Systems Engineering — Theory & Practice
Volume 28, Issue 1, January 2008
Online English edition of the Chinese language journal

Available online at www.sciencedirect.
ScienceDirect

Cite this article as: SETP, 2008, 28(1): 71–77

RESEARCH PAPER

Model of Customer Churn Prediction on Support Vector Machine

XIA Guo-en[1,*], **JIN Wei-dong**[2]

1. Department of Business Management, Guangxi University of Finance and Economics, Nanning 530003, China
2. School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China

Abstract: To improve the prediction abilities of machine learning methods, a support vector machine (SVM) on structural risk minimization was applied to customer churn prediction. Researching customer churn prediction cases both in home and foreign carries, the method was compared with artifical neural network, decision tree, logistic regression, and naive bayesian classifier. It is found that the method enjoys the best accuracy rate, hit rate, covering rate, and lift coefficient, and therefore, provides an effective measurement for customer churn prediction.

Key Words: customer churn; support vector machine (SVM); telecommunication industry

1 Introduction

Customer churn is a concern for several industries, but it is particularly acute in the strongly competitive and now broadly liberalized mobile telecommunication industry[1]. It is estimated that the average churn rate for the mobile telecommunication is 2.2% per month[2]. Losing customers not only leads to opportunity costs because of reduced sales, but also leads to an increased need of attracting new customers[2]. The cost of acquisition of a new customer is estimated to be ranging from $ 300 to $ 600[3], and it costs roughly 5∼6 times as much to sign on a new customer as to retain an existing one[4]. To predict the latent churn customers, several researchers have mainly presented the following two methods: the first is the traditional classification methods, including decision tree[5], logistic regression[6, 7], naive bayesian classifiers[8], clustering[9]. The methods can be used to analyze qualitative data and continuous data, and interpret the prediction models. However, it cannot guarantee the precision and generalization ability of the constructed models for large scale, high dimensionality, nonlinearity, and time series data etc.; the second is the artificial intelligence method, including artificial neural network (ANN)[10], self organizing maps (SOM)[11], and evolutionary leaning (EL)[12] etc. The methods can overcome the above difficulties, and the models using the methods have nonlinear mapping ability, strong robustness, and good prediction precision[10]. However, the methods based on empirical risk minimization always lead to low generalization ability and fuzzy construction of the models[13]. Above methods have been limited in real application. Therefore, the necessity to explore new prediction approaches is strong and urgent.

To solve the above problems, the article proposes the support vector machine (SVM) with structural risk minimization and the new model evaluation standard (including hit rate, covering, and lift coefficient) for customer churn prediction, and it is based on the researches from literature[13] using SVM at the earliest stage, and from literature[14] using the evaluation standard with the whole accuracy rate and by controlling wrong judge with the weight SVM. The method supposes that if the customer data can accurately be separated by the hyperplane, which is the closest to the data vector plane, optimal separating hyperplane and classification decision function can be guaranteed. If the data vectors are linear and inseparate, a new coefficient needs to be used to control the penalty for wrong separating samples. Based on selecting the appropriate parameters and kernel functions, SVM was compared with the artificial neural network, decision tree, logistic regression, and naive bayesian classifier for customer churn prediction in home and foreign telecommunication carriers. It is found that the method has better precision.

2 Principle and arithmetic based on structural risk minimization standard

2.1 Structural risk minimization

Customer churn prediction is based on the assumption that an unknown dependence relationship exists between churn variable y and customer information variable x, that is to say, there exists an unknown joint probability distribution $P(x, y)$. To minimize the anticipant expectation risk or the actual risk

$$R(w) = \int c(y, f(x, w))\mathrm{d}P(x, y) \tag{1}$$

Customer churn prediction on machine learning aims at computing the dependence relation using a optimal function

Received date: March 17, 2006
∗ **Corresponding author:** Tel: +86-15994467010; E-mail: gandlf007711@163.com
Foundation item: Supported by the National Natural Science Foundation of China (No.6057214); and Science Foundation for Young Scholar of Guangxi (No.0832102)

