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

An Experimental Study on Four Models of Customer Churn Prediction

Chao Zhu, Jiayin Qi
Economics and Management School
Beijing University of Posts and Telecommunications
Beijing, China
zhuchao840@sina.com; ssfqijiayin@gmail.com

Chen Wang
IBM China Research Laboratory
Beijing, China
wangcwc@cn.ibm.com

***Abstract*—Decision tree, neural network and logistic regression were applied frequently as models of customer churn prediction, but the application of them has been mature and they are difficult to be improved. In this paper, Bayesian Networks, Support Vector Machines, Rough Sets and Survival Analysis were selected for experimental comparison study. An integrated contrast among the four models from the applicability of model in theory and experimental comparison has been processed. Overall, of the four models the Bayesian network model performed best while the Survival analysis did worst.**

***Keywords*—customer churn, Bayesian networks, SVM, Rough sets, Survival analysis**

I. INTRODUCTION

How to predict and prevent customer churn has become a focus that many companies and scholars are concerning. As a result of the automation of operation flow, the enterprises have accumulated plenty of business data during the daily operating activities, which gives the data mining technology a good basis to work at. In the past decades, lots of algorithms and models have been used in this field and some scholars have worked on the comparison among different methods [1]. Actually, there is no method can be better than others in all indicators, because accuracy and concision can't appear in one method simultaneously.

Until today, many algorithms and models have been used in predicting customer churn. Some models such as Decision tree [2], Artificial Neural Network [3] and Logistic regression [4] have been used frequently and some other models such as Bayesian Network [5], Support Vector Machine [6], Rough Set [7] and Survival Analysis [8][9] less more.

By summarizing the related literatures, it seems that the first three models have been studied and applied maturely. The algorithms of every one of them have been improved for many times and are difficult to become better. As the operation of enterprises gets more and more complicated, the customer churn problems are more and more difficult to be solved, which request the appearance of some new models. This study tried to choose some novel models which have a big space to be improved to solve the customer churn problems and explore some new ideas foe the studies in this field. According to such a principle, we chose Bayesian Networks, Support Vector Machines, Rough Sets and Survival Analysis for our experimental models comparison study. All the four models are relatively new in the application of customer churn and can be improved greatly. So it's meaningful for exploring new ideas and building more efficient prediction models to process an experimental comparison study among them with the data of some operator in telecom industry.

II. REVIEW OF THE FOUR MODELS

In this part, a comparative review of the four models' developments and theory backgrounds were demonstrated, the advantages and disadvantages of the four models were also summarized.

A. Bayesian networks

A Bayesian network is a kind of graphics mode used in showing the joint probability among different variable. This model provides a natural way to describe the causality information which can be used in discovering the potential relations in data. The conception of Bayesian networks was first proposed by Judea Pearl (1986), as in [10], which systematically elaborated the related concepts and principles. As the development of artificial intelligence, Bayesian networks have been successively used in knowledge representation of expert system, data mining and machine learning. In recent years, the studies and application on Bayesian networks begin to cover most fields of artificial intelligence, including causal reasoning, uncertain knowledge representation, pattern recognition cluster analysis and etc.

A Bayesian network consists of many nodes representing attributes connected by some lines, so the problems are concerned that more than one attribute determine another one which involving the theory of multiple probability distribution. Besides, since different Bayesian networks have different structures and some conceptions in graph theory such as tree, graph and directed acyclic graph can describe these structures clearly, graph theory is an important theoretical foundation of Bayesian networks as well as the probability theory.

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

B. Support Vector Machines

Support Vector Machines are developed on the basis of statistical learning theory which is regarded as the best theory for the small sample estimation and predictive learning. The studies on the machine learning of finite sample were started by Vapnik in sixties of last century and a relatively complete theoretical system called statistical learning theory was set up in nineties. After that, Support Vector Machines, a new learning machine was proposed. SVM is built on the structural risk minimization principle that is to minimize the real error probability and is mainly used to solve the pattern recognition problems. Because of SVM's complete theoretical framework and the good effects in practical application, it has been widely valued in machine learning field.

