# Customer Churn Prediction Using Improved One-Class Support Vector Machine

Yu Zhao, Bing Li, Xiu Li, Wenhuang Liu, and Shouju Ren

Cims Research Center, Automation Department,
Tsinghua University, Beijing 100084, China
zhaoyu01@mails.tsinghua.edu.cn

**Abstract.** Customer Churn Prediction is an increasingly pressing issue in today's ever-competitive commercial arena. Although there are several researches in churn prediction, but the accuracy rate, which is very important to business, is not high enough. Recently, Support Vector Machines (SVMs), based on statistical learning theory, are gaining applications in the areas of data mining, machine learning, computer vision and pattern recognition because of high accuracy and good generalization capability. But there has no report about using SVM to Customer Churn Prediction. According to churn data set characteristic, the number of negative examples is very small, we introduce an improved one-class SVM. And we have tested our method on the wireless industry customer churn data set. Our method has been shown to perform very well compared with other traditional methods, ANN, Decision Tree, and Naïve Bays.

## 1 Introduction

According to Don Peppers and Martha Rogers, the marketing experts, most efficiency, and the incapacity of searching a specific item, enterprises average lost 25% customers annually. However, the cost of obtaining a new customer is five times higher than maintaining an existing customer [1].

In many industry fields, churn - that can be looked as the customer's decision to end the relationship and switch to another company - has become a major concern.

The churn rate for U.S. mobile carriers is 2 % to 3 % monthly, a major expense for the companies, which spend $400 to $500 to sign a single customer who typically generates about $50 in monthly revenue. Companies are now beginning to realize just how important customer retention is. In fact, one study finds that "the top six US wireless carriers would have saved $207 million if they had retained an additional 5% of customers open to incentives but who switched plans in the past year" [2]. Over the next few years, the industry's biggest marketing challenge will be to control churn rates by identifying those customers who are most likely to leave and then taking appropriate steps to retain them. The first step therefore is predicting churn likelihood at the customer level.

The Customer Churn Prediction problem has two major characteristics:

The first is that the number of churn customers (the negative examples) is small (2% in the total examples);

X. Li, S. Wang, and Z.Y. Dong (Eds.): ADMA 2005, LNAI 3584, pp. 300–306, 2005.

The second is accuracy. Consequently, for a carrier with 1.5 million subscribers, improving the monthly prediction accuracy rate 1% would yield an increase in annual earnings of at least $54 million.

Customer Churn Prediction generally can be considered as a binary classification problem, distinguishing between normal and churn. The standard support vector machine (SVM) is a classifier that finds a maximal margin separating two classes of data. There have been a lot of successful applications about that. But the data of Customer Churn Prediction are very special: the normal dataset is much larger than the abnormal. Therefore, the standard SVM does not work well on our task. We present an improving SVM method, which is based on one-class SVM described in [3] by Bernhrd Scholkopf et al. We used dataset provided by a wireless telecom company and included more than 150 variables describing more than 100,000 customers. We have performed experiments on the improved one-class SVM with the various kernel functions, and have compared the performance of SVM and other normal methods (such as ANN, Decision Tree, and Naïve Bays).

The rest of this paper is organized as follows. A brief description of the Customer Churn Prediction and SVM model will be described in Section 2. The improved one-class SVM will be introduced in Section 3. The dataset preparation and various experiments of improved one-class SVM and their results are presented in Section 4. Some concluding remarks and future work are given in Section 5.

## 2 Customer Churn and SVM Model

Customer churn – the propensity of customers to cease doing business with a company in a given time period – has become a significant problem for many firms. These include publishing industry, investment services, insurance, electric utilities, health care providers, credit card providers, banking, Internet service providers, telephone service providers, online services, and cable services operators.

There are numerous predictive modeling techniques for predicting customer churn. These vary in terms of statistical technique (e.g., neural nets versus logistic regression), variable selection method (e.g., theory versus stepwise selection), number of variables included in the model, and time spent in total on the modeling exercise as well as how a given time budget is allocated across various tasks in the model-building process [4].

