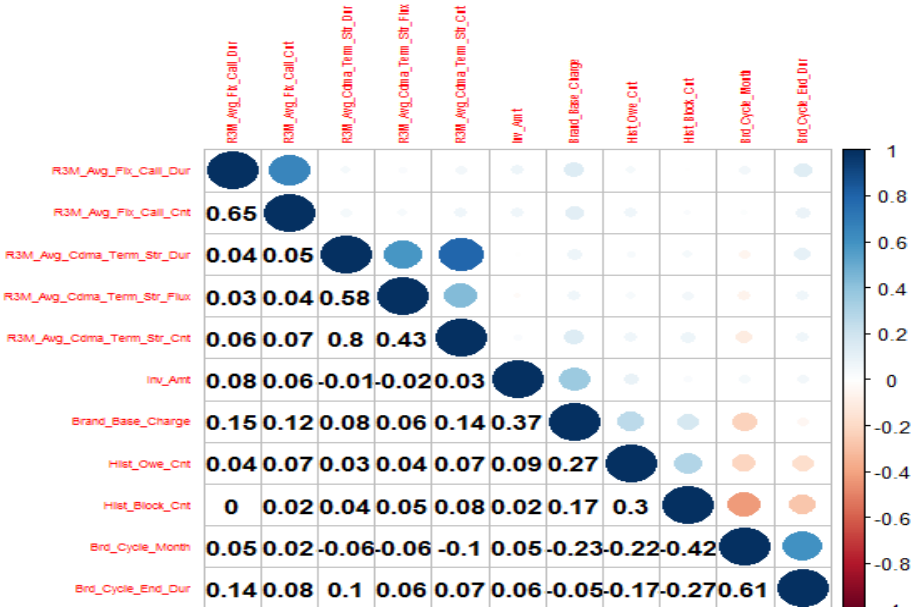

Figure 1. Correlation coefficient matrix

Introduction of Model Method

Decision Tree. Decision tree is a predictable model, which represents the mapping relations between object attributes and values [5]. It is a basic classifier for many integrated classification algorithms such as Boosting and Random Forests. The method divides the training records into groups using recursive partitioning by minimizing impurity at each step. If 100% of the observations in a node belong to a specific category of the target field, the node will be considered "pure". Decision tree construction process does not rely on domain knowledge; it uses attribute selection measures to divide tuple into different class attributes best.

Support Vector Machine. Support vector machine (SVM) has become a highly popular classification technology. This method selects a subset (that is the support vector which is very similar to the classification hyperplane) to represent the decision boundary from the training set. The standard SVM method is ultimately to solve a quadratic programming problem (QP) with constraints. By finding the hyperplane, the midpoint of the space is divided into two categories [6].

Random forest. The random forest (RF) was proposed by Leo Breiman in 2001. It uses the bootstrap resampling technique to generate a new training samples set by randomly sample K samples from the original data set N with replacement. Then the new sample set generates K number of categories to form random forest and the classification result of the new data is determined by the voting score of the classification tree [7].

The Prediction Model of Off-grid Customer Based on Random Forest

Attribute Creation. Based on the research and analysis of China telecom customer off-line early warning project, the basic attributes of consumer behavior such as number of calls, duration, number of messages, GPRS traffic, ARPU are extracted. In the meanwhile, a large number of case studies show that the off-grid probability of users is more closely related to the variation of the above attributes. Therefore, for better analysis, new attributes are created, such as ARPU' decline degree, call duration' decline degree, message count' decline degree, and GPRS decline degree based on raw attribute. The calculation formula is: (this month attributes value - last month attributes value) / last month attributes value.

Attribute Selection. In evaluating the importance of all Attributes, the random forest algorithm can avoid the multicollinearity problem that the general regression problem may face. It contains the algorithm for estimating missing values, and even if a portion of the data is lost, the random forest algorithm can still maintain a considerable degree of accuracy. At the same time, random forest algorithm is insensitive to outliers and robust in the case of interference. Mean-decrease-accuracy change the value of an attribute into a random number, the greater the number, the greater the importance of the variable. Mean-decrease-Gini calculates the effect of each attribute on the heterogeneity of observed values at each node of the classification tree through the *Gini* index. Fig. 2. shows the importance of the attributes [8].

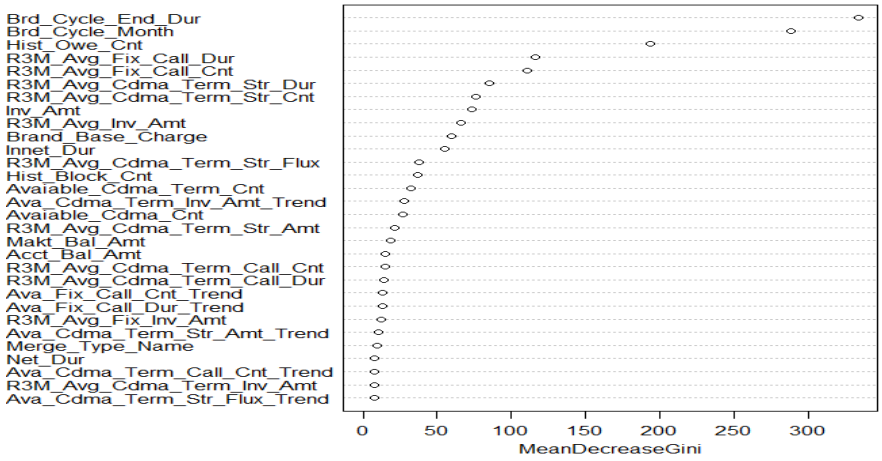

Figure 2. The importance of the original index

40 attributes were chosen as the input variables of the model, through the correlation coefficient and the importance degree.

