Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

A novel and convenient variable selection method for choosing effective input variables for telecommunication customer churn prediction model

Jiayin Qi
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, P.R. China
qijiayin@139.com

Yuanquan Li
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, P.R. China
Leo8410@gmail.com

***Abstract*—Customer churn prediction model is hot research topic in recent years. Most of the researchers have paid much attention on how to construct novelist data mining algorithm for the prediction model, while less research concerns the choosing of input variables for the churn prediction model. This paper focuses on how to select effective input variables for the telecommunications customer churn model. We proposed a procedure to select the input variables step by step, and proved the effect by comparative experiment using the data from one telecom carrier.**

***Keywords*—customer churn prediction, vatiable selcetion, telecommunication industry, customer relationship management**

I. INTRODUCTION

Though customer detaining management is very important in telecommunication operators' daily work, most researches are mainly conducted on constructing customer churn prediction models, while rarely focused on selecting reasonable input variables for the models. From the 1990s to date, a large number of international journals and papers in the international conferences have been done on designing churn prediction models and algorithms by using data mining technologies to construct more effective churn prediction models (Jiayin Qi, et al., 2009; Yangming Zhang, Jiayin Qi, et al., 2007; Yingying Zhang, Jiayin Qi, et al., 2007; Lian Yan, &Richard H. Wolniewicz, 2004; Kristof Coussement, &Dirk Van den Poel, 2006; AuW, Chan CC, &Yao X, 2003; Bloemer J, Brijis T, et al. 2002; Mozer MC,Wolniewicz R et al. 2000; Ng K,& Liu H, 2001), while only a few have involved in the variable selection problem with little details(L. Yan et al., 2003).

There're two important factors which affect the predictive accuracies of churn prediction models. One is the selection of model, and the other one is the design and selection of input variables (Wei and Chiu 1999; Yan et al 2004). The more powerful the input variables are, the better the customer churn model prediction effect will be. Unfortunately, most of the previous work done for predicting customer churn has either neglected a variable selection phase, or failed to document one. So, one of the most important directions suggested by churn management is to investigate the best variables for customer churn prediction. It's assumed that most of the present research would have benefited from the inclusion of a stage that identifies the best variables for predicting customer churn (Hadden et al 2007).

Hence, it's necessary to study how to select the most effective variables for customer churn prediction model.

II. TELECOMMUNICATION CUSTOMER CHURN PREDICTION

The core of telecommunication customer churn prediction is to construct effective churn prediction models to detect customers with high churn probabilities, which is the first step of customer detainment.

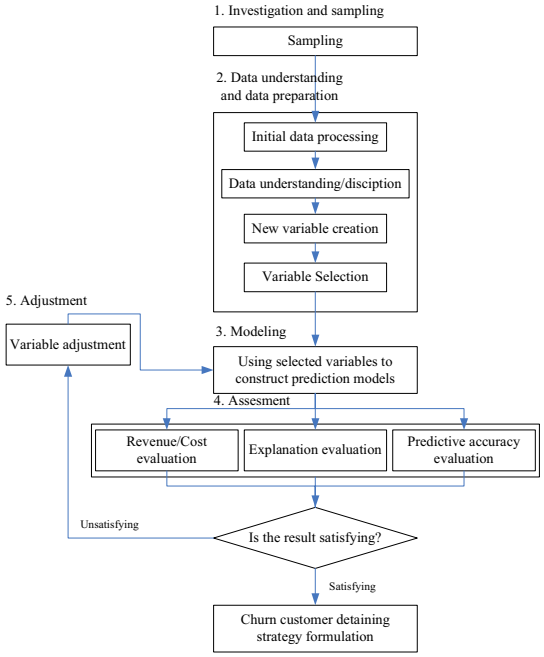

Figure 1. Telecommunications customer churn prediction process

Telecommunication customer churn prediction models apply data mining technologies to the build of mathematical models which use telecommunication customer characteristics

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

(e.g., customer basic information, bill information, call details, service logs, contractual information and credit information, etc.) to predict customer churn probabilities. Since telecommunication customer churn prediction employs data mining technologies to construct prediction models, its process can be categorized to 5 phases by referring to the data mining process, as shown in Fig.1.

