

CHURN PREDICTION IN THE TELECOMMUNICATIONS SECTOR USING SUPPORT VECTOR MACHINES

Ionut BRANDUSOIU¹, Gavril TODEREAN²

¹Technical University of Cluj-Napoca, ionut.brandusoiu@gmail.com

²Technical University of Cluj-Napoca, toderean@pro3soft.ro

Abstract—In these days, due to challenges resulted from global competition, customer churn represents one of the significant concerns for companies in different industries. With a churn rate of 30%, the telecommunication sector takes the first place on the list. In order to solve this problem, predictive models need to be implemented to identify customers who are at risk of churning. In this paper, an advanced methodology for predicting customers churn in mobile telecommunications industry is presented. The dataset used, contains call details records and has 21 attributes for each of its 3333 records. We use a Support Vector Machines algorithm with four kernel functions to implement the predictive models. The performance of the models is evaluated and compared using gain measure.

Keywords—Call details records, churn prediction, support vector machines, telecommunications.

I. INTRODUCTION

CUSTOMER retention is considered to be one of the main concerns of companies in the telecommunications industry. With the increasing competition and diversity of offerings on the market, many telecommunications companies take advantage of data mining techniques to predict customer churn [1].

In the mobile telecommunications industry, the churn term, also known as customer attrition or subscriber churning, refers to the phenomenon of loss of a customer [2]. It is measured by the rate of churn and is an important indicator for organizations. This process of movement from one provider to another is usually happening due to better rates or services, or due to different benefits that a competitor company offers when signing up.

In an almost saturated market, companies are using a defensive marketing strategy to keep their existing customers. In order to accomplish this, they need a method that can identify the customers who are most likely to churn, so that they can deploy proactive retention campaigns [3]. To maximize effectiveness and to reduce the high cost involved in these retention campaigns, churn prediction has to be extremely

accurate, to ensure that the incentives reach only the customers who are most likely to change their service providers.

Building a churn prediction model will ease the customer retention process, and in this way the mobile telecommunications companies will success in this constantly increasing competitive market. The churn prediction modeling process is strongly dependent on the data mining process and techniques due to an increased performance generated by machine learning algorithms compared to the statistical techniques for nonparametric data [4].

Data mining is the practice of digging data to find trends and patterns, and can provide you with answers to questions that you should have asked. Data mining methods lie at the intersection of artificial intelligence, machine learning, statistics, and database systems [5]. Data mining techniques can help building prediction models in order to discover future trends and behaviors, allowing organizations to make smart decisions based on knowledge from data.

The methodology used is called modeling. Modeling is the process of creating a mining model, an implementation of specific machine learning algorithms. These algorithms are mathematical functions that perform specific types of analysis on the associated dataset. Based on the business problems that need assistance, these data mining algorithms can be broadly classified into several categories [6]. The category which is of interest for us is called classification analysis. Classification is the process by which a model is built to categorize pre-classified instances (training examples) into classes. This classifier (classification model) is then used to classify future instances.

This paper is organized as follows. In Section 2, we describe the support vector machines algorithm that is later used to build the predictive model. In Section 3, we present the methodology and test the model using four commonly used kernel functions on a data set containing call details records. Finally, Section 4 concludes this research work.

II. SUPPORT VECTOR MACHINES

In this section, we present the support vector machines algorithm that we will use to build the predictive model in order to solve the subscribers churning problem.

The Support Vector Machines (SVM) algorithm is proposed by Boser, Guyon and Vapnik [7]. The SVM method generates a mapping function from a set of labeled training data given as $(\mathbf{x}_i, y_i)$, $\mathbf{x} \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$. The kernel functions transform the input data to a high-dimensional feature space, and thus the input data can be categorized, even when data are not linearly separable. The decision boundary with the largest margin could be drawn as a hyperplane. The model thus depends on only a subset of the training data near the class boundaries [8]. To illustrate the principle of SVM algorithm, in Fig. 1. we consider the case of a two-dimensional input space, i.e., $\mathbf{x} \in \mathbb{R}^2$.