SVM algorithm developed by Vapnik [5] is based on statistical learning theory. In some classification cases, we try to find an optimal hyper-plane that separates two classes. When the two classes of points in the training set can be separated by a linear hyper-plane, it is natural to use the hyper-plane that separates the two groups of points in the training set by the largest margin. In order to find an optimal hyper-plane, we need to minimize the norm of the vector w, which defines the separating hyper-plane. This is equivalent to maximizing the margin between two classes. [6]

Customer Churn is a problem of classification between “churn” and “no churn”. But when the number of the negative examples is too small, the generalization performance of SVM classifier must be weak, and the error rates is proved unsatisfactory.

## 3 Improved One-Class SVM

One-class SVM: Bernhard Scholkopf [7] et al. suggested a method of adapting the SVM methodology to the one-class classification problem. Essentially, after transforming the feature via a kernel, they treated the origin as the only member of the second class. By introducing "relaxation parameters", they separate the image of the one class from the origin.

Li present a One-Class SVM for anomaly detection [8]. The basic idea is to work first in the feature space, and assume that not only is the origin in the second class, but also that all data points "close enough" to the origin are to be considered as outliers or anomaly data points. If the input data match the selected samples, then they are regarded as anomaly data, i.e., that belongs to the anomaly class.

Here we introduce an improved One-Class SVM to predict customer churn:

Suppose we are given the training data:

$\{(x_1, y_1),(x_1, y_1),\cdots(x_1, y_1)\}$, where $x \in R^N$, $y \in \{-1,+1\}$, and $R^N$ is the feature space. This leads to the following quadratic programming problem:

$$\begin{aligned} &\min(R^2 + \sum_{y_i=1} C_+\xi_i + \sum_{y_i=-1} C_-\xi_i) \\ &s.t. \quad y_i(\left\|\Phi(x_i) - \alpha\right\|^2 - R^2) \le \xi_i \\ &\xi_i \ge 0, 1 \le i \le l \end{aligned} \tag{1}$$

Where $\xi_i$ are slack variables that are penalized in the objective function. The goal of introducing the slack variables is to allow some error during the training, where $\alpha$ and R are the center and radius of the hyper-sphere respectively, and $C_+ = \frac{l_-}{l_+ + l_-}C$, $C_- = \frac{l_+}{l_+ + l_-}C$ are penalty parameter. $l_+$ is the number of the positive examples. $l_-$ is the number of the negative examples.

The corresponding dual is:

$$\begin{aligned} &\min(\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(\Phi(x_i)\cdot\Phi(x_j)) - \sum_{i=1}^{l}\alpha_i y_i(\Phi(x_i)\cdot\Phi(x_j))) \\ &s.t. \ \sum_{i=1}^{l}\alpha_i y_i = 1, \\ &\quad 0 \le \alpha_i \le C_+, \quad y_i = 1, \\ &\quad 0 \le \alpha_i \le C_-, \quad y_i = -1, \end{aligned} \tag{2}$$

We can use some appropriate kernel $K(x_i, x_j)$ representing the inner product $\Phi(x_i)\cdot\Phi(x_j)$. The choice of the kernel functions depends on the experience and experiment.

For any input $x$, first we calculate the distance between the data point and the center of the hyper-sphere, if the following condition is true,

$$\left\|\Phi(x)-x\right\| \le R \tag{3}$$

The data point $x$ belongs to the hyper-sphere and regard it belongs to +1 class, otherwise it belongs to -1 class.

$$R^2 = 1-\frac{2}{n}\sum_{k,i} a_i y_i K(x_k, x_i) + \sum_{i,j} a_i a_j y_i y_j K(x_i, x_j) \tag{4}$$

Where $x_k$ are the bounded vectors and $n$ is the number of the bounded vectors.

The decision function can be written as:

$$\begin{aligned} f(x) &= \operatorname{sgn}(\sum_{i=1}^{l} \alpha_i y_i K(x, x_i) + b) \\ where\ \ b &= -\frac{1}{n}\sum_{k,i} \alpha_i y_i K(x_k, x_i) \end{aligned} \tag{5}$$

## 4 Data and Experiment Results

### 4.1 Characteristics of Input Data

Ultimately, churn occurs because subscribers are dissatisfied with the price or quality of service, usually as compared to the offerings of competing carriers. The main reasons for subscriber dissatisfaction vary by region and over time. We categorize our input variables as follows [9].