Model Evaluation by Precision, Recall and F value. Recall rate (that is, the real rate, the formula is R= TP / (TP + FN)), precision (the formula is P= (TP + FN)) and F-score (harmonic mean number of recall rate and precision, the formula is F=2* P* R/ (P + R)) are good indices evaluating the performance of classification model [9]. Tables 1 - 3 show the results of the three evaluation indices in Random forest, SVM and random forest.

Table 1 Decision tree

Predicted value / True value	Off-grid users	Non-off-grid users	Recall	Precision	F value
Off-grid users	3285	621	98.8%	84.1%	90.9%
Non-off-grid users	41	3869			

Table 2 SVM

Predicted value / True value	Off-grid users	Non-off-grid users	Recall	Precision	F value
Off-grid users	3747	159	98.9%	95.9%	97.4%
Non-Off-grid users	40	3870			

Table 3 Random Forest

Predicted value / True value	Off-grid users	Non-off-grid users	Recall	Precison	F value
Off-grid users	3890	16	99.9%	99.5%	99.3%
Non-Off-grid users	3	3907			

Model Evaluation by ROC Curve. The true value of whether the user is off-grid is the dichotomous variable, 0 and 1, whereas the predicted value for the model is the probability score, between 0 and 1. Therefore, the area under the ROC curve (AUC) provides a way to evaluate the average performance of the model [10]. If the model is perfect, then its AUC = 1, if a model is better than the other, then its AUC is relatively large. And the more convex the curve is (0, 1), the better the fitting effect. Fig. 3-5. shows the ROC curves for the three algorithms

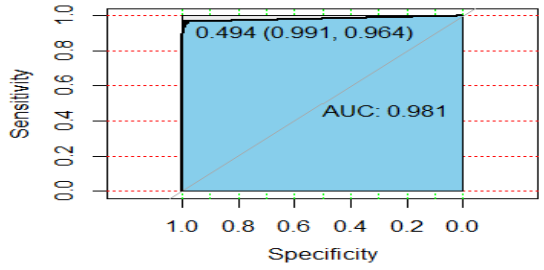

Figure 3. The ROC curve of SVM

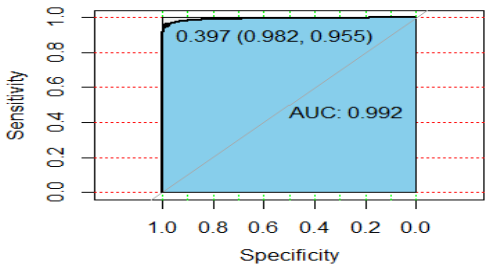

Figure 4. The ROC curve of SVM

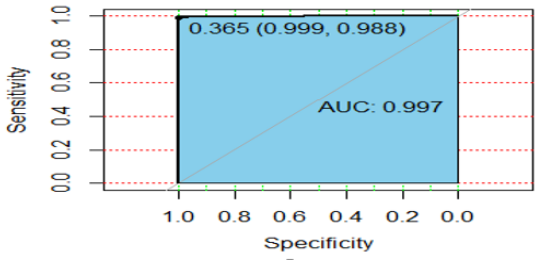

Figure 5. The ROC curve of RF

The results of the three algorithms are ranked as random forest, SVM and decision tree, therefore the proposed random forest is the best in all selected algorithms by comparing the precision, recall rate, F value and ROC curve.

Summary

This paper first makes a detailed analysis of the traditional data mining methods for researching users' behavior in telecom enterprises. Then, based on the characteristics of the data in China telecom, the extracted real data set is used to model and evaluate the potential user behavior. Further, by analyzing the influence of different attributes on experimental results, a better combination of attributes is selected. Finally, the comparison among the experimental results obtained by different models to select the optimal data model. In addition, in order to obtain more

accurate experimental results , this paper conduct some data processing such as the process of missing value and concentration value, attribute creation based on raw data and attribute selection using correlation and Boruta algorithm. Moreover, through the use of evaluation methods such as the precision, recall rate, F value and ROC curve, random forest algorithm is always better than other models. Although random forest algorithm is robust in the analysis of off-grid users, there is still space to improve the performance of the proposed RF model. Therefore, in the future work, how to balance the accuracy and forecasting time needs to be further investigated.

References

[1] P. Datta, B. Massand and D.R. Mani: Artificial Intelligence Review, Vol. 14 (2000) No.6, p.485.

[2] T.J. Gerpott, W. Rams and A. Schindler: Telecommunication Policy, Vol. 25(2001) No.4, p.249

[3] B. Huang, M.T. Kechadi and B. Buckley: Expert Systems with Applications, Vol. 39(2012) No.1, p.1414.

[4] P. Diggle and M.G. Kenward: Journal of the Royal Statistical Society, Vol. 43(1994) No.1, p.49.

[5] S.Y. Hung, D.C. Yen and H.Y. Wang: Expert systems with Applications, Vol. 31(2006) No.3, p.515.

[6] I. Robert and Kabacoff : *R in Action: Data Analysis and Graphics with R*(Manning Publications, American 2011), p.39.

[7] L.Breiman: Machine Learning, Vol. 45(2001), No.1, p.5.

[8] Y. Xie, X. Li, E.W.T. Ngai and W. Ying: Expert Systems with Applications, Vol. 7(2009) No.7, p.5445.

[9] C.F. Lin, S.D. Wang: IEEE Transactions on Neural Networks, Vol. 3(2002) No.2, p.464.

[10] J.Manyika, M.Chui, B.Brown, et al: Analytics (2011).