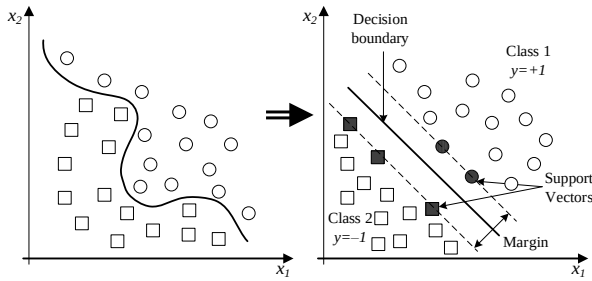


Fig. 1. The principle of SVM algorithm.

The main task of this algorithm implies finding the decision function (1) and, in order to do so, we need to solve the quadratic optimization problem (2)

$$f(x) = \text{sgn} \left(\sum_{i=1}^l y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \right) \quad (1)$$

where $\mathbf{x}_i$ represents the training vector, $y_i \in \{-1, 1\}$ the class label for i^{th} training sample, l the number of training samples, $K(\mathbf{x}_i, \mathbf{x}_j)$ the kernel function for the pair of samples i, j , α_i the coefficients for training samples, and b a constant.

$$\min f(\alpha) = \frac{1}{2} \alpha^T Q \alpha - \mathbf{e}^T \alpha \quad (2)$$

subject to $0 < \alpha_i < C$, $i = 1, \dots, l$, $\mathbf{y}^T \alpha = 0$, where $\mathbf{e}$ is the vector of all ones, C is the upper bound of all variables (chosen by the user, also known as regularization parameter), $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_l)^T$, Q is a $l \times l$ kernel matrix and $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$.

To train the SVM, which means to solve the above quadratic optimization problem (2), we use the fast algorithm proposed by Dong, Krzyzak, and Suen (2005) [9]. This algorithm is based on a divide-and-conquer approach, which divides the original problem into a set of small subproblems that can be solved by the sequential minimal optimization (SMO) algorithm [10]. For each

subproblem, the kernel matrix can be stored in a kernel cache defined as part of contiguous memory. The size of the kernel matrix will be large enough to hold all the support vectors in the whole training set and small enough to satisfy the memory constraint.

The SMO algorithm modifies a subset (working set B) of α per iteration, leading this way to a smaller problem to be minimized in each iteration. This algorithm restricts B to have only two elements. For the SMO decomposition method we use the working set selection using second order information algorithm proposed by Fan, Chen, and Lin (2005) [11]. This method determines the speed of convergence for the algorithm.

When the quadratic optimization problem is solved, we obtain an optimal hyperplane in a high-dimensional feature space to separate the data. We can get different support vector coefficients α_i :

- If $0 < \alpha_i < C$, the corresponding training sample is a *free support vector*.
- If $\alpha_i = C$, the corresponding training sample is a *boundary support vector*.
- If $\alpha_i = 0$, the corresponding training sample is a *non-support vector*.

When building the model, we will take into consideration four most commonly used kernel functions:

- Radial Basis Function kernel (RBF):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (3)$$

- Linear kernel (LIN):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \quad (4)$$

- Polynomial kernel (POL):

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j + r)^d \quad (5)$$

- Sigmoid kernel (SIG):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j + r) \quad (6)$$

For continuous input variables, we linearly scale each attribute to $[-1, 1]$ using the equation

$$V^{new} = \frac{V - V_{\min}}{V_{\max} - V_{\min}} (newmax - newmin) + newmin \quad (7)$$

For categorical input fields, if we have m categories, we use $(0, 1, 2, \dots, m)$ to represent the categories and scale the values as for continuous input variables.

For the output classes, the SVM model generates predictions and predicted probabilities. The predictions are based on the category with the highest predicted probability for each record. To choose a predicted value,

posterior probabilities are approximated using a sigmoid function as proposed by Platt (2000) [12].

In the next section we present the methodology used to build the predictive model using a dataset that contains call details records.

III. METHODOLOGY

In this section we will present a methodology which the data mining approach proposed in this paper is based on. The first part of this section implies understanding and preparing the dataset and the second one building and evaluating the models.

A. Data Preparation

The dataset is from University of California, Department of Information and Computer Science, Irvine, CA [13]. This dataset contains historical records of customer churn, how they turned out in hindsight, i.e., their previous behavior – if it turned out that they are churners or not.

This dataset has a total number of 3333 subscribers with 21 variables each. For each of these subscribers we can find information about their corresponding inbound/outbound calls count, inbound/outbound SMS count, and voice mail.

For the data preparation and model building part we decided to use IBM SPSS (Statistical Product and Service Solutions), a statistical and data mining software package used to build predictive models [14].