- Demographics: Geographic and population data of a given region.
- Usage level. Call detail records (date, time, duration, and location of all calls), peak / off-peak minutes used, additional minutes beyond monthly prepaid limit etc.
- Quality of Service (QOS): Dropped calls (calls lost due to lack of coverage or available bandwidth), and quality of service data (interference, poor coverage).
- Features / Marketing: Details of service bundle such as email, instant messaging, paging, rate plans offered by carrier and its competitors, recent entry of competitors into market, advertising campaigns, etc.

### 4.2 Data

The subscriber database provided by the carrier is stored in an Oracle database. It contains three relations which are listed in Table 1.

**Table 1.** Relations in the subscriber database

| Relation | Description |
|---|---|
| **Demographics** | Demographic records (Geographic and population data) |
| **Billing** | Billing records (fee, additional charges, etc) |
| **CDR** | Call detail records (date, time, duration, location, etc ) |

**Table 2**. Distribution of the data used in training and test in the simulation

| Training data set | | Testing data set | |
|---|---|---|---|
| Number of normal examples | Number of churn examples | Number of normal examples | Number of churn examples |
| 2134 | 152 | 824 | 67 |

The simulation data bases are summarized in Table 2. We select 2958 examples (2134 examples for training data set and 824 examples for testing data set) from the data consisted of 100,000 customers for whom there were 171 potential predictor variables. The data were compiled for a three month period and then whether or not the customers churned in the fifth month was recorded.

### 4.3 Experiment Results

We apply our improved one-class SVM to a set of Customer Churn data as described above. We use different Kernel function in SVM and we get different accuracy rate. The result of comparison of different Kernel function is shown in Table 3. Gaussian Kernel function shows the best performance, and the result also indicates that the separating hyper-plane is non-liner.

**Table 3.** Comparison of different Kernel function

| SVM Kernel | Linear | Polynomial | Gaussian |
|---|---|---|---|
| Accuracy rate | 72.28% | 77.65% | 87.15% |

We perform the experiment over the abstract data, by using ANN, Decision Tree, Naïve Bays, and compare them with the improved one-class SVM (Gaussian Kernel) given in this paper. The neural networks used in our experiments are multilayer perceptrons with a single hidden layer which contains 20 nodes and they were trained by the back propagation algorithm with the learning rate was set to 0.3 and the momentum term was set to 0.7. The result is given in Table 4. The experience shows that the improved one-SVM method has the best performance in detecting the churn.

**Table 4.** Comparison of different Algorithm

| Algorithm | ANN | Decision Tree | **Naïve** Bays | SVM (Gaussian) |
|---|---|---|---|---|
| Accuracy rate | 78.12% | 62% | 83.24% | 87.15% |