C. Rough set

Rough set is a data analysis theory proposed by Z. Pawlak. Its main idea is to export the decision or classification rules by knowledge reduction at the premise of keeping the classification ability unchanged. This theory has some unique views such as knowledge granularity which make Rough set theory especially suitable for data analysis.

Rough set is built on the basis of classification mechanism and the space's partition made by equivalence relation is regarded as knowledge. Generally speaking, it describes the imprecise or uncertain knowledge using the knowledge that has been proved. In this theory, knowledge is regarded as a kind of classification ability on data and the objects in the universe are usually described by decision table that is a two-dimensional table whose row represents an object and column an attribute. The attribute consists of decision attribute and condition attribute. The objects in the universe can be distributed into decision classes with different decision attributes according to the condition attributes of them. One of the core contents in the rough set theory is reduction that is a process in which some unimportant or irrelevant knowledge are deleted at the premise of keeping the classification ability unchanged. A decision table may have several reductions whose intersection was defined as the core of the decision table. The attribute of the core is important due to the effect to classification.

D. Survival analysis

Survival analysis is a kind of Statistical Analysis method to analyze and deduce the life expectancy of the creatures or products according to the data comes from surveys or experiments. It always combines the consequences of some events and the corresponding time span to analyze some problems. It was initially used in medical science to study the medicines' influence to the life expectancy of the research objects. The survival time should be acknowledged widely, that is, the duration of some condition in nature, society or technical process. In this paper, the churn of a customer is regarded as the end of the customer's survival time. In the fifties of last century, the statisticians began to study the reliability of industrial products, which advanced the development of the survival analysis in theory and application. The proportional hazard regression model is a commonly used survival analysis technique which was first proposed by Cox in 1972. The basic model without Time-dependent variables can be written as follows:

$$h(t) = \lambda_0(t)\exp\{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in}\}. \quad (1)$$

E. Advantages and disadvantages

These four models have their own applicable scopes, so their respective advantages and disadvantages should be defined before using them to solve some practical problems in order to exert their strengths and evade their weakness. Their concrete features are as follows:

TABLE I. COMPARISON OF THE FOUR MODELS

	advantages	disadvantages
Bayesian networks	1.able to deal with incomplete data sets 2.able to study causality 3.able to consider prior knowledge 4.able to effectively prevent over-fitting	the structure learning of the Bayesian networks will be too difficult if the data set is large
Support Vector Machines	1.fit for the finite samples 2.able to get the global optimization point but not the local extremum 3.the samples' dimension can't affect the algorithm complexity	1.there are some difficulties in theory 2.SVM has many types and is not easy to choose a fitting one
Rough set	1.any preparatory or additional information is unnecessary 2.easy to remove the data noises 3.having a good ability of knowledge reduction, be complementary with other models	only the data after discretization can be used
Survival analysis	having unique advantages in dealing with time-series data	the data with a big time span and good time continuity is necessary

III. ANALYSIS OF THE APPLICABILITY

From the introduction above, it is obvious that the four models are different in theory and have different output forms. However, they are all capable to predict the customer churn so they can be compared with each other. An analysis of the four models' applicability is given as follows:

A. Bayesian networks

As mentioned above, the learning of Bayesian networks has two main tasks. The first one is to get a complete directed acyclic graph whose nodes represent the condition attributes and decision attributes in the customer data. The connections among nodes represent the conditional probability among attributes. The second one is to get a conditional probability table (CPT) which describes each node's probability when its father nodes were assigned some values. After the two tasks, the structure and the parameters can be defined and at the same time the relations of different customer's attributes and how these attributes affect the customers' class (churn or not) can be clear. At last, the probability of customers' churn can be predicted after putting the conditional attributes in validation set into the obtained Bayesian networks.