$f(x, w_0)$ in a set of $\{f(x, w)\}$ when there exist l independent identical distribution samples $(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)$. Note that $\{f(x, w)\}$ is the decision function set; w is the general parameter; and $c(y, f(x, w))$ is the loss function using $f(x, w)$ to predict y. However, unknown $P(x, y)$ cannot be directly computed for minimization expectation risk by only using sample information. Therefore, the traditional methods use empirical risk

$$R_{emp}(w) = \frac{1}{l}\sum_{i=1}^{l} c(y_i, f(x, w)) \tag{2}$$

minimization to replace actual risk minimization, that is, empirical risk minimization standard.

Since the computing of the actual risk is very difficult, the researchers can only resolve the empirical risk minimization problem in several years. The relationship between empirical risk and expectation risk has been opened out because of the emergence of the statistics learning theory. If loss function $c(y, f(x, w))$ is a generic boundary nonnegative real function (using 0-1 loss function, that is to say, the expression of 0-1 loss function is $c(y, f(x, w)) = \hat{c}(y - f(x, w))$ for decision function $f(x, w)$, note that $\hat{c}(\xi) = \begin{cases} 0, & \xi = 0; \\ 1, & \xi \neq 0. \end{cases}$), and there exists leastwise $1 - \delta$ probability between empirical risk $R_{emp}(w)$ and actual risk $R(w)$ to define

$$R(w) \leq R_{emp}(w) + \sqrt{\frac{h(\ln(\frac{2l}{h}) + 1) - \ln(\frac{\delta}{4})}{l}} \tag{3}$$

Note that h is the VC dimension of function set, and l is the number of samples. Eq.(3) defines the quantitative estimation of empirical risk. The second item $\sqrt{\frac{h(\ln(\frac{2l}{h})+1)-\ln(\frac{\delta}{4})}{l}}$ of the equation's right is the VC confidence interval or confidence risk, and the summation of the right is structural risk. The equation is the upper bound of empirical risk $R(w)$. According to Eq.(3), the upper bound is the summation of empirical risk and confidence risk, where empirical risk depends on the selection of $f(x, w)$, and confidence risk is the increasing function h of VC dimension.

Further, when $\{f(x, w)\}$ is biggish, the appropriate $f(x, w)$ leads to the lesser $R_{emp}(w)$, but the biggish h of VC dimension can acquire the biggish confidence risk. Contrarily, when function set is lesser, h of VC dimension and confidence risk is lesser, but $R_{emp}(w)$ is biggish. Therefore, the conflicting tendency exists in empirical risk and confidence risk. To select the appropriate assumption $f(x, w)$, structural risk minimization is introduced.

Vapnik proposed a new strategy[13], according to which the function subset series in function set was constructed on the size of the VC dimension in all subsets; to acquire actual minimization risk, there exists the tradeoff between empirical risk and confidence risk in the subsets by searching least empirical risk (Figure 1). The idea is named as structural risk or sequence risk minimization.

2.2 Support vector machine principle and algorithm

SVM is a new machine learning method based on structural risk minimization, and the kernel contents were proposed in 1992–1995[15], which are still under constant improvement.

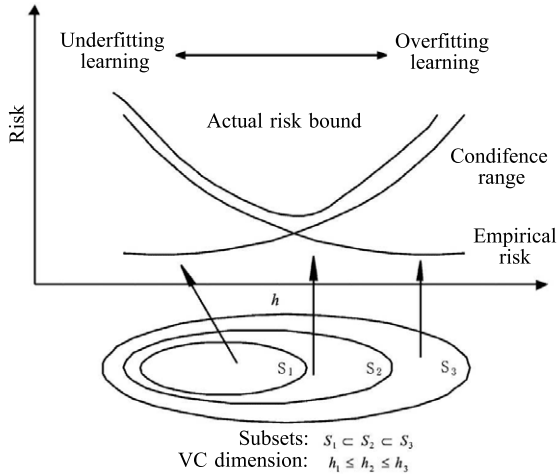

Figure 1. Structural risk minimization standard

SVM was initially proposed for the linearly separable binary problems of pattern recognition. We assume that the linearly separable sample set is $(x_i, y_i), i = 1, 2, \cdots, l, x_i \in R^n, y_i \in \{-1, 1\}$ (y_i is the ith class label).

