Study on call detail records of family members based on classification model

Xingguo WU

Department of Finance and Economics, Dongguan Polytechnic, DongGuan, 523808, China

email: wxg319@126.com

Keywords: Family Relation, calling patterns, machine learning, classifier models

Abstract. In telecommunication industry, machine learning techniques have been applied to the Call Detail Records (CDRs) for predicting customer behavior. To further investigate the information buried in huge amounts of CDRs, family relationship among mobile users can be identified, which helps the effective targeted marketing behavior, it is significantly important for increasing profitability. We use the information extracted from the CDRs analysis to identify customer calling patterns. then Customers calling patterns are given to a classification algorithm to generate a classifier model for predicting the family relation of a customer. We apply different machine learning techniques to build classifier models and compare them in terms of classification accuracy and computational performance. The reported test results demonstrate the applicability and effectiveness of the proposed approach.

Introduction

Telecommunications companies have accumulated huge amounts of call detail records, which contain much information, such as the call duration, call frequency, call time, call address and so on. How to grab useful information from a sea of data has already become a hot topic to research. For companies in the telecommunications industry, to understand their customers and their families, and to design better packages to suit their customers needs become increasingly important.

Many scholars have in-depth research on the call detail records, which aims at studying the social network analysis. The analysis leads to valuable discoveries that may have essential social and economical impact. On the one hand, the sociologists can not only distinguish terrorist groups, family relationships, friendship, etc. on basis of the analysis, but also can study the process of information flow of the relative network. Nathan Eagle from Massachusetts Institute of Technology and his research team can accurately judge whether two persons are friends or not by analyzing the calling patterns. Moreover, the accuracy rate is more than 95 %[1](Nathan Eagle, 2009). Some researchers use call detail records to trail and identifying terrorist groups [2](e.g., Nasrullah & Larsen, 2006)

On the other hand, the telecommunications workers adopt different research methods to mine information of the call records, like Wu Sen[3] (Wu Sen, 2004), constitutes the largest fully connected set with CABDP algorithms and does cluster analysis through these telephone numbers to explore the possibility of the cluster clients; Liu Xia[4] (Liu Xia, 2008) builds a collection of call records and uses the fuzzy set of semantic association algorithm to analyze the correlations of different users; Cai Xin[5] (Cai xin, 2009) employs direct data mining for family relationships through decision tree algorithms, but he is confined to mining the family numbers of the internal network. Retrieving information of call graphs (where people are the nodes and calls are the edges) obtained from the Call Detail Records (CDRs) can provide major business insights to mobile telecom operators for designing effective strategies.

The main theme is to analyze the logs that reflect social communication between different parties. Identifying social communities is an emerging research area that has already attracted the attention of many research groups. In particular, we investigate customer relationships by analyzing call detail records obtained from a telecommunication company. In this paper we concentrate more attention on the latter perspectives. Firstly, the researchers collect family information of 1843 clients and then analyze 195520 communication logs of them in two months. Secondly, they conclude specific call

modes of domestic customers. Lastly, they mine family relationship models of the clients. Through these models, the researchers can accurately judge whether two clients have family relationship between each other or not and design the family package more pertinently.

This paper is organized as follows: section 2 surveys the related work and introduces the relative concepts; section 3 describes the framework of GCRM and data preparation; section 4 presents new algorithms in GCRM and the features of calling patterns; section 5 elaborates model evaluation. The system consists of three major phases: data preprocessing, clustering, and classification.

Notation and Terminology

We assume there are a set of n customers, and that each user has contacted at least one. i represents the customer groups, while U indicates the assemblage of the customer groups：

(1) $R(j|i)$ represents the assemblage of all the customers j in contact with Customer i;

(2) N_i represents the number of Customer j in contact with Customer i in the assemblage $R(j|i)$;

$x_k(i, j|i)$ represents the attributes or features of a pair of Customer i and j. During the designated period of time, Customer j kept in touch with Customer i at least once. For different Customer i, the value of j is different, depending on the number of the customers contacting Customer i.