B. SVM

The process of the prediction of customer churn using SVM can be described as follows: Each item of the customer data

can be regarded as a sample point in a multidimensional space, and each point contains all the conditional attributes a customer should have. So it's a direct method to distinguish the normal customers and the abnormal ones to find a hyper-plane that makes the two types of customers' sample points be separated at the two sides of it. Between the two types of points there is a pair of mutually parallel hyper-planes. One of them only goes through the normal customer points, while the other only goes through the abnormal customer points. This distance between the two planes is called the margin of classification. The hyper-plane in the right middle of the two planes is called optimal margin hyper-plane. A decision function can describe this hyper-plane and if the points in validation sets are put into the function, the return value can tell the type of each point.

C. *Rough set:*

The rough set classification divides the samples into different equivalent classes, each of which have the same conditional attributes. If the samples in some class are all abnormal customers, one rule will be generated. There will be a lot of rules by comparing the class of all equivalent classes. The rules can be used as the foundation to classify customers.

D. *Survival analysis*

The learning target of survival analysis is a hazard function which tells the customers' hazard rate at each time point. This hazard rate was determined by baseline hazard rate and covariate (or conditional attributes). The baseline hazard rate is only related to survival time and each attribute's regression coefficient can be given by learning the models. Then a complete hazard function is built, according to that each customer's hazard rate can be calculated. The customers' hazard rate is positively related to their churn probability. If an appropriate threshold value can be found, the customers with a high churn probability will be predicted.

Table Ⅱ summarizes the applicability analysis of the four models:

TABLE II. APPLICABILITY ANALYSIS OF THE FOUR MODELS

	Analysis of the applicability
Bayesian networks	able to give a customer's churn probability in a form of a graph and the main influence factors
Support Vector Machines	SVM itself is a good classifier and applicable to classify different kind of customers
Rough set	Rough set is good at knowledge reduction and can remove some attributes that weakly correlated with customers' class to generate some rules
Survival analysis	customers' churn can be regarded as the end of their survival time and Survival analysis is expert in solving this kind of problem

IV. EXPERIMENTS AND RESULTS

A. *Data Preparation and Preprocessing*

The data used in this experiment comes from a fixed network carrier's customers' information. This experiment chooses 2000 records from original data including 300 records from customers which have churned. 32 attributes which have strong classification ability and little mutual information have been selected.

The whole data has been divided into two parts: training set and validation set. The former one accounts for 60% and was used to train the models while the latter one accounts for 40% and was used to predict the abnormal customers in order to test the models' accuracy. Besides, there were some specific preprocessing for different models and that will be introduced in the specific experiments.

The main tools used in the experiments are Rosetta(A rough set toolkit for analysis of data), Matlab (including SVM toolbox and ex-link toolbox) 、 BN PowerSoft and SPSS(Statistical Product and Service Solutions).

B. *Modeling and Experimental process*

a) Bayesian networks: The training set was first discretized by the Data PreProcessor tool, and then the modeling and classification learning could be carried on using BN PowerPredictor tool. Figure 1 gives the obtained Bayesian network, and the node in a rough wire frame represents the class of the customers. After the prediction to the customers in validation set, every customer's churn probability can be returned. Define the customers with a probability more than 0.5 as the churn customer and 77.5% of all the churn customers have been predicted. The details will be introduced later.

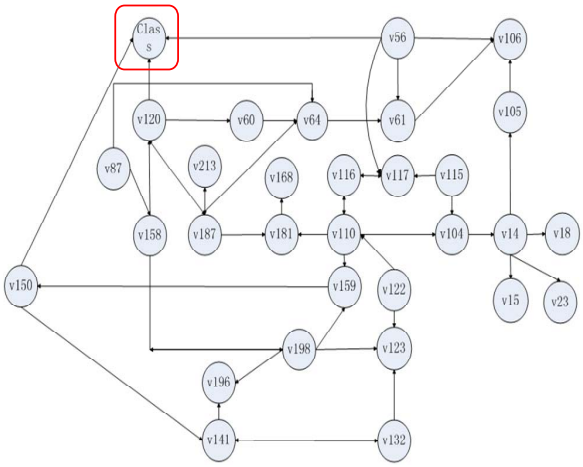

Figure 1. The bayesian networks

b) SVM: Since the SVM has a high request for the scale of data, in order to reduce the experiment time, the data set was evenly divided into four parts, three of which were used as training sets and the other one validation set. During such a process, every record in the validation has three prediction values; the one appears two or more times was selected. This paper used the SVM toolbox in the Matlab to build models; the main functions used in modeling are as follows:

$$[nsv, alpha, bias] = svc(A, A_1, \ker, E). \quad (2)$$

$$predictY = svcoutput(A, A_1, D, \ker, alpha, bias). \quad (3)$$

The first function gave the classifier and the second one returned each customer's predicted classes (predictY).

