Applied Soft Computing 35 (2015) 740–748

Contents lists available at ScienceDirect

Applied Soft Computing

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

Profit-based feature selection using support vector machines – General framework and an application for customer retention

Sebastián Maldonado[a,*], Álvaro Flores[b], Thomas Verbraken[c], Bart Baesens[c,d,e], Richard Weber[b]

[a] Universidad de los Andes, Mons. Álvaro del Portillo, 12455 Las Condes, Santiago, Chile
[b] Department of Industrial Engineering, FCFM, University of Chile, Républica 701, Santiago, Chile
[c] Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[d] School of Management, University of Southampton, United Kingdom
[e] Vlerick Leuven-Gent Management School, Belgium

ARTICLE INFO

Article history:
Received 10 August 2014
Received in revised form 2 February 2015
Accepted 16 May 2015
Available online 4 July 2015

Keywords:
Data mining
Feature selection
Support vector machines
Churn prediction
Customer retention
Maximum profit

ABSTRACT

Churn prediction is an important application of classification models that identify those customers most likely to attrite based on their respective characteristics described by e.g. socio-demographic and behavioral variables. Since nowadays more and more of such features are captured and stored in the respective computational systems, an appropriate handling of the resulting information overload becomes a highly relevant issue when it comes to build customer retention systems based on churn prediction models. As a consequence, feature selection is an important step of the classifier construction process. Most feature selection techniques; however, are based on statistically inspired validation criteria, which not necessarily lead to models that optimize goals specified by the respective organization. In this paper we propose a profit-driven approach for classifier construction and simultaneous variable selection based on support vector machines. Experimental results show that our models outperform conventional techniques for feature selection achieving superior performance with respect to business-related goals.

1. Introduction

Classification is a very relevant task in many profit-driven applications, such as e.g. credit scoring or customer retention [3]. It has been shown that the performance of a classifier can be improved by concentrating on the most relevant features used for classifier construction. Such variable selection has important advantages: first, a low-dimensional representation of the objects enhances the predictive power of classification models by decreasing their complexity. Having less features also leads to more parsimonious models which in turn contributes to reduce the risk of overfitting [9] caused by the *curse of dimensionality* [19,30].

Additionally, it allows a better interpretation of the classifier, which is particularly important in business analytics. Many machine learning approaches are usually labeled as *black boxes* by practitioners, who therefore tend to be reluctant to use the respective methods [10]. A better understanding of the process that generates the data is therefore of crucial importance in business analytics for decision-making, e.g., by identifying those attributes that permit explaining customers' decisions [7].

In the past, statistically inspired techniques have been the most frequently used approaches to validate both classifiers as well as feature selection methods. Recently, profit-based measures have been suggested for classifier validation [35]. In this paper we go one step further and adapt the idea of profit-driven metrics also to the task of feature selection by introducing several embedded methods combining the method Holdout support vector machine (HOSVM) [26] with various validation measures.

To the best of our knowledge, profit-driven feature selection is a novel approach that has not yet been covered in the data mining and machine learning literature. Most of the work in business analytics and feature selection applies traditional, statistically grounded techniques without taking into account profit-related issues. Our experiments underline that the proposed methods outperform alternative techniques and provide classifiers with highly relevant features, thus reducing the risk of overfitting while increasing the related profit at the same time.

The remainder of the paper is organized as follows: Section 2 describes the cost benefit analysis in the context of customer retention. Section 3 presents support vector machines for classification

[*] Corresponding author. Tel.: +56 226181874.
E-mail address: smaldonado@uandes.cl (S. Maldonado).

http://dx.doi.org/10.1016/j.asoc.2015.05.058
1568-4946/© 2015 Elsevier B.V. All rights reserved.

and the feature selection techniques studied in this work. The proposed profit-based approach for feature selection and classification is presented in Section 4. Section 5 provides experimental results using real-world datasets. A summary of this paper can be found in Section 6, where we provide its main conclusions and address future developments.

2. The cost benefit analysis framework for customer retention

Trying to retain customers that are about to leave the company is one of the most important tasks in the service industry, mainly in the banking and telecommunications sector. This is driven by the increasing number of customers willing to change their provider, and the strong competition for attracting new ones. Therefore, there is an urgent need to develop and apply accurate models in order to identify current customers who are most likely to leave the company in a given period of time. Churn can be observed in two different ways, *voluntary*, meaning that the customer decides to terminate the contract, or *involuntary*, where the company decides to finish the contract with the customer [4]. In the present work we focus on churn as a voluntary decision.

