Statistical Methodology 22 (2015) 47–57

Contents lists available at ScienceDirect

Statistical Methodology

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

Incorporating auxiliary information for improved prediction using combination of kernel machines

Xiang Zhan [a,*], Debashis Ghosh [a,b]

[a] *Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA*

[b] *Department of Public Health Sciences, Pennsylvania State University, University Park, PA 16802, USA*

ARTICLE INFO

Article history:
Received 2 October 2013
Received in revised form
10 June 2014
Accepted 6 August 2014
Available online 26 August 2014

Keywords:
Auxiliary information
Combination of kernels
Hybrid predictor
Kernel ridge regression
Mean squared prediction error

ABSTRACT

With evolving genomic technologies, it is possible to get different measures of the same underlying biological phenomenon using different technologies. The goal of this paper is to build a prediction model for an outcome variable Y from covariates X. Besides X, we have surrogate covariates W which are related to X. We want to utilize the information in W to boost the prediction for Y using X. In this paper, we propose a kernel machine-based method to improve prediction of Y by X by incorporating auxiliary information W. By combining single kernel machines, we also propose a hybrid kernel machine predictor, which can yield a smaller prediction error than its constituents. The prediction error of our kernel machine predictors is evaluated using simulations. We also apply our method to a lung cancer dataset and an Alzheimer's disease dataset.

1. Introduction

Biomarkers in cancer research are considered to be central to prevention, detection and monitoring of the disease. With continual development of genomic technologies, one consequence is that different data with biomarkers measured by different technologies are available. As a motivating example, we consider data from a lung cancer study in Chen et al. [8]. One of the main scientific goals in Chen

[*] Corresponding author. Tel.: +1 8143213493.
E-mail addresses: xyz5074@psu.edu (X. Zhan), ghoshd@psu.edu (D. Ghosh).

http://dx.doi.org/10.1016/j.stamet.2014.08.001
1572-3127/© 2014 Elsevier B.V. All rights reserved.

et al. [8] focuses on predicting survival time in patients with lung cancer. Affymetrix gene expression data were obtained on 439 tumor samples. As a follow-up, a subset of 47 samples was measured again using a quantitative real-time polymerase chain reaction (qRT-PCR) platform. While both technologies measure gene expression, the Affymetrix data are regarded as a noisy version the qRT-PCR data since qRT-PCR technology is more generalizable and clinically applicable. The goal is to develop prognostic models for the survival outcome from qRT-PCR data. The question we consider in this paper is how the auxiliary information in the Affymetrix data can be used to improve the prediction of survival time given qRT-PCR data. Let Y denote the survival time, X denote qRT-PCR measurement of gene expression and W be Affymetrix measurement. Depending on whether a tumor sample is measured by qRT-PCR, samples are divided into two parts A and B. Subsample A consists of the 47 tumor samples which are measured by qRT-PCR in a follow-up study. And the remaining tumor samples form subsample B. They are denoted by (Y^A, X^A, W^A) and (Y^B, W^B) respectively. The goal in this paper is to use auxiliary information in W to boost the prediction of $Y|X$.

Boonstra et al. [2] first considered this non-standard prediction problem assuming the following models:

$$Y = \beta_0 + X^T\beta + \epsilon; \qquad W = \phi I_p + \nu X + \varepsilon, \tag{1}$$

where Y is a continuous response, X and W are p-dimension biomarker measurements, β is a $p \times 1$ vector, I_p is a $p \times p$ identity matrix, $\epsilon \sim N(0, \sigma^2)$, $\varepsilon \sim N_p(0, \tau I_p)$, β_0, ϕ, ν, σ and τ are scalars. They proposed a general class of targeted ridge (TR) estimators which include ridge regression (Hoerl and Kennard [11]) as a special case. Ridge regression estimator shrinks the least squares estimator toward zero. And TR estimators shrink the least squares estimator to certain targets derived from W^B and Y^B, which is how the auxiliary information in subsample B is used. More details can be found in Boonstra et al. [2]. Generally speaking, TR estimators are biased. It is possible that TR can have a better prediction performance by largely reducing variance to offset the introduced bias (Boonstra et al. [2]).

However, we observe two major disadvantages of TR. First, TR fails when the dimension of X is not equal to that of W. The formulas proposed in Section 2 of [2] do not work when X and W are of different dimensions. Second, the prediction performance of TR may not be good when the true underlying functional relationship in Eq. (1) is not linear. To address those two issues, we propose a kernel method based on kernel ridge regression (Cristianini and Shawe-Taylor [9]) to solve the aforementioned prediction problem. One general model consistent with the context is:

$$Y = f(X) + \epsilon; \qquad X = h(W) + \varepsilon, \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$ and $\varepsilon \sim N(0, \tau^2)$. Functions $f(\cdot)$ and $g(\cdot)$ are in some Hilbert functional spaces spanned by kernel functions. More details can be found in Section 2. If one takes W as an error-prone version of X, $X = h(W) + \varepsilon$ can be viewed as a weakly parametric measurement error model (Carroll et al. [6]). Kernel machine regression has been widely used in recent works (see [5, 14,16,15] for more details). It is flexible and allows for complicated relationships between response and predictor, which is desirable in practice. The fact that kernel machine makes few assumptions can give it an advantage in certain scenarios. For example, the TR class of estimators are inefficient when the linearity assumption is violated. The goal of this paper is to establish a prediction model for new observations X_{new}. The performance of the predictive model is typically measured by the mean squared prediction error (MSPE):

$$MSPE(\hat{f}) = E[Y_{new} - \hat{f}(X_{new})]^2 = \sigma^2 + E[f(X_{new}) - \hat{f}(X_{new})]^2. \tag{3}$$

The rest of the paper is organized as follows. The main result of this paper is presented in Section 2. We first review some useful facts about reproducing kernel Hilbert space (RKHS) and kernel ridge regression. Based on that, we propose a kernel machine predictor for the prediction problem considered in this paper. In the end, a hybrid kernel machine predictor is also proposed based on combination of single kernel machine predictors. In Section 3, we present a simulation study to compare our kernel method with the TR method proposed in [2]. A lung cancer dataset and a GRIN2B gene dataset