c) Rough set: The Rosetta is selected as the software in rough set experiment. There are three steps to finish the modeling process. The first step is discretization, the second one is to reduce the attributes, and at last generate rules. After the three steps, the prediction can be carried on using the rules. Before that, the data in the validation set should be discretized in the same standard as the training set. Only the rules for the churn customers were selected, so the unselected rules were regarded as ones for normal customers.

d) Survival analysis: This paper uses Cox Regression process in SPSS to carry on the survival analysis experiment. Before that, the survival time should be defined first. The time span from a customer entering the net to his stopping consuming is regarded as his or her survival time. Table Ⅲ shows the variables in the hazard function, and Figure 2 shows the curve of the survival function. It's obvious that as the time goes by, the cumulative survival rate declines. With the value of proportional hazard regression model, each customer's churn hazard rate can be obtained. Take the customer with ID6135780 for example, his survival time is 10 months and the corresponding baseline cum hazard is 0.716, so his hazard rate is as follows:

$$h_i(t) = 0.716\exp(-0.004 * v106 - 0.049 * v108 + 0.007 * v23 + 0.003 * v15 + 0.018 * v14 - 0.004 * v213 + 0.023 * v120 - 1.31 * v196 + 0.008 * v187 + 0.95 * v117 - 0.708 * v56) = 0.9724$$

This hazard rate is not a probability, so it needn't to be less than 1.

TABLE III. .VARIABLES IN THE HAZARD FUNCTION

Variables in the Equation							
		B	*SE*	*Wald*	*df*	*Sig.*	*Exp(B)*
Step 3	v106	-0.004	0.002	2.956	1	0.086	0.996
	v108	-0.049	0.007	51.098	1	0.000	0.952
	v23	0.007	0.002	9.375	1	0.002	1.007
	v15	0.003	0.001	10.396	1	0.001	1.003
	v14	0.018	0.003	31.882	1	0.000	1.018
	v213	-0.004	0.001	40.168	1	0.000	0.996
	v120	0.023	0.005	19.638	1	0.000	1.024
	v196	-1.310	0.311	17.735	1	0.000	0.270
	v187	0.008	0.002	24.665	1	0.000	1.008
	v117	0.950	0.517	3.373	1	0.066	2.585
	v56	-0.708	0.177	16.099	1	0.000	0.493

C. Evaluation and Comparison of models

The ROC and three indicators reflecting the preciseness of models are selected to evaluate the prediction effects of the models. ROC (receive operating characteristic curve) is a comprehensive indicator to reflect the sensitivity and specificity of continuous variables. The accuracy, captured response and lift value are the main indicators to evaluate the preciseness of the models. The detailed evaluation is as follows:

a) ROC-evaluation: the four models' ROC are drawn in Figure 3, Table Ⅴ shows the values of areas below curves. The values of areas are positively related to the prediction effects of the models.