A. Investigation and sampling

Different telecommunication operators may have different definitions of churn customer. Customer data storage in different operators may also vary. Hence, model builders should investigate the telecommunication operators comprehensively before modeling, in order to get clear definitions of churn customer and positive customer based on their practical work. Meanwhile, builders should understand the customer condition of available data in detail, so that a feasible and practical data sampling solution can be proposed in the next step. Usually, a data sampling solution should depict the size of the sample, the list of fields (attributes) to be extracted, the time scale of data, and the sampling proportions of actual churn customers and positive customers.

B. Data understanding and data preparation

Data understanding and data preparation mainly include four phases: initial data processing, data understanding/description, new variable creation and variable selection. The task of initial data processing is to determine and collect initial data; the task of data understanding/description is to describe the data which has been acquired, and then conduct initial data exploration and examine the quality of the data; new variable creation is to process the initial data, including selecting data for prediction model, cleaning and transforming data, as well as integrating and reducing data for modeling tools; Variable selection is to select variables which are useful for prediction as inputs of churn prediction models, from the variables generated from last step.

C. Modeling

Customer Churn Alarm Model is based on 2 basic hypotheses:

1. Different customer groups have different churn tendencies;

2. Certain customer who churned from a company will have certain unusual consuming behaviors.

Customer Churn Prediction Model is eventually fitted into a churn probability function $P(X)$, in which X stands for the eigenvector of customer (includes information stored in database which reflects the churn tendency, e.g. demographic information, account information and customer calling behavior). $P(X)$ is defined within $[0,1]$ and describes the customer churn rate if there is no detain measure for this customer in a certain coming period. Many methods are used for fitting $P(X)$. The most common methods are multiple regressions, neural network, decision tree and the combination of these 3 methods. What needs to be mentioned is there is so few churned customer for a company that the ratio of regular customers to churned customers is high skew distribution in the whole sample. This point must be taken into fully consideration when modeling and feasible measures must be used. But there are many previous researches in this field so it won't be mentioned in this paper.

D. Model Evaluation

Normally, assertion for classification of customers is not provided by prediction models. Models return a probability from 0 to 1 to say customer belonged to a certain class rather than a Yes or No conclusion to answer whether customer will churn. Model users should take all constraints (e.g. different resources of the company) into consideration, set a threshold for churn probability accordingly and finally classify customers. A high probability stands for a high possibility it belongs to a class. So there should be a probability threshold p_0, if $p \geq p_0$, this customer is identified as churner and vice versa. Churn probability threshold can be set by maximum earning principle which we will discuss in another paper.

Prediction and explanation are two main functions of churn prediction model. Issues of resource allocation, always somewhat thorny, offer still more challenges. Through prediction, the enterprises can focus on a small scale of customers (i.e. the most likely churners), and then pool resources to conduct customer retention with pertinence. Through explanation, the enterprises can find some regular patterns of customers' behavior, which will help the enterprises to better understand which factors have great impact on customers' loyalty and churn. As for prediction, response rate, captured rate, lift value, ROC value and revenue curve are the common indexes to evaluate the prediction effect of the churn model.

Response rate indicates that the percentage of churned customer in this situation:

1. Use churn prediction model to get the churn possibility of all customers;

2. Sort all customers by churn possibility;

3. Within a certain percentage of all customers (e.g. 10%), response rate stands for the percentage for which real churned customers account.

Captured response means the percentage that after sorting all customers by predicted churn possibility, how much real churned customers within a certain range of customers (e.g. 10%) account for in all real churned customers.

Lift value indicates the ratio of real churned customers within top n (n=10, 20…100) customers who has been sorted by churn possibility according to prediction model to real churned customers within top n (n=10, 20…100) customers who has been sorted by churn possibility according to subjective judgment.

That all indicators above are higher stand for a better model prediction performance.

E. Model Adjustment

If model evaluation provides a satisfied result, there is no need adjusting the model and customer retention strategy

should be the next step; if not, model should be adjusted or some other models should be selected.

III. THE NOVEL AND CONVENIENT INPUT VARIABLE SELECTION METHOD

A. *Significance of variable selection*

The principle of variable creation is to get most information out of the customer data. With this principle, the number of the variables is bound to be large. However, there could be positive variables and redundant variables for prediction. Using all variables as inputs to train churn prediction models will bring burdens to the training process of models; furthermore, some variables may have negative impact on the predictive abilities of models. Hence, variable selection is a quite important step.

