Expert Systems with Applications 39 (2012) 1687–1697

Contents lists available at ScienceDirect

Expert Systems with Applications

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

QBoost: Predicting quantiles with boosting for regression and binary classification

Songfeng Zheng *

Department of Mathematics, Missouri State University, 901 S. National Ave., Springfield, MO 65897, USA

ARTICLE INFO

Keywords:
Quantile regression
Boosting
Functional gradient algorithm
Binary classification

ABSTRACT

In the framework of functional gradient descent/ascent, this paper proposes Quantile Boost (QBoost) algorithms which predict quantiles of the interested response for regression and binary classification. Quantile Boost Regression performs gradient descent in functional space to minimize the objective function used by quantile regression (QReg). In the classification scenario, the class label is defined via a hidden variable, and the quantiles of the class label are estimated by fitting the corresponding quantiles of the hidden variable. An equivalent form of the definition of quantile is introduced, whose smoothed version is employed as the objective function, and then maximized by functional gradient ascent to obtain the Quantile Boost Classification algorithm. Extensive experimentation and detailed analysis show that QBoost performs better than the original QReg and other alternatives for regression and binary classification. Furthermore, QBoost is capable of solving problems in high dimensional space and is more robust to noisy predictors.

1. Introduction

Classical least square regression aims to estimate the conditional expectation of the response Y given the predictor (vector) $\mathbf{x}$, i.e., $E(Y|\mathbf{x})$. However, the mean value (or the conditional expectation) is sensitive to the outliers of the data (Koenker, 2005). Therefore, if the data is not homogeneously distributed, we expect the traditional least square regression to give us a poor prediction.

The τth quantile of a distribution is defined as the value such that there is $100\tau\%$ of mass on its left side. Compared to the mean value, quantiles are more robust to outliers (Koenker, 2005). Let $I(\cdot)$ be the indicator function with $I(\cdot) = 1$ if the condition is true, otherwise $I(\cdot) = 0$. Let $Q_\tau(Y)$ be the τth quantile of random variable Y. It can be proved (Hunter & Lange, 2000) that

$$Q_\tau(Y) = \arg\min_c E_Y[\rho_\tau(Y - c)],$$

where $\rho_\tau(r)$ is the "check function" (Koenker, 2005) defined by

$$\rho_\tau(r) = rI(r \geqslant 0) - (1 - \tau)r. \tag{1}$$

Given training data $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$, with predictor vector $\mathbf{x}_i \in \mathbf{R}^d$ and response $Y_i \in \mathbf{R}$, let the τth conditional quantile of Y given $\mathbf{x}$ be $f(\mathbf{x})$. Similar to the least square regression, quantile regression (QReg) (Koenker, 2005; Koenker & Bassett, 1978) aims at estimating the conditional quantiles of the response given a predictor vector $\mathbf{x}$ and can be formulated as

$$f^*(\cdot) = \arg\min_f \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - f(\mathbf{x}_i)). \tag{2}$$

Compared to least square regression, quantile regression is robust to outliers in observations, and can give a more complete view of the relationship between predictor and response. Furthermore, least square regression implicitly assumes normally distributed errors, while such an assumption is not necessary in quantile regression. Since being introduced in Koenker and Bassett (1978), quantile regression has become a popular and effective approach to statistical analysis with wide applications in economics (Hendricks & Koenker, 1992; Koenker & Hallock, 2001), survival analysis (Koenker & Geling, 2001), and ecology (Cade & Noon, 2003), to name a few.

The quantile regression model in Eq. (2) can be solved by linear programming algorithms (Koenker, 2005) or Majorize-Minimize algorithms (Hunter & Lange, 2000). However, when the predictor $\mathbf{x}$ is in high dimensional space, the aforementioned optimization methods for QReg might be inefficient. High dimensional problems are ubiquitous in applications such as image analysis, gene sequence analysis, etc. To the best of our knowledge, the problem of high dimensional predictor is not sufficiently addressed in QReg literature, and this paper proposes a method for QReg which can work in high dimensional spaces.

The proposed algorithm for QReg is based on boosting (Freund & Schapire, 1997), which is well known for its simplicity and good performance. The powerful feature selection mechanism of boosting makes it suitable to work in high dimensional spaces. Friedman, Hastie, and Tibshirani (2000) developed a general statistical framework which yields a direct interpretation of boosting as a

* Tel.: +1 417 836 6037; fax: +1 417 836 6966.
E-mail address: SongfengZheng@MissouriState.edu

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.06.060

method for function estimation, which is a "stage-wise, additive model".