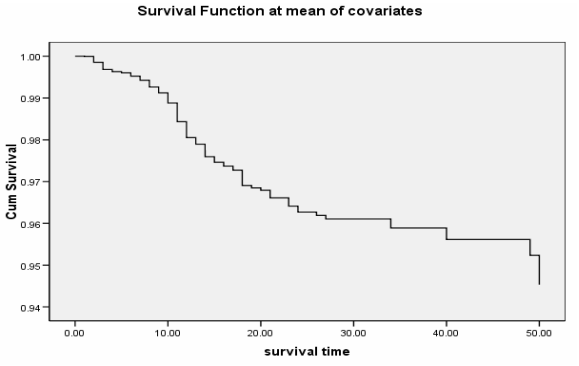

Figure 2. The curve of the survival function

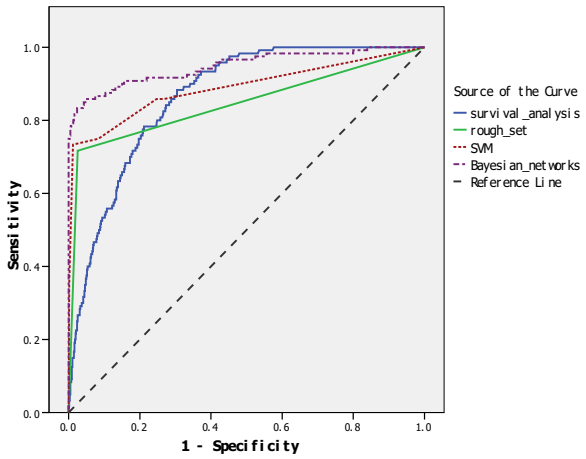

Figure 3. The ROC curve

TABLE IV. AREA UNDER THE CURVE S

Area Under the Curve					
				Asymptotic 95% Confidence Interval	
Test Result Variable(s)	*Area*	*Std. Error(a)*	*Asymptotic Sig.(b)*	*Upper Bound*	*Lower Bound*
Survival analysis	0.865	0.015	0.000	0.837	0.894
Rough set	0.846	0.025	0.000	0.797	0.895
SVM	0.889	0.022	0.000	0.847	0.931
Bayesian networks	0.948	0.014	0.000	0.921	0.975

It is obvious from Figure 3 and Table IV that Bayesian network has a best prediction effects (its corresponding area is

0.948), the second one is SVM (0.889), and then survival analysis (0.865), and the worst one of them is rough set (0.846).

b) Accuracy evaluation: Figure 4 shows the accuracy of the four models

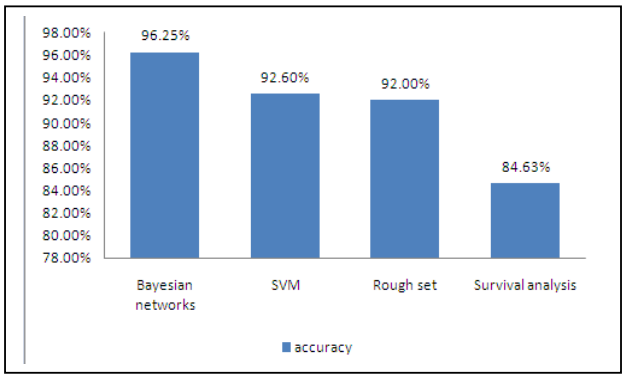

Figure 4. Accuracy of the models

From this indicator these models' prediction effects can be ranked as follows: Bayesian networks (96.25%), SVM (92.60%), rough set (92%), survival analysis (84.63%).

c) Captured response evaluation: Figure 5 intuitively shows each model's captured response. The figures on the abscissa are recorded as n, and ordinate m, then each point on the curve means that after the customers are sorted descending by the predicted churn probability, the real churn customers in the former *n%* of all the customers account for *m%* of all the real churn ones.

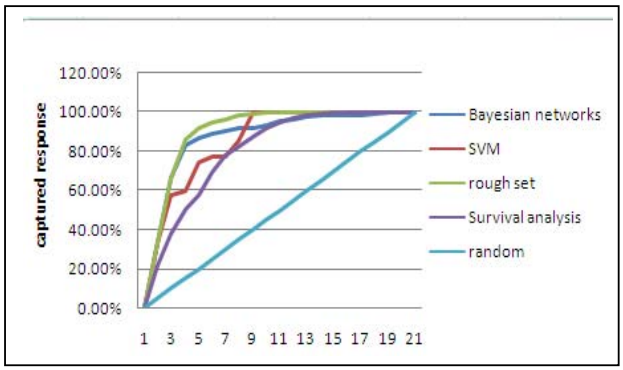

Figure 5. Captured response of the models

To observe the captured respond, Bayesian network and rough set are better than the other two models. As for Bayesian network and rough set, the former 20% of the predicted churn customers can capture about 80% of the real churn customers, which is a good effect.

d) Lift-value evaluation: Figure 6 is the broken line graph of these models' lift-value. The figures on the abscissa are recorded as n, and ordinate m, then each point on the curve means that after the customers are sorted descending by the predicted churn probability, the real churn customers in the former n% of all the customers are m times as many as those in a corresponding group without using any models.

