Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 1102–1108
Expert Systems with Applications
www.elsevier.com/locate/eswa

Response modeling with support vector regression

Dongil Kim, Hyoung-joo Lee, Sungzoon Cho *

Seoul National University, San 56-1, Shillim-dong, Kwanak-gu, Seoul 151-744, Republic of Korea

Abstract

Response modeling has become a key factor to direct marketing. In general, there are two stages in response modeling. The first stage is to identify respondents from a customer database while the second stage is to estimate purchase amounts of the respondents. This paper focuses on the second stage where a regression, not a classification, problem is solved. Recently, several non-linear models based on machine learning such as support vector machines (SVM) have been applied to response modeling. However, there is a major difficulty. A typical training dataset for response modeling is so large that modeling takes very long, or, even worse, modeling may be impossible. Therefore, sampling methods have been usually employed in practice. However a sampled dataset usually leads to lower accuracy. In this paper, we employed an ε-tube based sampling for support vector regression (SVR) which leads to better accuracy than the random sampling method.

Keywords: Response modeling; Customer relationship management; Direct marketing; Support vector machines; Regression; Pattern selection

1. Introduction

A response model, given a mailing campaign, predicts whether each customer will respond or how much each customer will spend based on the database of customers' demographic information and/or purchase history. Marketers will send mails or catalogs to customers who are predicted to respond or to spend large amounts of money. A well-targeted mail increases profit while a mistargeted or unwanted mail not only increases marketing cost but also may worsen a firm's relationship with its customers (Gönül, Kim, & Shi, 2000; Potharst et al., 2000). Various methods have been used for response modeling such as statistical techniques (Bentz & Merunka, 2000; Haughton & Oulabi, 1997; Ling & Li, 1998; Suh, Noh, & Suh, 1999), machine learning techniques, (Cheung, Kwok, Law, & Tsui, 2003; Chiu, 2002; Shin & Cho, 2006; Viaene et al., 2001; Wang, Zhou, Yang, & Yeung, 2005; Yu & Cho, 2006) and neural networks (NN) (Bentz & Merunka, 2000; Potharst et al., 2000; Zahavi & Levin, 1997).

In general, there are two stages in response modeling. The first stage is to identify respondents from a customer database while the second stage is to estimate purchase amounts of the respondents. The response modeling task in the first stage has been usually formulated as a binary classification problem. The customers are divided into two classes, respondents and non-respondents. A classifier is constructed to predict whether a given customer will respond or not. However, as pointed out in KDD98 Cup (1998) for the KDD-CUP-98 task, there is an inverse correlation between the likelihood to buy and the dollar amount to spend (Wang et al., 2005). This is because the more dollar amount is involved, the more cautious a customer becomes in making a purchase decision. Hence, one may need a regression model, in a second stage, to estimate the purchase amount of responding customers.

Support vector machine (SVM) has been recently spotlighted with great generalization performances by employing the structural risk minimization (SRM) principle (Vapnik, 1995). Support vector regression (SVR), a regression version of SVM, was developed to estimate regression functions

* Corresponding author. Tel.: +82 2 880 6275; fax: +82 2 889 8560.
E-mail addresses: dikim01@snu.ac.kr (D. Kim), imhjlee@gmail.com (H.-j. Lee), zoon@snu.ac.kr (S. Cho).

0957-4174/$ - see front matter
doi:10.1016/j.eswa.2006.12.019

(Drucker, Burges, Kaufman, Smola, & Vapnik, 1997). Like SVM, SVR is capable of solving non-linear problems using kernel functions and has been successful in various domains (Drucker et al., 1997; Müller et al., 1997). However, there is a difficulty to train SVR on real-world dataset. As the number of training patterns increases, SVR training takes much longer with a time complexity of $O(n^3)$ where n denotes the number of training patterns. So far, many algorithms such as Chunking, SMO, SVM^{light} and SOR have been proposed to reduce the training time. However, their training time complexity is still strongly related to the number of training patterns (Platt, 1999). We take another direction called pattern selection which focuses on reducing the number of training patterns. Neighborhood property based pattern selection (NPPS) proposed by Shin and Cho (2003) is a powerful pattern selection method for SVM, but it is not for regression, but for classification. Recently, a pattern selection method based on the ε-tube (SVR-ε) was proposed which is specifically designed for SVR (Kim & Cho, 2006). Thus, we employ SVR as a response model to predict an amount of money spent by each respondent. One can improve the performance of a response model by identifying profitable respondents instead of just respondents among all customers. Hence, after applying a classification model that predicts likelihoods to buy, one needs a regression model that predicts amounts to buy. Since classification is not our main concern, we assume that a perfect classifier exists which can identify all respondents without false positive (FP) errors. So, an SVR model is constructed on a subset of customers which consists only of respondents. The procedure of constructing the proposed response model is depicted in Fig. 1. In order to reduce the training time of SVR, we employ the pattern selection method. The DMEF4 dataset from the Direct Marketing Educational Foundation (DMEF) is used in our experiments. A small dataset is used to measure improvement in efficiency. However, for a very large dataset, some kind of sampling is inevitable, anyway.