B. *Principles of variable selection*

As mentioned above, variable creation means to extract potential customer churn character information as much as possible. So this information (large amount of variables) unavoidably contains lots of overload information. Input variable selection achieves both data cleaning and data reduction by selecting important features and omitting redundant, noisy, or less informative ones (L. Yan et al., 2003). So, it is necessary to extract useful and brief ones from so many variables.

- The classifying abilities of the variables should be high

Classifying ability here means the ability of a variable to classify/predict churn customers and positive customers. Many researches have paid much attention to this problem: It has been suggested by Meyer-Base and Watzel that neural networks can be used for feature selection (Meyer-Base and Watzel, 1998). Ng and Liu have performed feature selection by running an induction algorithm on the dataset.(Ng K, Liu H, 2001) A method suggested by Datta et al. involves initially finding a subset of features from the data warehouse by manually selecting those that appear most suitable for the task.(Datta et al., 2001)

We use AUC to measure the predictive ability in our study which was proposed by Yan et al. (Yan L, Wolniewicz R, 2004).

- AUC（Area Under ROC Curve）Method

AUC is the area between an ROC (Receiver Operating Characteristic) curve and the X axis. $x(p_0)$ is the abscissa, while $y(p_0)$ is the Y-axis. According to the definition of ROC curve, $y(p_0)$ is the sensitivity of the model for a given probability cutoff point p_0. The sensitivity is a measure of accuracy for predicting target A that is equal to the number of correctly predicted target as A divided by the total number is actual target A under cutoff point p_0.

p_0 is the number of incorrectly predicted target A for a given probability cutoff point divided by the number of non-A. An example of ROC curve is shown in Fig.2.

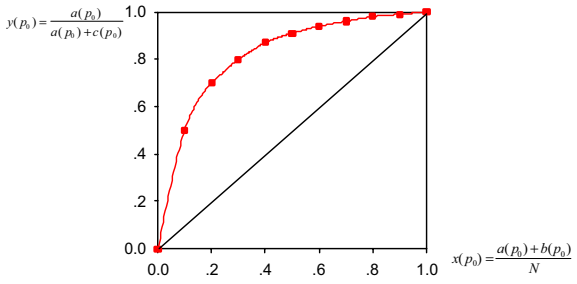

Figure 2. An example of ROC curve

ROC curve is a graphical display that gives the initiative measure of classifying accuracy. If the ROC curve in the lower left corner has a steep upward trend, it means the prediction model has a high sensitivity even with strict selection criteria. Thus the model is proved to have high accuracy. The closer the ROC curve is to the upper-left, the higher classifying accuracy the model is. The area under ROC curve-AUC is a frequently used index to evaluate the classifying accuracies of models.

AUC can be calculated in the form

$$U = \frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} I(x_i, y_j)}{mn} \tag{1}$$

Where,

$$I(x_i, y_j) = \begin{cases} 1 & if(x_i > y_j) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$\{x_0, x_1, ..., x_{m-1}\}$ is the set of different predicted values for churn customers, while $\{y_0, y_1, ..., y_{n-1}\}$ is the one of different predicted values for positive customers.

In practical applications, a variable is considered to be useful for churn prediction only when its AUC is larger than 0.5. In this way, the number of the variables can be determined. Two variables are kept only when the mutual information between the two variables is smaller than 0.5 and then the number of the variables used for building up prediction models can be determined. The variables selected can be used as inputs for initial churn prediction models. If the performances of the models are satisfying, the variables could be the final variables for prediction, otherwise future adjustment of the cutoff values of AUC and mutual information should be made to get better performances.

- Mutual information between variables should be relatively low

Mutual information is a concept to measure how much one variable tells about another one. A large mutual information between two variables means the two variables share similar information. In this case, we can deselect some redundant

variables, to make sure the mutual information between variables is relatively low.