To evaluate their prediction effects by the lift-value, Bayesian network and rough set are better than the other two. The performances of each model in the experiment are shown in Table V.

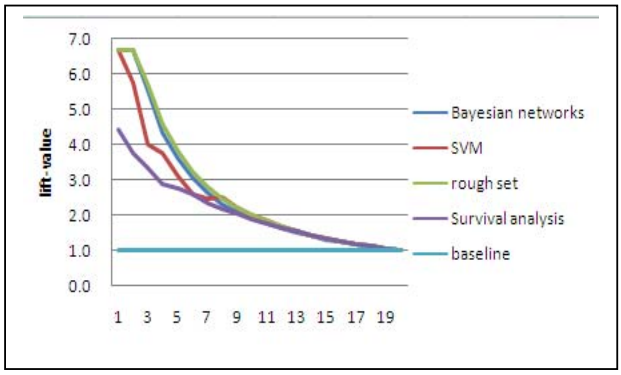

Figure 6. Lift- value of the models

TABLE V. PERFORMANCES OF EACH MODEL IN THE EXPERIMENT

Items	Bayesian networks	SVM	Rough set	Survival analysis
output	a directed acyclic graph and a CPT	a function to judge whether or not a customer will churn	the attributes after reduction and rules for classification	a proportional hazards regression model and each customer's hazard rate
complexity	complex to learn the network's structure	complex when the data scale is large	the numbers of attributes will improve its complexity	a kind of statistic method ,not very complex
the need for guidance	the prior knowledge is not necessary but can help to build model	Data-driven, don't need guidance	Data-drive and ,don't need guidance	customers' survival time are needed
easy or not to be explained	easy to be explained and have causal reasoning ability	difficult to be explained	easy to be explained and can provide if-then rules	not easy but better than SVM

The four models have different forms of outputs, and most intuitive one is Bayesian networks' graphical structure which also contains the most information. The rules of rough set can be ranked second, and the other two are relatively worse.

As for the need for guidance, only survival analysis needs the survival time as guidance. Prior knowledge can help Bayesian networks improve the accuracy but it is unnecessary for modeling. SVM and rough set are both data-driven and don't need any prior knowledge.

In this study, all the experiment models are not carried on in one platform, so it's difficult to compare their complexity. According to the time that individual model has spent, they can be ranked as follows: SVM (about 430s), Bayesian networks

(about 300s), rough set (55s) and survival analysis (less than 5s).

Interpretability of the output is an important factor. Rough set can provide if-then rules that can be easily explained. The hazard function of Survival analysis can show the importance of every attribute by the regression coefficients, but not very easily. SVM's output is too simple and difficult to explain.

V. CONCLUSIONS AND PROSPECTS

In this paper, Bayesian Networks, Support Vector Machines, Rough Sets and Survival Analysis were selected for an experimental comparison study. An integrated contrast among the four models from the applicability of model in theory and experimental comparison has been processed. According to their performances in the experiment and the evaluation of their prediction results, the Bayesian networks model has the best effect, and the Survival analysis has the worst.

Based on stage work and conclusions above, there are some prospects below in this field:

1. Since the Bayesian networks have a good prediction effect, its structure and parameter learning algorithms can be improved in order to improve its efficiency and accuracy and make it more applicable to data set with large scale and many attributes.

2. When the data set is too large, the efficiency of SVM will be relatively worse, so some work can be done to solve this problem.

3. In our experiment, rough set doesn't have a good performance, which is partly because it has been used only as a classifier. Actually, if it is used as a tool of knowledge reduction to work with other models such as decision tree, SVM, the effects may be better.