$V(i, j) = \{x_1(i, j|i), x_2(i, j|i), \ldots, x_m(i, j|i)\}$

V(i,j) is a characteristic vector, which represents the connection of Customer i with Customer j, (i, j, V), $\{i \in U, j \in R(j|i)\}$

In the equation, i, j and V are respectively a Customer i index in U ， a Customer j index in $R(j|i)$ and a characteristic vector of the connection between Customer i and Customer j，$S(i, j|i)$ represents the family relation fraction composed of Customer i and Customer j contacting i.

Data Preparation

In the communication industry, the detailed data can be obtained among the reliable and abundant telephone numbers. The data mainly comes from two parts: basic data and derived data. Basic data includes：calling times，called times，total times，calling time，called time, total talk time, etc. Derived data refers to the attributes and variables with the outstanding characteristics gained through data analysis or statistical methods. Sometimes derived data benefits the enhancement of the accuracy of the model. Adding fields that represent relationships considered important by experts in the field is a way of letting the mining process benefit from that expertise.

Linear scaling works well when the maximum and minimum values are known. In view of the individual difference of different families in communication behaviors, the variables should be normalized and redistributed. The ratio can be obtained from the scale of the two interconnected customers' attribute value and the maximum attribute value of all the customers contacting the target customer. As follows is the formula definition:

$$x_k^{'}(i, j|i) = \frac{x_k(i, j|i) - \min\limits_{j \in R(j|i)} (x_k(i, j|i))}{\max\limits_{j \in R(j|i)} (x_k(i, j|i)) - \min\limits_{j \in R(j|i)} (x_k(i, j|i))} \quad (1)$$

Rank:Sort the data by the first column x_k Create a new column r_k and assign it the ranked values $1, 2, 3, \cdots, n$. Given the attribute K between Customer I and Customer J, $x_k(i, j|i)$, $\{i \in U, j \in R(j|i)\}$, that is, a column with n blocks (each block includes N_i record) calculates the ranks within each block. If there are tied values, assign to each tied value the average of the ranks that would have been assigned without ties. Replace the data with a new column, where the entry $r_k(i, j|i)$ is the rank of $x_k(i, j|i)$ within block i.

Calling Patterns

The family customers are different from the nonfamily customers in call times, call duration, call time interval, call frequency and so on. These differences are beneficial for us making right choices.

Call duration of the family members mostly concentrates the scope from 25 seconds to 75 seconds, which is longer than the nonfamily members'. But the proportion of the nonfamily call duration is relatively much higher. According to the graph， it has semi-heavy tails. It was verified by [6]. Duration of the phone calls is the log-logistic distribution. The basic formulas for the log-logistic distribution and, consequently:

$$PDF(x) = \frac{\exp(z(1+\sigma)-\mu)}{(\sigma(1+e^{z}))^{2}} \tag{2}$$

$$CDF(x) = \frac{1}{1+\exp(-(\ln(x)-\mu)/\sigma)} \tag{3}$$

$$z = (\ln(x)-\mu)/\sigma \tag{4}$$

Where μ is the location parameter and σ the shape parameter. There are examples in the literature of the use of the log-logistic distribution to model the distribution of wealth, flood frequency analysis and software reliability. All of these examples present a modified version of the well known "rich gets richer" phenomenon.