By exploring the data we discovered that this is a complete dataset, meaning that for each subscriber there is no attribute missing. We also discovered that between some variables there is a perfect correlation with the R-squared statistic precisely 1. We have four *charge* variables which are linear functions of the *minutes* variables, so we decided to arbitrarily eliminate all four of them to avoid incoherent results. The *area code* variable contains only three different values (from CA only) for all the records, while the *state* variable contain all 51 states, so we decided not include these two variables as well, as it can be bad data. The *phone* variable has been eliminated because it does not contain relevant data that can be used for prediction; it is useful only for identification purposes. We have therefore reduced the number of predictors from 20 to 13. The target variable is *Churn*, which has two values, one of them for each subscriber: *yes* or *no*, telling if a subscriber is a churner or not. The following variables together with their desired corresponding type are used: categorical – *international plan*, *voice mail plan*; and continuous – *account length*, *# voice mail messages*, *# day minutes*, *# day calls*, *# evening minutes*, *# evening calls*, *# night minutes*, *# night calls*, *# international minutes*, *# international calls*, *customer service calls*.

B. Model Building

This phase involves choosing the modeling algorithms and the modeling architecture, and finally building and

evaluating the model. Because we intend to build a classification model based on SVM we must train the algorithm first and then test it, and therefore we must partition the dataset in two: a training and a testing set [15]. After multiple partitioning attempts we concluded that the best performance is obtained if we randomly partition the training set to be approximately 80% of the original data set, consisting of 2651 subscribers, and the testing set to be approximately 20%, consisting of 682 subscribers.

The churn distribution within both sets is as follows: in the testing set we have 105 churners (15%) and 577 non-churners (85%), and in the training set we have 378 churners (14%) and 2273 non-churners (86%). Because the SVM algorithm requires to have a nearly equal distribution of “yes-es” and “no-s” to train properly, we need to clone the “yes-es” until they are approximately equal to the “no-s” by balancing by boosting [16] the training set. Therefore, we will obtain a distribution of 2273 records for churners and 2273 records for non-churners, a total of 4546 records within the training set.

To train the SVM we use the algorithm presented in Section 2 with four kernel functions (see Section 2). We tested multiple configurations of all four SVM algorithms and the architectures of the predictive models that drove to better performance, are listed in Table 1.

TABLE I
ARCHITECTURES FOR SVM

Parameter	Kernel Functions			
	RBF	LIN	POL	SIG
C	10	9	3	3
Gamma	0.29	-	1	2.5
Bias	-	-	0	0.25
Degree	-	-	3	-

We can visualize the performance of the models on the testing set, by looking at the confusion matrix in Table 2. The cells on the diagonal of the cross-classification of cases are correct predictions, whilst those off the diagonal are incorrect predictions.

TABLE II
CONFUSION MATRIX FOR SVM

Kernel Type	Observed (Churn?)	Predicted		
		No	Yes	% correct
RBF	No	498	79	86.31%
	Yes	19	86	81.90%
	Overall %	75.8%	24.2%	85.63%
LIN	No	448	129	77.64%
	Yes	22	83	79.05%
	Overall %	68.9%	31.1%	77.86%
POL	No	519	58	89.95%
	Yes	20	85	80.95%
	Overall %	79%	21%	88.56%
SIG	No	406	171	70.36%
	Yes	30	75	71.43%
	Overall %	63.9%	36.1%	70.53%

This suggest that our first three SVM models that use the RBF, linear, and polynomial kernel functions, will correctly classify 8 out of 10 subscribers as churners. The

last model which uses the sigmoid kernel function performs the worst (7 out of 10).

Another way of comparing the performance of our SVM models is based on gain measure. The gain chart shows on the vertical axis the percentage of positive responses, while on the horizontal axis the percentage of customer contacted. Gains are defined as the proportion of respondents that occurs in each percentile, relative to all responders in the whole sample. The cumulative gain chart shows how better the prediction rate produced by our models is, compared to the random expectation (diagonal line).

Fig. 2. depicts the curves corresponding to SVM models with all four kernel functions. One can find that the models that use RBF and polynomial kernel functions perform better than the other two in the 20th percentile (approximately 80% for RBF and POL, compared to 60% for LIN and 50% for SIG).