4. Survival analysis has been widely used in the researches of medical science but seldom used in customer churn prediction. It is related to the features of the data in different industries. Based on this, the next work can be the improvement of this method to make it more applicable to solve the customer churn problem in some specific industry such as telecommunications and financial industry.

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] He Dong, Guangyi Rong. Analysis and Comparison of Classification Algorithms for Data Mining. Journal of Jilin Normal University (Natural Science Edition).2008,4:107-143

[2] Bin Luo, Peiji Shao, Juan Liu. Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service. Service Systems and Service Management International Conference. June, 2007:1-5

[3] Xu E, Liangshan Shao, Xuedong Gao, Zhai Baofeng. An Algorithm for Predicting Customer Churn via BP Neural Network Based on Rough Set. Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06):47-50

[4] T. Mutanen. Customer churn analysis – a case study. Research Report.2006:1-19 A .Knott, A.Hyaes, S. A. Neslin. Next-Product-to-buy models for cross-selling applications. Journal of Interactive Marketing, 2002, 16(3):59-75.

[5] Ming Guo, Huili Zheng, Yuwei Lu. An Analysis of Customer Loss with Bayesian Networks M ethod. Journal of Nanjing University of Posts and Telecommunication. 2005,25(5):79-81

[6] Jing Zhao, Xinghua Dang. Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example. Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08:1-4.

[7] Dengf Hu. The Study on Rough Set Theory for Customers Churn. Wireless Communications, Networking and Mobile Computing. WiCOM '08: 1-4.

[8] Guozheng Zhang. Customer Segmentation Based on Survival Character. Wireless Communications, Networking and Mobile Computing. WiCom '07:3391-3396

[9] Junxiang Lu. Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS. (In)M.J.A Berry. Data Mining Techniques.

[10] Pearl, Judea. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc,1988

[11] J.B. Ferreira, M.Vellasco, M.A.Pachcco. Data mining techniques on the Evaluation of wireless churn [J]. ESANN' 2004 proceedings—European Symposium on Artificial Neural Networks Bruges (Belgium).2004,4:483-488.

[12] James J.H. Liou. A novel decision rules approach for customer relationship management of the airline market. Expert Systems with Applications 2009, 36: 4374–4381

[13] Jiayin Qi , Huaizu LI, Huaying Shu. A New Method to Determine Cumulate Customer Retention Model. Journal of Industrial Engineering Management.2004,18(4):60-63

[14] Ping Li; Jiayin Qi, Huaying Shu. The design of the detain-value model of mobile the lossing customers in communication. Journal of Beijing University of Posts Telecommunications.2005,7(3):39-43.

[15] Jiayin Qi, Yangming Zhang, Yingying Zhang, Shuang Shi. TreeLogit Model for Customer Churn Prediction. Services Computing, 2006. APSCC '06. IEEE Asia-Pacific Conference. 12,2006 :70-75.

[16] Jiayin Qi, Li Da Xu, Huaying Shu, Huaizu Li. Knowledge management in OSS - an enterprise information system for the telecommunications industry. Systems Research and Behavioral Science.2006,23(2): 177-190.

[17] Piotr Sulikowski. Mobile Operator Customer Classification in Churn Analysis.(In) SAS Global Forum 2008.

[18] Gregory F. Cooper, Edward Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning.9.1992:309-347

[19] Qing He, Zhongzhi Shi. A Novel Classification Method Based on Hypersurface. Mathematical and Computer Modelling 38. 2003. 395-407

[20] ShinYuan Hung, David C. Yen, Hsiu-Yu Wang. Applying data mining to telecom churn management [J].Expert Systems with Applications.2005, 9:1-10.

[21] J. Ross Quinlan. Learning Efficient Classification Procedures and Their Application to Chess and Games. in R. S. Michalski, J. G. Carbonell, and T. M. Mitchell(eds.). Machine learning: An Artificial Intelligence Approach. Tioga, Palo Alto, CA, 1983:463-482

[22] A .Knott, A.Hyaes, S. A. Neslin. Next-Product-to-buy models for cross-selling applications. Journal of Interactive Marketing, 2002, 16(3):59-75.