If a company is able to identify potential churners, the next step is to develop marketing campaigns, and retention strategies focusing on this particular group, thus enhancing customer loyalty and leading to major benefits, such as e.g.:

- Loyal engaged customers, can generate 1.7 times more revenue than other customers [18].
- A direct impact on profitability: a 5% increment in the customer retention rate may lead to a 18% reduction in operational costs [18].
- A decrease of money misspending, focusing resources on churn candidates instead of the whole customer database, reducing marketing and operational costs [15].

According to this, the churn rate is explicitly included in the following customer lifetime value (CLV) formula [4]:

$$CLV = \sum_{t=1}^{\infty} \frac{m(1-c)^{t-1}}{(1+r)^{t-1}} = m\frac{(1+r)}{(r+c)} \tag{1}$$

where c is the annual churn rate and m stands for the mean of the annual profit contribution per customer. Parameter r is the annual discount rate. There are two classical approaches to determine this value. The first one is the company's weighted average cost of capital (WACC). The second one is to use the discount rate of the particular industrial sector. Given this formula, and understanding the CLV as the net present value of the profit for a customer, a decreasing churn rate will impact heavily on the company's profitability.

Churn phenomena can be modeled either with time-dependent techniques [4], or with single period future predictions. In the first category, this kind of models tries to not assume that the churn will occur in a given period, determining probabilities of churning up to a number of months, and taking into consideration time-varying covariates [4]. In the latter, we find approaches aiming to predict if a customer decides to churn in the next period, where the most common approaches are based on statistical methods, such as logistic regression [8,23,29], non-parametric statistical models such as k-nearest neighbor [13], decision trees [39], and other machine learning techniques [15,36]. A review on customer churn prediction modeling can be found in [37]. Here we use SVM classifiers to predict churn in a single period. Churn rates usually are below 5% [35] for this kind of classification models, leading to the class-imbalance problem as will be seen in Section 5.

3. Feature selection for SVM

In this section we present the foundations of SVM for binary classification and the different feature selection strategies available in the literature, and we provide a brief description of each method used in this work.

3.1. Binary classification with support vector machine

Among existing classification methods, support vector machine provides several advantages such as adequate generalization to new objects due to the *structural risk minimization* principle, absence of local minima via convex optimization, and representation that depends on only a few data points (the *support vectors*). All these features reduce the risk of overfitting in classification [34]. Additionally, the introduction of kernel functions for nonlinear classification enhances performance via flexible classifiers, in contrast to traditional techniques such as logistic regression.

Let $\mathcal{F}$ be the feature set, $\mathbf{x}_i \in \Re^{|\mathcal{F}|}$ the feature vector and $y_i \in \{-1, 1\}$ the class label of object i, $i = 1, \ldots, N$. $\mathcal{T} = \{(\mathbf{x}_i, y_i); i = 1, \ldots, N\}$ denotes the training set.

In our case, $\mathbf{x}_i$ is the feature vector describing customer i and y_i indicates his/her class label; churners (non-churners) are identified by $y_i = 1$ ($y_i = -1$).

Linear SVM constructs an optimal hyperplane $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ which tries to correctly separate one class from the other. To achieve this optimal hyperplane, SVM aims to maximize its *margin*, defined as the sum of the distances (with a given metric) between the hyperplane to the closest positive and negative training patterns. This is equivalent to minimizing the Euclidian norm of $\mathbf{w}$ [34]. Given that a perfect separation between the two classes is not always possible, a slack variable ξ_i is introduced for each training vector $\mathbf{x}_i$, $i = 1, \ldots, N$ whereby C is used as a penalization parameter to control the training error [34] as shown in model (2).

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y_i \cdot (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N, \\ & \xi_i \geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{2}$$

The previous formulation can be extended to nonlinear classifiers by using the *kernel trick*: the training samples are mapped into a higher dimensional domain $\mathcal{H}$ through the function $\phi : \mathbf{x} \rightarrow \phi(\mathbf{x}) \in \mathcal{H}$ [31]. A kernel function $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \cdot \phi(\mathbf{y})$ defines an inner product in space $\mathcal{H}$, leading to the following dual formulation:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,s=1}^{N}\alpha_i\alpha_s y_i y_s K(\mathbf{x}_i, \mathbf{x}_s) \\ \text{s.t.} \quad & \sum_{i=1}^{N}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, N. \end{aligned} \tag{3}$$

In this work we use both linear SVM as well as the kernel-based formulation with Gaussian kernel, which usually achieves very good results and is a common choice in the literature [26,31]. This kernel function has the following form:

$$K(\mathbf{x}_i, \mathbf{x}_s) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_s||^2}{2\sigma^2}\right) \tag{4}$$

where $\sigma > 0$ controls the kernel width.