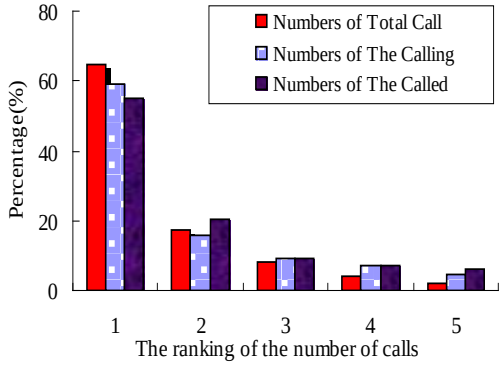

Figure 1: the ranking of the number of total call, the calling, the called of two family customers

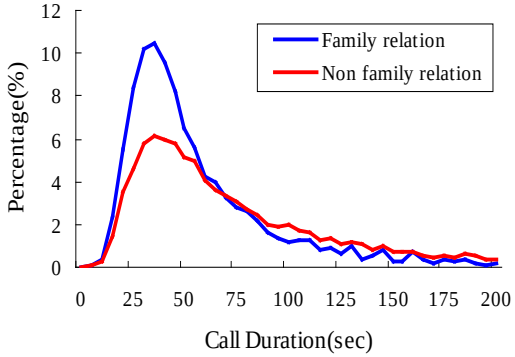

Figure 2: the average call duration distribution of the family and the nonfamily customers.

Make a statistical statement of different call time distribution. Figure 3 shows that the call rate of the family relation from 17:00 to 19:00 pm is higher than the nonfamily relation'. This result matches up the practical situation, because 18:00 pm is exactly the off-duty time. However, at about 9:00 am, the call rate of the nonfamily relation is higher than the family relation', because there is a work habit of five-to-nine in most areas of china.

a comparison is made between the family relation and the nonfamily relation about the calls on weekdays and weekends. Figure 4 clearly shows the number of the calling of the family relation is larger than the nonfamily relation's on weekdays, which also indicates the valuables on weekdays and

weekends are helpful in forecasting family relation.

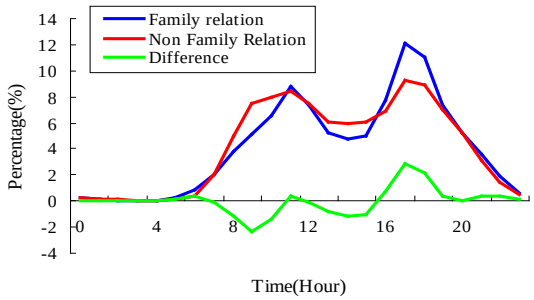

Figure 3. A simple comparison of the time distribution between the family relation and nonfamily relation.

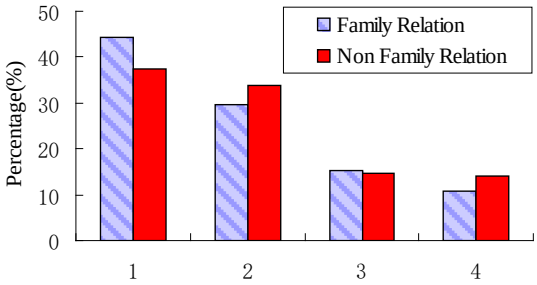

Figure 4. 1 2 3 4 respectively represent weekdays & calling, weekdays & called, weekend & calling, weekend & called

Model Evaluation

Two customers with family relations are regarded as modeling positive samples, while two customers with nonfamily relations as modeling negative samples. Through using the TOP-K idea of recommendation algorithm, the customer with the highest marks was chosen as the representative of family customers. After collecting several thousand of family data, we made a data cleansing in this situation that the sample size meets the demand, and finally obtained the relatively purer family data of 1,843 customers. In this evaluation, data set was split into 70% and 30% for training and testing on cross-validated data. We compared the C5.0 decision tree algorithm , support vector machine(SVM), Chi-squared automatic interaction detection (CHAID), Neural network algorithm (BP-Back Propagation） and classification and regression trees (C&RT). We compared our proposed optimal SVM (OSVM) method with decision-tree algorithm (DTA),naïve Bayes classifier and K-nearest neighbor algorithm(KNN)[7].