Consider the problem of function estimation

$$f^*(\mathbf{x}) = \arg\min_f E[l(Y, f(\mathbf{x}))|\mathbf{x}],$$

where $l(\cdot,\cdot)$ is a loss function which is typically differentiable and convex with respect to the second argument. Estimating $f^*(\cdot)$ from the given data, $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$, can be performed by minimizing the empirical loss, $n^{-1}\sum_{i=1}^n l(Y_i, f(\mathbf{x}_i))$, and pursuing iterative steepest descent in functional space. This leads us to the generic functional gradient descent algorithm (Friedman, 2001; Mason, Baxter, Bartlett, & Frean, 2000). Fig. 1 shows the version summarized in Bühlmann and Hothorn (2007).

Many boosting algorithms can be understood as functional gradient descent with appropriate loss function. For example, if we choose $l(Y, f) = \exp(-(2Y-1)f)$, we would recover AdaBoost (Friedman et al., 2000), and L_2 Boost (Bühlmann & Yu, 2003) corresponds to $l(Y, f) = (Y - f)^2/2$.

Motivated by the gradient boosting algorithms (Friedman, 2001; Mason et al., 2000), this paper estimates the quantile regression function by minimizing the objective function in Eq. (2) with functional gradient descent. In each step, we approximate the negative gradient of the objective function by a base function, and grow the model in that direction. This results in the Quantile Boost Regression (QBR) algorithm. In the binary classification scenario, we define the class label via a hidden variable, and the quantiles of the class label can then be estimated by fitting the corresponding quantiles of the hidden variable. An equivalent form of the definition of quantile is introduced, whose smoothed version is employed as the objective function for classification. Similar to QBR, functional gradient ascent is applied to maximize the objective function, which yields the Quantile Boost Classification (QBC) algorithm. The obtained Quantile Boost (QBoost) algorithms are computationally efficient and converge to local optima. More importantly, they enable us to solve high dimensional problems efficiently.

The rest of this paper is organized as follows: in Section 2, we first apply the functional gradient descent to QReg, yielding the QBR algorithm, and then we proceed to propose the approximation of the objective function for binary classification and to maximize the objective function with functional gradient ascent in order to obtain the QBC algorithm; Section 3 discusses some computational issues in the proposed QBR and QBC algorithms and introduces implementation details; Section 4 presents the experimental results of the proposed QBR and QBC algorithms on benchmark regression and binary classification datasets, and in-depth discussions of the results are also presented in Section 4; finally, Section 5 summarizes this paper with a brief discussion for future research directions.

2. Methods

The methods proposed in our research are presented in this section. We first directly apply the functional gradient descent to the quantile regression model, yielding the quantile boost regression algorithm. We then propose a smooth approximation to the optimization problem for the quantiles of binary response, and based on this we further propose the quantile boost classification algorithm with some discussions of the related methods.

2.1. Quantile boost regression

We consider the problem of estimating quantile regression function in the general framework of functional gradient descent with the loss function

$$l(Y, f) = \rho_\tau(Y - f) = (Y - f)I(Y - f \geqslant 0) - (1 - \tau)(Y - f).$$

A direct application of the algorithm in Fig. 1 yields the Quantile Boost Regression (QBR) algorithm, which is shown in Fig. 2.

Similar to AdaBoost, QBR enables us to select most informative predictors if an appropriate base learner is employed, and this will be demonstrated experimentally in Section 4.1.1.

There is a large volume of literature applying boosting to regression problems, for example in Duffy and Helmbold (2002), Freund and Schapire (1997), and Zemel and Pitassi (2001). However, all these methods estimate the mean value of the response, not quantiles. Langford, Oliveira, and Zadrozny (2006) proposed to use classification technique for estimating the conditional

0. Set iteration number $m = 0$, initialize $f^{[0]}(\cdot) = 0$ or

$$f^{[0]}(\cdot) = \arg\min_c \frac{1}{n}\sum_{i=1}^n l(Y_i, c).$$

1. Increase m by 1. Compute the negative gradient $-\frac{\partial}{\partial f} l(Y, f)$ at $f^{[m-1]}(\mathbf{x}_i)$:

$$U_i = -\left.\frac{\partial l(Y_i, f)}{\partial f}\right|_{f = f^{[m-1]}(\mathbf{x}_i)}, \quad i = 1, \cdots, n.$$

2. Fit the negative gradient vector $U_1, \cdots, U_n$ to $\mathbf{x}_1, \cdots, \mathbf{x}_n$ by a base procedure (e.g., the weak learner in AdaBoost):

$$\{(\mathbf{x}_i, U_i), i = 1, \cdots, n\} \rightarrow g^{[m]}(\cdot).$$

3. Update the estimation by $f^{[m]}(\cdot) = f^{[m-1]}(\cdot) + \eta_m g^{[m]}(\cdot)$, where η_m is a step size factor.

4. Check the stopping criterion; if not satisfied, go to step 1.

Fig. 1. Generic functional gradient descent.