The linear discriminant function is defined as $g(x) = w \cdot x + b$ in n dimension space, where $w \cdot x$ is the inner product of $w \in R^n$ and $x \in R^n$ vectors. The classification plane equation can be written as

$$w \cdot x + b = 0 \tag{4}$$

To let two classes of samples meet $|w \cdot x_i + b| \geq 1, i = 1, 2, \cdots, l$, the discriminant function is normalized by adjusting w and b proportionably. Then, since the classification margin is $2/||w||$, the computing of the maximum margin is converted to that of the minimum $||w||$. When the samples meet $|g(x)| = 1$ and the distance is least between the samples and classical plane, the samples are defined as support vectors and construct the optimal classification plane.

Further, the optimization classification problem turns into

$$\min \phi(w) = \frac{1}{2}||w||^2 = \frac{1}{2}(w \cdot w) \tag{5}$$

$$\text{s.t.} \quad y_i((w \cdot x_i) + b) - 1 \geq 0, \quad i = 1, 2, \cdots, l.$$

Then, the lagrange function is introduced in

$$L(w, b, \boldsymbol{a}) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} a_i(y_i((w \cdot x_i) + b) - 1), \tag{6}$$

where, $\boldsymbol{a} = (a_1, \cdots, a_l)^{\mathrm{T}} \in R_+^l$ are lagrange multipliers. Therefore, Eq.(5) can be written as

$$\min Q(a) = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{l} a_i \tag{7}$$

$$\text{s.t.} \quad a_i \geq 0, \ i = 1, 2, \cdots, l, \ \sum_{i=1}^{l} y_i a_i = 0.$$

$b^* = y_i - \sum_{i=1}^{l} y_i a_i (x_i \cdot x_j)$ is given by computing the optimal solution $a^* = (a_1^*, \cdots, a_l^*)^{\mathrm{T}}$ of Eq.(7), computing $w^* = \sum_{i=1}^{l} y_i a_i x_i$, and selecting positive a_i^* of a_i. Then,

by constructing discriminant equation $(w^* \cdot x) + b^* = 0$, we obtain decision function

$$f(x) = sgn((w^* \cdot x) + b^*) = sgn\left(\sum_{i=1}^{l} a_i^* y_i (x \cdot x_i) + b^*\right). \tag{8}$$

Vapnik introduced the kernel space theory for the linearly non-separable problem[13]. The above SVMs are designed for binary classification. The SVM for multiclass classification is by constructing and combining several binary classifiers. The former approach mainly includes three methods: One-Against-All (OAA)[16], One-Against-One (OAO)[17] and Binary-Tree-Architecture (BTA)[18].

3 Empirical research of customer churn prediction on support vector machine in telecommunication

3.1 Data processing

We collected data sets from the machine learning UCI database of University of California[19] (data set 1) and a home telecommunication carry (data set 2).

(1) Data set 1

In the data set, the definition of customer churn was a cellular phone customer who does not enjoy all services of telecommunication carry, where data window is continuously 3 months. When the fourth month customer statuses were defined as the output of model, the input of the model was obtained by converting call record, customer bill, and services etc. into the value of all attributes. Currently, the indexes of customer churn were basic character, call behavior, contact information, contract information, and product characteristic, etc. According to the Delta strategy model of customer churn management from Wei, etc.[20], influence factors, and data acquisition, we selected the indexes of call behavior and contact information (Table 1). The anomaly data were eliminated by the twice and thrice standard deviation test. By over sampling, we finally obtained 3333 training data from data set 1 (i.e., 2850 non-churn customers and 483 churn customers) and 1667 testing data at the same time of training data from data set 2 (i.e., 1443 non-churn customers and 224 churn customers). By factor analysis in SPSS11.5 for training data, the number interpreting factors were 7 and the accumulative variance was 74.08% when the eigenvalue was more than 1. Further, the indexes were defined as night charge willingness factor, day charge willingness factor, international long call charge willingness factor, evening charge willingness factor, usage time factor, customer care factor, and call frequency factor. (Table 1)