The remaining of this paper is organized as follows. The concept of SVR is briefly introduced and the main idea of the pattern selection method is presented in Section 2. In Section 3, the DMEF dataset and our experimental settings are described in detail. Section 4 presents the experimental results. Section 5 concludes this paper with remarks on limitations and future research directions.

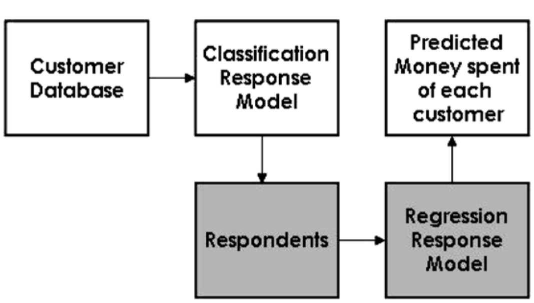

Fig. 1. The procedure of constructing the proposed response model. After identifying respondents using a classification model, dollar amounts to be spent by them are predicted using a regression model. This paper focuses on the two dark boxes.

2. Pattern selection for support vector regression

2.1. Support vector regression

For a brief review of SVR, consider a regression function $f(x)$ which is estimated using a hyperplane $\mathbf{w}$ based on training patterns $\{(\mathbf{x}_i, y_i) | i = 1, 2, \cdots, n\}$ as follows:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \quad \text{with } \mathbf{w}, \mathbf{x} \in \mathbb{R}^N, \; b \in \mathbb{R}, \quad \text{where } \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \in \mathbb{R}^N \times \mathbb{R}. \tag{1}$$

By the SRM principle, the generalization accuracy is optimized over the empirical error and the flatness of the regression function which is guaranteed on a small $\mathbf{w}$. Therefore, the objective of SVR is to include training patterns inside an *ε-insensitive tube* (ε-tube) while keeping the norm $\|\mathbf{w}\|^2$ as small as possible. An optimization problem can be formulated as the following soft margin problem:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \\ &\text{Subject to} \quad y_i - \mathbf{w} \cdot \mathbf{x}_i - b \leqslant \varepsilon + \xi_i, \\ &\qquad \mathbf{w} \cdot \mathbf{x}_i + b - y_i \leqslant \varepsilon + \xi_i^*, \\ &\qquad \xi_i, \xi_i^* \geqslant 0, i = 1, \cdots, n, \end{aligned} \tag{2}$$

where C, ε, and $\xi(\xi^*)$ are a trade-off cost between the empirical error and the flatness, the size of the ε-tube, and slack variables, respectively. By adding Lagrangian multipliers α and α^*, the QP problem can be optimized as a dual problem. Also, SVR can estimate a non-linear function by employing a kernel function, $k(\mathbf{x}_i, \mathbf{x}_j)$. The regression function estimated by SVR can be written as the following kernel expansion:

$$f(\mathbf{x}) = \sum_{i=1}^{ns}(\alpha - \alpha^*)k(\mathbf{x}_i, \mathbf{x}) + b, \tag{3}$$

where ns is the number of support vectors. The RBF kernel is used in this paper.

2.2. Pattern selection for support vector regression

The training time complexity of SVR is $O(n^3)$. If the number of training patterns increases, the training time increases more radically, i.e. in a cubic proportion. It takes too long to train SVR directly with a marketing dataset since marketing databases usually consist of over one millions of customers and hundreds of input variables. Thus, we apply a pattern selection method to reduce the training time of SVR as proposed in our previous research (Kim & Cho, 2006).

SVR learns patterns based on the ε-loss function foundation. SVR makes an ε-tube around the training patterns. The patterns within the ε-tube are not counted as errors, while patterns outside of the ε-tube become support vectors