Mutual information is computed by:

$$I(T,a) = H(T) - H(T/a) \tag{3}$$

Where:

$$H(T) = -\sum_{i=1}^{n} p(a_i)\log p(a_i) \tag{4}$$

$H(T)$ is the entropy of T

$$H(T/a) = -\sum_{j=1}^{m}\sum_{i=1}^{n} p(a_i b_j)\log p(a_i/b_j) \tag{5}$$

$H(T/a)$ is the conditional entropy of T, given the value of a;

The former one reflects the amount of "disorder" of target variable T, while the later one reflects the amount of "disorder" of target variable T after knowing a. Hence, mutual information is the amount of information that a provides to classify T.

According to mutual information, the variables can be deleted by the following principles.

Firstly, to standardize the mutual information using the equation (6).

$$StandardMI_{ij} = \frac{MI_{ij} - MinMI}{MaxMI - MinMI} \tag{6}$$

Where, $StandardMI_{ij}$ is the standardized mutual information between the ith variable and the jth variable; MI_{ij} is the mutual information between the ith variable and the jth variable; $MinMI$ is the minimum mutual information among all the mutual information between any two variables.

$MaxMI$ is the maximal mutual information among all the mutual information between any two variables.

Secondly, to judge the correlation level between any two variables. Using table I, the correlation level between any two variables can be judged by their mutual information.

TABLE I. REFERENCE VALUE FOR JUDGING CORRELATION

Interval Value for Mutual Information	Correlation Level
(0, 0.2]	Extremely Weak Correlation
(0.2, 0.4]	Weak Correlation
(0.4, 0.6]	Moderate Correlation
(0.6, 0.8]	Strong Correlation
(0.8, 1.0]	Extremely Strong Correlation

Thirdly, to get the all the variables whose mutual information is bigger than 0.4. All these variables form set High_MuIn.

Fourthly, to delete the variables whose AUC is smaller than 0.5 in set High_MuIn, and to remain the variables which are in different type of customer characteristics.

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

C. *Variable Selection Process*

There are five steps to complete the variable selection process, which is shown in Fig.3.

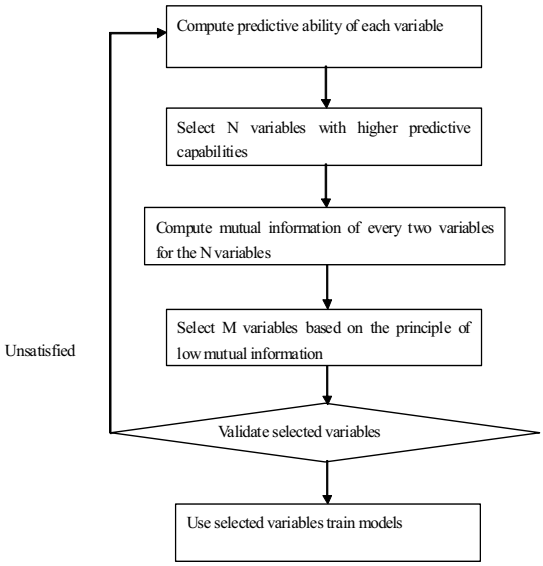

Figure 3. Variable Selection Process

Step1. To Compute the predictive ability of each variable. The predictive ability of a variable equals to its AUC value.

Step 2. To select N variables with higher predictive ability. The variables whose AUC value is bigger than 0.5 will be selected to get set High_AUC.

Step 3. To Compute the mutual information of every two variables in set High_AUC.

Step4. To Select M variables based on the principle of low mutual information. In this Step, High_MuIn is gotten from set High_AUC. At the same time, all the other variables whose mutual information is smaller than 0.4 forms set Low_MuIn from set High_AUC. Then, to remain only one variable for

each different type of customer characteristics (such as customer basic information, customer service information, customer calling information, and customer billing information etc.) in High_MUIn. This only one variable should be the one whose AUC is the highest among all the variables in this type of customer characteristics and whose mutual information with other variables are relative low. All these Variables form set R. The number of the elements in R can be expressed as .

Setp5. To use the set Low_MuIn as the input variable set, and complete customer churn prediction. If the evaluation of the prediction is good, then the set Low_MuIn can be used as the final input variables for the prediction model. And the variable selection process is over. If the evaluation of the prediction is not good, each variable combination in set R, which has $\sum_{l=1}^{r} C_r^l$ types of combination, should be added to the set Low_MuIn respectively. Let each new set Low_MuIn as the input variables, and the prediction results are tested individually. The new set corresponding to the best prediction result is the selected variables for input variables of the prediction model. If there is still no acceptable prediction results after the adding any variable combination in set R, then the prediction model should be re-built, and this is not the topic of this paper.