(2) Data set 2

In the data set, the definition of customer churn is that customers with personal access phone removed the phones or cancelled the phone numbers. The data were collected from July to November in 2006, where, the data window was from July to September and the delay time was October. When customer statuses in November were defined as the output of the model, the input of the model was obtained by converting the call record, customer bill, and services etc. into the value of all the attributes in the data window.

Among the customers, if the samples were interior employee, group customer, the customers whose time of usage were less than 4 months and had no consumption for the continuous 2 months in the above 4 months could be excluded. The index selection was based on data set 1.

Owing to the instability of collecting data, there existed several data with missing value. Then, the indexes with more than 30% missing values needed to be deleted; the missing items were supplemented using the equal value of the indexes; The anomaly data were eliminated by the twice and thrice standard deviation test. By over sampling, we finally obtained 1474 training data from data set 2 (i.e., 852 non-churn customers and 622 churn customers) and 966 testing data at the same time of training data from data set 2 (i.e., 534 non-churn customers and 432 churn customers). By factor analysis in SPSS11.5 for training data, the number of interpreting factors was 3 and the accumulative variance was 89.75% when the eigenvalue was more than 1. Further, the indexes were defined as voice call factor, message sending factor, and message receiving factor (Table 2).

Table 1. Factor analysis results of data set 1

Factor	Index	Variance contributing rate (%)
Night charge willingness factor	Night call charges, night call minutes	13.34
Day charge willingness factor	Day call charges, day call minutes	13.34
International long charge willingness factor	International long call charges, international long call minutes	13.34
Evening charge willingness factor	Evening call charges, evening call minutes	13.33
Usage time factor	Serves months, day call times	7.02
Customer care factor	Customer care times, message times	6.86
Call frequency factor	Evening call times, international long call times, night call times	6.85

Table 2. Factor analysis results of data set 2

Factor	Index	Variance contributing rate (%)
Voice call factor	Call minutes, call charge minutes, call charge times, call times, call charges	43.20
Message sending factor	Message sending times, message sending times from customer to user, message sending charges	29.99
Message receiving factor	Message receiving times, Message receiving times from user	16.56

Table 3. Kernel function

Name of kernel	Function expression
Radial basis kernel function	$K(x,y)=e^{-u\sum_{i=1}^{n}(x_i-y_i)^2}$
Symmetry triangle kernel function	$K(x,y)=\prod_{i=1}^{n}\max(1-u\lvert x_i-y_i\rvert,0)$
Cauchy kernel function	$K(x,y)=\prod_{i=1}^{n}\frac{1}{1+(u(x_i-y_i)^2)}$
Laplace kernel function	$K(x,y)=\prod_{i=1}^{n}e^{-u\lvert x_i-y_i\rvert}$
Square sine kernel function	$K(x,y)=\prod_{i=1}^{n}\frac{\sin^2[u(x_i-y_i)]}{u^2(x_i-y_i)^2}$
Polynomial kernel function	$K(x,y)=(\sum_{i=1}^{n}x_iy_i+1)^u$
Hyperbola secant kernel function	$K(x,y)=\prod_{i=1}^{n}\frac{2}{e^{u(x_i-y_i)}+e^{-u(x_i-y_i)}}$
Linear kernel function	$K(x,y)=\sum_{i=1}^{n}x_iy_i$