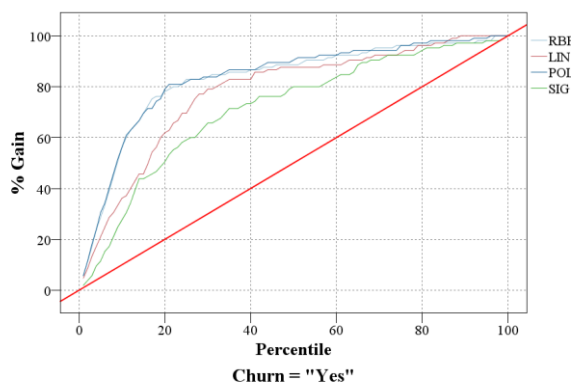


Fig. 2. Cumulative gain chart for SVM.

If, for instance, a mobile telecommunications company wants to send offers containing incentives to stay with them, they can easily select, for example, the top 20% subscribers (20th percentile), and expect that 80% of the contacted subscribers to be churners (if they use the SVM models with RBF and Polynomial kernel functions).

IV. CONCLUSION

In this paper we built four predictive models for subscribers churn in mobile telecommunications companies, using SVM algorithm with different kernel functions. By evaluating the results, from the technical point of view, we observe that for predicting both churners and non-churners, the model that uses the polynomial kernel function performs best, having an overall accuracy of 88.56%.

From the practical point of view, the first three models (RBF, LIN, and POL) have a very good performance (around 80%) in predicting churners in a mobile telecommunications company. Decision making

employees can build different marketing approaches to retain churners based on the predictors that have higher importance in scoring the model performance. All these churn prediction models can be used in other customer response models as well, such as cross-selling, up-selling, or customer acquisition.

REFERENCES

- [1] M. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge, "Knowledge management and data mining for marketing," *Decision Support Systems*, Vol. 31, no. 1, 2001, pp. 127-137.
- [2] C. P. Wei and I. T. Chiu, "Turning telecommunications call details to churn prediction: a data mining approach," *Expert Systems with Applications*, Vol. 23, 2002, pp. 103-112.
- [3] J. H. Ahn, S. P. Han, and Y. S. Lee, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," *Telecommunications Policy*, Vol. 30, Issues 10-11, 2006, pp. 552-568.
- [4] V. García, A. I. Marqués, and J. S. Sánchez, "Non-parametric Statistical Analysis of Machine Learning Methods for Credit Scoring," *Advances in Intelligent Systems and Computing*, Volume 171, 2012, pp. 263-272.
- [5] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, and W. Wang, "Data Mining Curriculum: A Proposal," Version 1.0, 2006.
- [6] F. Gorunescu, *Data Mining Concepts, Models and Techniques*, Springer-Verlag Berlin Heidelberg, 2011.
- [7] B. E. BOSER, I. M. GUYON, and V. N. VAPNIK, "A training algorithm for optimal margin classifiers," In: *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, New York, NY, USA: ACM Press, 1992, pp. 144-152.
- [8] L. Wang, *Support Vector Machines: Theory and Applications*, Studies in Fuzziness and Soft Computing, Volume 177, Springer, 2005.
- [9] J. Dong, C. Y. Suen, and A. Krzyzak, "Fast SVM training algorithm with decomposition on very large data sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 2005, pp. 603-618.
- [10] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," In *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. J. C. Burges, and A. J. Smola, editors, Cambridge, MA, MIT Press, 1998.
- [11] R. E. Fan, P. H. Chen, and C. J. Lin, "Working set selection using the second order information for training SVM. Technical Report," Taipei, Taiwan: Department of Computer Science, National Taiwan University, 2005.
- [12] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," In: *Advances in Large Margin Classifiers*, A. J. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, eds. Cambridge, MA: MIT Press, 2000, pp. 61-74.
- [13] C. L. Blake and C. J. Merz, "Churn Data Set", *UCI Repository of Machine Learning Databases*, <http://www.ics.uci.edu/~mlearn/MLRepository.html>, University of California, Department of Information and Computer Science, Irvine, CA, 1998.
- [14] IBM SPSS Modeler, IBM Corporation, <http://www-01.ibm.com/software/analytics/spss/>
- [15] S. Sumathi and S. N. Sivanandam, *Introduction to Data Mining and its Applications*, Studies in Computational Intelligence, Vol. 29, Springer, 2006.
- [16] R. Nisbet, J. Elder IV, and G. Miner, *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press, 2009.