Determination of Algorithms
Making Balance Between Accuracy and Comprehensibility in Churn Prediction Setting

Hossein Abbasimehr, K. N. Toosi University of Tech, Iran
Mohammad Jafar Tarokh, K. N. Toosi University of Tech, Iran
Mostafa Setak, K. N. Toosi University of Tech, Iran

ABSTRACT

Predictive modeling is a useful tool for identifying customers who are at risk of churn. An appropriate churn prediction model should be both accurate and comprehensible. However, reviewing the past researches in this context shows that much attention is paid to accuracy of churn prediction models than comprehensibility of them. This paper compares three different rule induction techniques from three categories of rule based classifiers in churn prediction context. Furthermore logistic regression (LR) and additive logistic regression (ALR) are used. After parameter setting, eight distinctive algorithms, namely C4.5, C4.5 CP, RIPPER, RIPPER CP, PART, PART CP, LR, and ALR, are obtained. These algorithms are applied on an original training set with the churn rate of 30% and another training set with the churn rate of 50%. Only the models built by applying these algorithms on a training set with the churn rate of 30% make balance between accuracy and comprehensibility. In addition, the results of this paper show that ALR can be an excellent alternative for LR, when models only from accuracy perspective are evaluated.

Keywords: C4.5, Churn Prediction, PART, RIPPER, Rule Based Classifier

1. INTRODUCTION

The importance of Customer relationship management (CRM) is well known for firms in industries with high competitive and saturated market. One of the key components of CRM is customer retention. Customer retention is an important strategy to keep and satisfy existing customers. Existing customer is a valuable asset for most companies (Athanassopoulos, 2000; Jones, Mothersbaugh, & Beatty, 2000; Thomas, 2001). Acquisition of new customer costs more than keeping existing customers (Coussement & Van den Poel, 2008a). It is stated that attracting new customer costs 12 times more than retain-
ing existing one (Coussement, Benoit, & Van den Poel, 2010).

Benefits of increasing customer loyalty or retention with in a current customer base are:

1. There is less need to offer incentive to loyal customers
2. Loyal customers are less price sensitive
3. Loyal customer will recommend the company to other people
4. Individual revenue of each customer grows as trust increase (Chaffey, Ellis-Chadwick, Mayer, & Johnston, 2006)

To keep existing customer, a company must identify likely churners (customers tend to stop their business with company). Churn prediction is a useful tool to predict customer at churn risk.

Finding churn drivers of customer churn or model building for customer churn prediction are the aims of researches in this field (Coussement & Van den Poel, 2009). In past researches, many modeling techniques had been used for customer churn prediction; however, in this regards, only the accuracy of the constructed models is considered and less attention has been paid to the comprehensibility of the developed model (Verbeke, Martens, Mues, & Baesens, 2011). Accuracy is not the only important criterion in evaluating a churn prediction model (Verbeke et al., 2011). Comprehensibility of model causes it to disclose some valuable knowledge about churn drivers of customers. Such knowledge can be extracted in the form of “if then” rules which allows marketing managers to develop right strategies for retaining the churners (Verbeke, Martens, Mues, & Baesens, 2011). Also, being accurate, a model can recognize churner and non-churner well. In this study, we have concentrated on both accuracy and comprehensibility of churn prediction models.

Data mining techniques had been used widely in churn prediction context. From data mining view point, churn prediction is a binary classification task. The aim of churn prediction model is classifying of customers into two classes as churner and non-churner (Coussement, Benoit, & Van den Poel, 2010).

Basic techniques of data classification are decision tree classifiers, Bayesian classifiers, Bayesian belief networks and rule based classifiers, k-nearest-neighbor classifiers, case-based reasoning, neural network technique and so on (Han & Kamber, 2006).

In this study, the main motivation is the comparison of three different algorithms from three categories of rule-based classifier in churn prediction context. These algorithms include: C4.5 decision tree (Han & Kamber, 2006; Tan, Steinbach, & Kumar, 2005; Witten & Frank, 2005) is belonging to divide-and-conquer algorithms category, RIPPER (Reduced Incremental Pruning to produce Error Reduction) (Han & Kamber, 2006; Tan, Steinbach, & Kumar, 2005; Witten & Frank, 2005) which is belonging to separate-and-conquer algorithms, PART (Witten & Frank, 2005) algorithm which is a combinational method of divide-and-conquer and separate-and-conquer algorithms. Although divide-and-conquer and separate-and-conquer strategies previously had been used by researchers in churn prediction setting, to the best of our knowledge this is the first study in a customer churn context that evaluates and uses algorithms, which are a combinational method of divide-and-conquer and separate-and-conquer algorithms. Although divide-and-conquer and separate-and-conquer strategies previously had been used by researchers in churn prediction setting, to the best of our knowledge this is the first study in a customer churn context that evaluates and uses algorithms, which are a combinational method of divide-and-conquer and separate-and-conquer algorithms. The performance of these algorithms is measured by both accuracy and comprehensibility metrics. Moreover, we have used Logistic regression and additive logistic regression. Besides in this paper we have identified a good class distribution for mining churn data.

The reminder of this paper is organized as follows. Firstly, a literature review about churn prediction is presented in Section 2. Executed methods are described in Section 3. In Section 4, the evaluation metrics is illustrated. In Section 5 the data preprocessing, model building and results of a series of experiments are discussed. Conclusions are considered in Section 6.
Outline Shape Retrieval Using Textual Descriptors and Geometric Features
www.igi-global.com/article/outline-shape-retrieval-using-textual-descriptors-and-geometric-features/90442?camid=4v1a