3.2 Construction of support vector machine model

According to the above analysis, the sample set (x,y) was constructed using the input data from x and the output data from y, where non-churn customers were rewritten as $y=-1$ and churn customers were rewritten as $y=1$. The kernel functions must not only meet the Mercer condition[21] but also reflect the distribution characteristic of training data in fact. Since there are no methods of selecting kernel functions for building the model, the selection of kernel functions needs to depend on transcendent information[22] and empiricism. In addition, to adjust the generalization ability of the model, the proportion between the confidence range and the empirical risk can be adjusted by parameter C. When C is small, the model is underfitting because of small empirical error, simple classification plane, and big training error; When C is big, the model is overfitting because of big empirical error, and the small weight of the classification plane normal. Therefore, it is important to select an appropriate C. Compared with the functions of Table 3[13] using MATLAB6.5 to select kernel functions and the parameters, it is found that the SVM model based on $w\cdot x+b=0$ can acquire good results using the radial basis kernel function with $u=0.12$, $C=3$ in data set 1 (Table 4); in addition, the SVM model using cauchy kernel function with $u=11$, $C=0.3$ can acquire good results in the data set 2 (Table 5).

3.3 Analysis of empirical results

By selecting the best parameters and precision of the prediction model, Tables 7 and 8 show the prediction results of the SVM model compared with the other methods in dataset 1 and dataset 2. ANN used BP arithmetic, 8 and 5 hidden layers, 0.1 error, and discriminant: $f(\sum_{j=1}^{n_H}w_{if}(\sum_{i=1}^{d}w_{ji}x_i+w_{j0})+w_0)$ where, $f(\cdot)$ was sigmoid function; w_{ji} was the weight from input cell i to hidden cell j; n_H was the number of hidden layers; d was the dimension of x; w_{j0} and w_0 were the values of inner cells. The attributes of the decision tree C4.5 were computed using the discriminant: $P(w_0|A=0)$, i.e., the information gain method, where, $P(\cdot)$ was condition probability; and w_i was the classification type. The information gain ratio $Gaintatio(X,A)=\frac{Gain(X,A)}{SplitIF(X,A)}$ was maximum when A was choice attribute (note that X was split into $S_1,S_2,\cdots,S_s$ subsets on S different values of A, where, $Gain(X,A)$ was information gain by splitting S; $SplitIF(X,A)$ was information quantity). The discriminant of logistic regression was $1/(1+\exp(1-(b_0+$

Table 4. SVM prediction results using the different kernel functions in dataset 1

Function	C	u	AR	HR	CR	HC
Radial basis kernel function	3	0.12	0.9088	0.8333	0.4018	6.2186
Symmetry triangle kernel function	3	0.12	0.9046	0.8218	0.3705	6.1328
Cauchy kernel function	3	0.12	0.9058	0.819	0.3839	6.1119
Laplace kernel function	3	0.12	0.9022	0.7699	0.3884	5.7455
Square sine kernel function	3	0.90	0.9022	0.7699	0.3884	5.7455
Polynomial kernel function	1	3	0.8716	0.7273	0.0714	5.4276
Hyperbola secant kernel function	3	1	0.8914	0.6838	0.3571	5.1029
Linear kernel function	1.5	—	0.7972	0.2808	0.3259	2.0955

Notes: linear kernel function has no parameter u; C and u are the parameters of the model; AR is the accuracy rate; HR is the hit rate; CR is the coverage rate; HC is the lift coefficient.

Table 5. SVM prediction results using the different kernel functions in dataset 2

Function	C	u	AR	HR	CR	HC
Radial basis kernel function	2	1	0.5911	0.7126	0.1435	1.5942
Symmetry triangle kernel function	11	0.025	0.5880	0.7073	0.1343	1.5823
Cauchy kernel function	11	0.3	0.5963	0.7141	0.1620	1.5975
Laplace kernel function	11	0.03	0.5828	0.6381	0.1551	1.4275
Square sine kernel function	9	2	0.5932	0.7010	0.1574	1.5682
Polynomial kernel function	20	4	0.5538	0.6000	0.0069	1.3423
Hyperbola secant kernel function	10	0.5	0.5694	0.7000	0.0648	1.5660
Linear kernel function	0.71	—	0.5538	1.0000	0.0023	2.2371

Notes: linear kernel function has no parameter u; C and u are the parameters of the model; AR is the accuracy rate; HR is the hit rate; CR is the coverage rate; HC is the lift coefficient.