IV. EXPERIMENTAL RESEARCH

A. Sampling

When we started the research, we got a chance to construct customer churn prediction model for one telecommunications operator named S. We were invited to do the work for its fixed line customers in wholesale market. All the customers are business men/women. The company S tailored special service packages for this customer segment to convenient its communication in business. But, S did not see the churn rate became lower after the special service packages had been implemented. So, S wanted to explore which customers are more likely to churn. After detailed investigation for the company S, we got the analysis samples. The features of the samples are shown in tableII.

TABLE II. FEATURES FOR THE SAMPLES

Feature	description
Time period	Bill information from Jan. 2007 to Oct. 2007 Detailed sound service consumption information from Aug. 2007 to Oct. 2007 Basic customer information updated in Oct. 2007
Sampling method	All the fixed line customers in wholesale market of S. The number is about 12 thousand.
Definition for Churned customer	Those customers who discarded their fixed line service formally.
Churn rate	The churn rate is 1% in Nov. 2007
Churn month	Taking Nov. 2007 as the churn month, we use the previous months as history record to do predict

B. Initial Data Sets

The initial data sets include customer basic information, customer bill information, customer calling detailed record information and customer called detailed record information. The later two kinds of record information consists the whole detailed sound service consumption information..

C. Initial Variable Sets

Based on the initial data, we can get initial input variable sets and all the initial variables. From the customer basic information, we got two kinds of customer basic characteristic variable sets, one is customer demographic characteristic variable set, the other one is customer's contract related characteristic variable set. From the billing information, we can get the absolute characteristic variable set about customer bill, the construct characteristic variable set about customer bill and the changing characteristic variable set about customer bill. From the detailed customer sound service consumption record information, we got the absolute characteristic variable set about customer consumption behavior, the construct characteristic variable set about consumption behavior and the changing characteristic variable set about consumption behavior.

Finally, we got 3 kinds of initial input variable sets and the total number of the initial input variables is 216.

D. Variable Selection

Firstly, the AUC index is used to test each variable in the initial variable set. If the AUC value of one variable is bigger than 0.5, then the variable is useful for prediction; otherwise, it's useless for the prediction purpose and will be discarded. By using AUC index for each variable in the initial variable set, we got 155 variables for the next step.

Secondly, we measure mutual information for each two variables in the 155 variables. All the variables whose mutual information with other variables are smaller than 0.4 should be selected as the final selected input variables. Then, we get 60 variables as the final selected input variables for the churn prediction model.

E. Result

For the same churn prediction model, we used the initial 216 variables and the final selected 59 variables as the input variables respectively, and compared the computing resource consumptions and predictive effects caused by the two difference input variable sets.

As for computing resource consumption, we built the experiment environment as shown in TableIII:

TABLE III. EXPERIMENT ENVIRONMENT

Experiment Environment	
Data Set	2000 customers (246 are churned, 1754 are normal)
Customer Churn Model	Decision Tree based on Chi-Square Test

Software Platform	SAS 9.1 Enterprise Miner, Microsoft Windows XP Professional SP3
Hardware Platform	CPU: Pentium(R) Dual-Core E5300 (2.6GHz*2), Memory: 1.98GB

As the predictive result, we use ROC curve to compare the predictive ability between Tree60 and Tree216 which mean Tree model trained by 60 input variables and 216 input variables respectively. See Figure 4.

