Simulation Modelling Practice and Theory 55 (2015) 1–9

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

A comparison of machine learning techniques for customer churn prediction

T. Vafeiadis [a,*], K.I. Diamantaras [b], G. Sarigiannidis [a], K.Ch. Chatzisavvas [a]

[a] mSensis S.A., VEPE Technopolis, Bld C2, P.O. Box 60756, GR-57001 Thessaloniki, Greece
[b] Department of Information Technology, TEI of Thessaloniki, GR-57400 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 13 January 2015
Received in revised form 20 February 2015
Accepted 10 March 2015
Available online 3 April 2015

Keywords:
Churn prediction
Machine learning techniques
Boosting algorithm

ABSTRACT

We present a comparative study on the most popular machine learning methods applied to the challenging problem of customer churning prediction in the telecommunications industry. In the first phase of our experiments, all models were applied and evaluated using cross-validation on a popular, public domain dataset. In the second phase, the performance improvement offered by boosting was studied. In order to determine the most efficient parameter combinations we performed a series of Monte Carlo simulations for each method and for a wide range of parameters. Our results demonstrate clear superiority of the boosted versions of the models against the plain (non-boosted) versions. The best overall classifier was the SVM-POLY using AdaBoost with accuracy of almost 97% and *F*-measure over 84%.

1. Introduction

Customer Relationship Management (CRM) is a comprehensive strategy for building, managing and strengthening loyal and long-lasting customer relationships. It is broadly acknowledged and extensively applied to different fields, e.g., telecommunications, banking and insurance, retail market, etc. One of its main objectives is customer retention. The importance of this objective is obvious, given the fact that the cost for customer acquisition is much greater than the cost of customer retention (in some cases it is 20 times more expensive [1]). Thus, tools to develop and apply customer retention models (churn models) are required and are essential Business Intelligence (BI) applications. In the dynamic market environment, churning could be the result of low-level customer satisfaction, aggressive competitive strategies, new products, regulations, etc. Churn models aim to identify early churn signals and recognize customers with an increased likelihood to leave voluntarily. Over the last decade there has been increasing interest for relevant studies in areas including telecommunication industry [2–10], banking [1,11–13], insurance companies [14], and gaming [15], among others. Several, very popular in research community machine learning algorithms have been proposed in order to tackle the churning prediction problem. Such methods include Artificial Neural Networks [4,5,16–18], Decision Trees learning [4,5,7,9,16,17], Regression Analysis [16], Logistic Regression [5,9], Support Vector Machines [17], Naïve Bayes [4,19], Sequential Pattern Mining and Market Basket Analysis [20], Linear Discriminant Analysis [13], and Rough Set Approach [21].

This work constitutes a comparison of five of the most widely used classification methods on the problem of customers' churning in the telecommunication sector. In particular, we compare the performance of multi-layer Artificial Neural

* Corresponding author.
E-mail addresses: thanvaf@gmail.com (T. Vafeiadis), kdiamant@it.teithe.gr (K.I. Diamantaras), g.sarigiannidis@msensis.com (G. Sarigiannidis), kchatz@msensis.com (K.Ch. Chatzisavvas).

http://dx.doi.org/10.1016/j.simpat.2015.03.003
1569-190X/

Networks, Decision Trees, Support Vector Machines, Naïve Bayes classifiers, and Logistic Regression classifiers, compared to their boosting versions in an attempt to further improve their performance. The motivation behind our study is to evaluate the suitability of the state of the art machine learning methods on the problem of churning. This investigation is performed using Monte Carlo simulation at different settings of each classification method. We use the churn dataset originally from the UCI Machine Learning Repository (converted to MLC++ format[1]), which is now included in the package *C50* of the *R* language,[2] in order to test the performance of classification methods and their boosting versions. Data are artificial based on claims similar to the real world. The dataset has been used in numerous publications [22–28].

The remainder of the paper is organized as follows. In Section 2, we give a brief presentation of the machine learning techniques that were evaluated. Evaluation criteria and the proposed boosting algorithm are presented in Sections 3 and 4. The simulation setup and results are given in Section 5, and in Section 6 we draw our conclusions.

2. Machine learning techniques-classification methods

In the following, we briefly present five well established and popular techniques used for churn prediction, taking into consideration reliability, efficiency and popularity in the research community [4,7,9,16,17,19,29,30].

2.1. Artificial Neural Network

Artificial Neural Networks (ANNs) is a popular approach to address complex problems, such as the churn prediction problem. Neural networks can be hardware-based (neurons are represented by physical components) or software-based (computer models), and can use a variety of topologies and learning algorithms. One popular supervised model is the Multi-Layer Perceptron trained with variations of the Back-Propagation algorithm (BPN). BPN is a feed-forward model with supervised learning. In the case of the customer churn problem, Au et al. [31] have shown that neural networks achieve better performance compared to Decision Trees. Also, experimental results showed that ANN outperformed Logistic Regression and C5.0 for churn prediction [30].

2.2. Support Vector Machines

Support Vector Machines (SVMs), also known as Support Vector Networks, introduced by Boser, Guyon, and Vapnik [32], are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. SVM is a machine learning technique based on structural risk minimization. Kernel functions have been employed for improving performance [4]. Research on selecting the best kernels or combinations of kernels is still under way. In the churn prediction problem, SVM outperform DT and sometimes ANN, depending mainly on the type of data and data transformation that takes place among them [29,17].

2.3. Decision trees learning

Decision Trees (DTs) are tree-shaped structures representing sets of decisions capable to generate classification rules for a specific dataset [33], or as Berry and Linoff noted *"a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules"* [34]. More descriptive names for such tree models are Classification Trees or Regression Trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. DT have no great performance on capturing complex and non-linear relationships between the attributes. Yet, in the customers churn problem, the accuracy of a DT can be high, depending on the form of the data [16].

2.4. Naïve Bayes

A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. In simple terms, a Naïve Bayes (NB) classifier assumes that the presence (or absence) of a particular feature of a class (i.e., customer churn) is unrelated to the presence (or absence) of any other feature. The NB classifier achieved good results on the churn prediction problem for the wireless telecommunications industry [19] and it can also achieve improved prediction rates compared to other widely used algorithms, such as DT-C4.5 [4].

2.5. Regression analysis-logistic regression analysis

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or

[1] Available at http://www.sgi.com/tech/mlc/db/.
[2] http://cran.r-project.org/web/packages/C50.