Table 6. Classification matrix

Customer state	Prediction churn	Prediction non-churn
Actual churn	A	B
Actual non-churn	C	D

Table 7. Prediction results compared with all the methods in dataset 1

Model type	Accuracy rate	Hit rate	Coverage rate	Lift coefficient
SVM	0.9088	0.8333	0.4018	6.2186
ANN	0.8983	0.7538	0.3625	5.6256
Decision tree C4.5	0.8386	0.3869	0.3437	2.8876
Logistic regression	0.8716	0.6190	0.1160	4.6198
Naive bayesian classifiers	0.8782	0.7142	0.1562	5.3305

Table 8. Prediction results compared with all the methods in dataset 2

Model type	Accuracy rate	Hit rate	Coverage rate	Lift coefficient
SVM	0.5963	0.7141	0.1620	1.5975
ANN	0.5569	0.7500	0.0139	1.6779
Decision tee C4.5	0.5248	0.4657	0.4236	1.0417
Logistic regression	0.5890	0.7012	0.1412	1.5686
Naive bayesian classifiers	0.5549	0.6250	0.0116	1.3982

$\sum_{i=1}^{d} b_i x_i))) - \frac{1}{2} = 0$, where, b_0 was constant; d was dimension; and b was regression coefficient, which was the contributing quantity from x for the discriminant. The discriminant of naive bayesian classifiers was $p(x|w_i)P(w_i) - p(x|w_j)P(w_j) = 0$, where, w_i and w_j were the classification types; $p(x|w)$ was the condition probability density; and $P(w)$ was the transcendent probability. Table 6 shows the evaluation standard of the model: model accuracy rate=(A+D)/(A+B+C+D); hit rate=A/(A+C); coverage rate= A/(A+B); lift coefficient=hit rate/churn rate of the testing. In Tables 7 and 8, the accuracy rate (0.9088), hit rate (0.8333), coverage rate (0.4018), and lift coefficient (6.2186) in dataset 1 using the SVM model are superior to the other methods; the accuracy rate (0.5963), hit rate (0.7141), coverage rate (0.1620), and lift coefficient (1.5975) in dataset1 using the SVM model are superior to the other methods except that ANN in the hit rate and decision tree C4.5 in the cover-age rate are slightly superior to the proposed method, where, the main reasons are that the coverage rate (0.0139) of ANN leads to the overfitting phenomena, and there exists little influence in the rare classes for the decision tree. In addition, the reason why the prediction precision in dataset1 is superior to that in dataset 2 is that there exist less data indexes, indistinct churn trend, serious missing record, long delay time, and so on, in dataset 2. The good accuracy rate shows that the SVM model has strong integration prediction ability in the whole datasets; the good lift coefficient and coverage rate show that the model can keep more latent churn customer with less cost in the telecommunication markets with different churn rates. The reason why the good results are acquired using the SVM model, which have appropriate kernel function and parameter is that the SVM model on structural risk minimization includes empirical risk minimization and confidence minimization.

4 Conclusions

SVM is a general learning arithmetic based on the statistics learning theory, and it can solve the nonlinearity, high dimension, and local minimization problems, which are insoluble for traditional methods in the customer churn prediction of telecommunication. The article used SVM to predict customer churn in the telecommunication, and compared with BPANN, decision tree C4.5, logistic regression, and naive bayesian classifiers. It then draws the conclusion that the traits of SVM are the simple classification plane, strong generation ability, and good fitting precision etc. from the methodology. From data condition and structure, when there exist several samples (abundant support vectors), abundant attribute, big churn rate, less missing record, and nonlinearity data, SVM has good prediction precision. In addition, the article also finds according to the above analysis that the customers with big churn rate have the following traits in dataset 1: strong charge willingness, long mobile service, and considerable customer care; the customers with big churn rate have the following charactersitics in dataset 2: regular call and message usage. The results show that when telecommunication carries keep the customers with long call usage, the carries need to use favourable sales promotion work, increase the times and quality of customer care, and improve the increment service to keep the customers with several call charges. However, since the above methods are initially used to predict customer churn, there exist some problems, for example, how to select fitting kernel function and parameter; how to weigh customer samples. In addition, since the traits of customer churn data in the telecommunication are of large scale, high dimension, nonlinearity, nonnormality, time series, and rare class, SVM can also be developed in the bank industries using similar customer churn data.