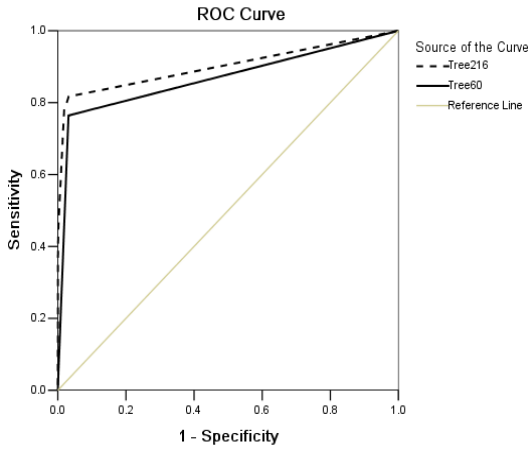

Figure 4. ROC Curve of Tree60 and Tree216

The area under ROC curve (AUC) represents prediction ability for a certain model or method. As shown in Fig. 6, AUC of Tree60 is 0.87 and AUC of Tree216 is 0.90. So, there is no obvious difference between Tree60 and Tree216 in prediction ability.

However, large differences exist in resource consumption between these two models.

TABLE IV. RESOURCE CONSUMPTION

Resource Consumption		
	Tree60	Tree216
CPU Time	0.37s	0.45s
Memory	1331KB	4653KB

As shown in Table 4, Tree216 costs 50% more CPU time and 250% more Memory than Tree60. This trend will be magnified when the input data set is much larger than our experiment data set (The real data set commonly includes hundreds millions of customers).

V. CONCLUSION

In this paper, the authors provided a simple method to select effective input variables for customer churn prediction model. The method was gotten from one of our consulting projects. It was proved to be practical. However, the method is not a systematic research for the selection of input variables for customer churn prediction model, and this will be the future research direction for the issue.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Project No.: 70701005), the Specialized Research Fund for the Doctoral Program of Higher Education of the People's Republic of China (Project No.: 20070013014), and the Open Research Fund between Beijing University of Posts and Telecommunications and IBM.

REFERENCES

[1] Au W, Chan CC, & Yao X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction, IEEE Transactions on Evolutionary Computation, 7, 532–545.

[2] Bloemer, J., Brijis, T., Vanhoof, K., & Swinnen, G. (2002). Comparing complete and partial classification for identifying customers at risk. International Journal of Research in Marketing, 20, 117–131.

[3] Datta P, Masand B, Mani DR, & Li B. (2001). Automated cellular modeling and prediction on a large scale, Issues on the Application of Data Mining, 485–502.

[4] Huayin Shu, & Jiayin Qi. (2004). Telecommunication Customer Life Cycle Management, Beijing University of Posts and Telecommunications Press.

[5] Jiayin Qi, Li Zhang, et al. ADTreesLogit model for customer churn prediction. Annual of operation research, 2009, 168(1): 247-265

[6] Kristof Coussement, & Dirk Van den Poel. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert Systems with Applications, 34(1), 313-327

[7] L. Yan. (2003). Optimizing Classifier Performance via an Approximation to the Wilcoxon-Mann-Whitney Statistic, Proc. 20th Int'l Conf. Machine Learning, AAAI Press, 848–855.

[8] Lian Yan, Richard H. Wolniewicz, & Robert Dodier. (2004). Predicting Customer Behavior in Telecommunications, Intelligent Systems, 19(2), 50-58,

[9] Meyer-Base A, & Watzel R. (1998). Transformation radial basis neural network for relevant feature selection, Pattern Recognition Letters, 19, 1301–1306.

[10] Mozer MC, Wolniewicz R, & Grimes DB. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Transactions on Neural Networks 11(3), 690–696.

[11] Ng K, & Liu H. (2001). Customer retention via data mining, Issues on the Application of Data Mining, Artificial Intelligence Review, 14(6) 569–590.

[12] Wei C, & Chiu I. (2002). Turning telecommunications call details to churn prediction: a data mining approach, Expert Systems with Applications, 23, 103–112.

[13] Yan L, Wolniewicz R, & Dodier R. (2004). Predicting customer behaviour in telecommunications, IEEE Intelligent Systems, 19, 50–58.

[14] Yangming Zhang, Jiayin Qi, Huaying Shu, Jiantong Cao. A hybrid KNN-LR classifier and its application in customer churn prediction. 2007 IEEE International Conference on Systems, Man and Cybernetics , Oct. 7-10, 2007,3265-3269.

[15] Yingying Zhang, Jiayin Qi, Huaying Shu. Case study on CRM: detecting likely churners with limited information of fixed-line subscriber. Proceedings of IEEE International Conference on Service Systems and Service Management (IEEE SSSM'06), 2: 1495-1500.