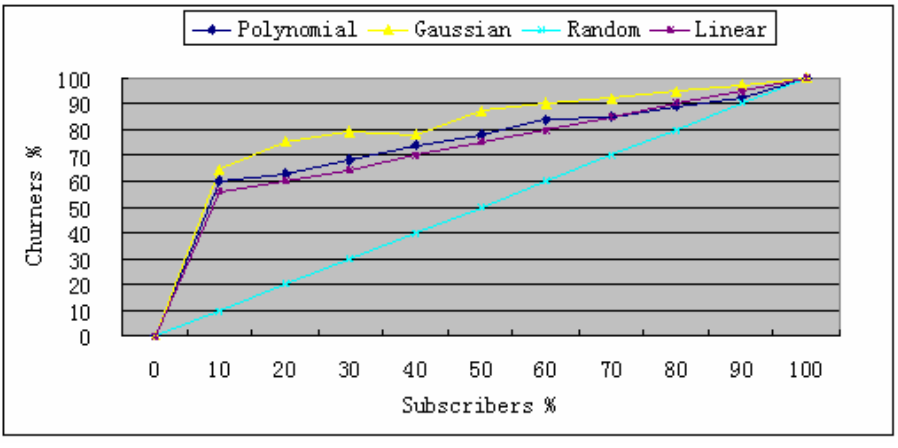


**Fig. 1.** Lift curve of different kernel function (Polynomial, Gaussian and Linear)

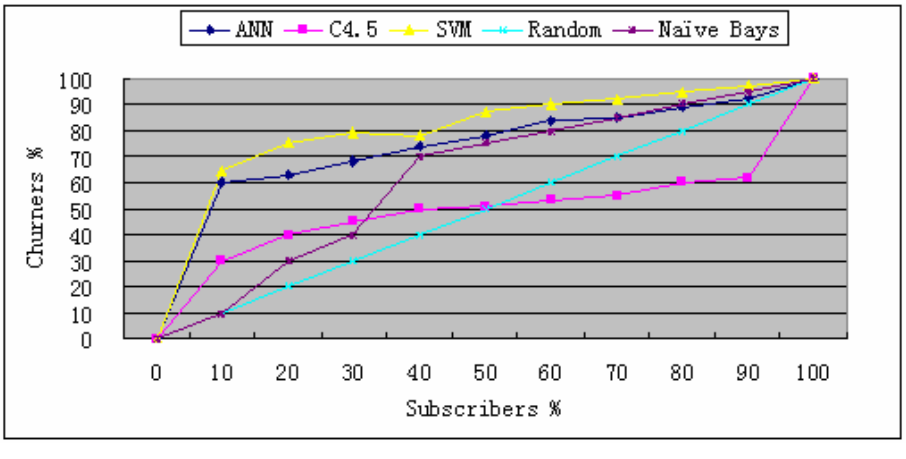


**Fig. 2.** Lift curve of different algorithm: ANN, C4.5, Naïve Bays and SVM

In the telecommunications industry, the "churn" and "no churn" prediction is usually expressed as a lift curve. The lift curve plots the fraction of all churners having churn probability above the threshold against the fraction of all subscribers having churn probability above the threshold. The lift curve indicates the fraction of all churners can be caught if a certain fraction of all subscribers were contacted.

Fig 1 shows the lift curves of different kernel function of improved one-class SVM. SVM with Gaussian kernel function can detect more churners than which with Polynomial and Linear kernel. Fig 2 shows the lift curves of different algorithm.

## 5 Conclusion and Future Work

In this paper, we introduce Customer Churn Prediction and use an improved one-class SVM method to wireless industry data set. The performance of different kernel functions in the improved one-class SVM has been investigated, and the result shows that RBF kennel function get highest accuracy. The classification accuracy of SVM, 87.15%, is better than of ANN, Decision Tree, and Naïve Bays. Support vector machines hold high potential against traditional approaches due to their scalability, faster training and running times. Application of support vector machines to the task of customer churn prediction shows promising results and this work is a contribution to the researches done in the field.

Some more research should be done in how to choose appropriate kernel parameters and input features for better accuracy.

## Acknowledgement

Foundation item: Project supported by the National Natural Science Foundation of China (Grant No.70202008) and the National High-Tech. R&D Program for CIMS, China (Grant No.2003AA414021).

## References

1. Ding-An Chiang, Yi-Fan Wang, Shao-Lun Lee, Cheng-Jung Lin, Goal-oriented sequential pattern for network banking churn, Expert Systems with Applications 25 (2003) 293-302
2. Duke Teradata, Teradata Center for Customer Relationship Management. Retrieved on: Nov 7, (2002).
3. Bernhard scholkopf et al., Estimating the support of a High-Dimensional Distribution, Technical Report, Department of Computer Science, University of Haifa, Haifa, (2001)
4. Scott A. Neslin, Sunil Gupta Wagner Kamakura Junxiang Lu Charlotte Mason, Defection Detection: Improving Predictive Accuracy of Customer Churn Models
5. V. N. Vapnik, The Nature of Statistical Learning Theory. New York: Springer, (1995)
6. Trafalis, Theodore B. Support vector machine for regression and applications to financial forecasting, Proceedings of the International Joint Conference on Neural Networks, v 6, (2000) 348-353
7. B. Scholkopf, J. C. Platt, J. T. Shawe, A. J. Smola, R. C. Williamson, "Estimation the support of a high-dimensional Distribution", Technical Report MSR-TR-99-87, Microsoft Research
8. Kunlun Li, Houkuan Huang, Shengfeng Tian, Wei Xu, Improving one-class SVM for Anomaly detection, Proceedings of the second international conference on machine learning and cybernetics, Xi'an, 2-5 November, (2003)
9. Nath, Shyam V., Behara, Ravi S.Customer churn analysis in the wireless industry A data mining approach, Proceedings - Annual Meeting of the Decision Sciences Institute, (2003) 505-510