References

[1] Keaveney S M. Customer switching behavior in service industries: An exploratory study. Journal of Marketing, 1995, 59(2): 71–82.

[2] Mozer M C, Wolniewicz R, Grimes D B, et al. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Transactions on Neural Networks, 2000, 11(3): 690–696.

[3] Athanassopoulos A D. Customer satisfaction cues to support market segmentation and explain switching behavior. Journal of Business Research, 2000, 47(3): 191–207.

[4] Bhattacharya C B. When customers are members: Customer retention in paid membership contexts. Journal of the Academy of Marketing Science, 1998, 26(1): 31–44.

[5] Chih P W, Chiu I T. Turning telecommunications call details to churn prediction: A data mining approach. Expert Systems with Applications, 2002, 23(2): 103–112.

[6] Kim H S, Yoon C H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy, 2004, 28(9): 751–765.

[7] Rosset S, Neumann E. Integrating customer value considerations into predictive modeling. Third IEEE International Conference on Data Mining, 2003: 1–8.

[8] Nath S V. Data warehousing and mining: Customer churn analysis in the wireless industry. A thesis submitted to the faculty of the college of business in partial fulfillment of the

requirements for the degree of master of business administration, May 2003.

[9] Yi Ming, Hui Wan, Lei Li, et.al. Multidimensional model-based clustering for user behavior mining in telecommunications industry. Proceeding of the third international conference on machine learning and cybernetics, Shanghai, 2004: 26–29.

[10] Yan L, Miller D J, Mozer M C, et al. Improving prediction of customer behavior in nonstationary environments. Proc of the International Joint Conference on Neural Net-works, 2001: 2258–2263.

[11] Ultch A. Emergent self-organizing feature maps used for prediction and prevention of churn in mobile phone markets. Jornal of Targeting, 2002, 4(10): 401–425.

[12] Au W, Chen K C C, Yao X. A novel evolutionary data mining algorithm with applications to churn prediction. Evolutionary Computation, IEEE Transactions, 2003, 7(6): 532–545.

[13] Vapnik V N. The nature of statistical learning theory. Beijing: Publishing House of Electronics Industry, 2004.

[14] Xia G E, Jin W D. Tradeoff of errors of two types in customer churn prediction. Journal of Marketing Scinece, 2006, 2(4): 1–7.

[15] Cortes C, Vapnik V. Support vector networks. Machine Learning, 1995, 20: 273–297.

[16] Hsu C W, Lin C J. A comparison of methods for multiclass support vector machines. IEEE Transactions On Neural Networks, 2002, 13(2): 415–425.

[17] Platt J C, Cristianini N, Shawe T J. Large margin DAG's for multiclass classification. Advances in Neural Information Processing Systems, 2000(12): 547–553.

[18] Heong S M, et al. Support vector machines with binary tree architecture for multiclass classification. Neural Information Processing Letters and Reviews, 2004, 2(3): 47–51.

[19] Merz C J, Murphy P M. UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/ MLRe-pository.html]. Irvine, CA: University of California,Department of Information and Computer Science.

[20] Wei Y, Jutla D N, Sivakumar S C. A churn strategy alignment model for managers in mobile telecomcommunication. Networks and Services Research Conference, Proceedings of the 3rd Annual16-18 May 2005: 48–53.

[21] Muller K R, Mika S, Rrtsch G, et al. An introduction to kernel based learning algorithms. IEEE Transactions on Neural Networks, 2001, 12(2): 181–202.

[22] Smla A J. Learining with kernels. Berlin: Fechnical University of Berlin, 1998.