To compare the performance of the classification methods, we look at a set of standard performance measures. We use the F1 measure introduced by Ref.[3], this measure is the harmonic mean of precision and recall,which combines recall and precision in the following way:

$$Recall = \frac{\text{number of correct positive prediction}}{\text{number of positive examples}} \tag{5}$$

$$Precision = \frac{\text{number of correct positive prediction}}{\text{number of positive prediction}} \tag{6}$$

$$F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall+Precision}} \tag{7}$$

In addition, we computed and displayed a receiver-operating characteristic (ROC) curve, which represents the performance of a classifier under any linear utility function.

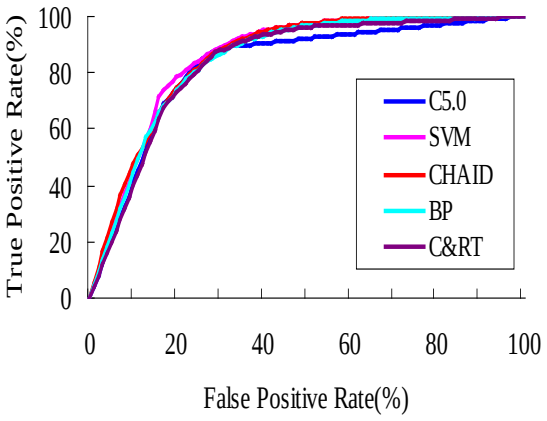

Table 1. Model Performance

Method	Recall	Precision	F1
C5.0	**84.76%**	**50.56%**	**0.63**
SVM	**86.45%**	**48.92%**	**0.62**
CHAID	**87.66%**	**46.19%**	**0.61**
BP	**86.69%**	**43.88%**	**0.58**
C&RT	**86.15%**	**46.89%**	**0.61**

Figure 5. C5.0,SVM,CHAID,BP, ROC figure of C&RT model

According to the effect, the ROC curves of these models are quite close to one another, whose F1 values have little differences. Considering model implementation and efficiency, C5.0 will be recommended to use, because the distribution of Recall and Precision respectively reach 84.76% and 50.56%.

Acknowledgments

Discussions with Docter cheng wanyou,professor Li contributed significantly to the evolution of the ideas expressed in this paper. China Telecom Corporation suggested the application of Methods to the business. China Telecom (guangdong province) Corp. Ltd. provided the Call Detail Records and feedback on classification results. Finally, This paper is supported by the National Natural Science Foundation of China under Grant No.11101081.

References

[1] Nathan Eagle1，Alex (Sandy) Pentland，David Lazer，Inferring friendship network structure using mobile phone data，PNAS，2009，Volume(106)：15274-15278；

[2] Amit A. Nanavati,Siva Gurumurthy,Gautam Das,Dipanjan Chakraborty,Anupam Joshi，On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications，In: Proceedings of the 15th ACM International Conference on Informationand Knowledge Management, 2006:435-444；

[3] Wu Sen, Cui Huan Rong, an clustering algorithm Based on paired data, Conference Proceedings of the Eighth Industrial Engineering and Enterprise Information, Telecommunications Technology (Industrial Engineering), 2004:253-257(in Chinese)

[4] Liu xia.Application of Fuzzy-association Rule to Telecom Data Mining, Jisuanji Yu Xiandaihua, 2008,154(6)：36-38(in Chinese)

[5] Cai xin.Telecom Integrated Package Marketing Base on Family Relation Identifying Model Telecommunications Science ,2009 25(9): 34-37(in Chinese)

[6] Willkomm, D., Machiraju, S., Bolot, J., Wolisz, A.: Primary users in cellular networks: A large-scale measurement study. In: New Frontiers in Dynamic Spectrum Access Networks,2008. DySPAN 2008. 3rd IEEE Symposium on. pp. 1–11 (October 2008)

[7] F.J.Provost, T.Fawcett, Robust classification for imprecise environments, Machine Learning, Vol.42, No.3,pp.203-231,March 2001.