Journal of Biomedical Informatics 60 (2016) 431–445

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Object-oriented regression for building predictive models with high dimensional omics data from translational studies

Lue Ping Zhao [a,b,*], Hamid Bolouri [c]

[a] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, United States
[b] Department of Biostatistics and Epidemiology, University of Washington School of Public Health, Seattle, WA, United States
[c] Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, United States

ARTICLE INFO

Article history:
Received 5 October 2015
Revised 23 February 2016
Accepted 1 March 2016
Available online 10 March 2016

Keywords:
Big data
Clustering analysis
Gene expression
High dimensional data
LASSO
Lung cancer
Nearest neighbor approach
Penalized regression
Generalized linear model

ABSTRACT

Maturing omics technologies enable researchers to generate high dimension omics data (HDOD) routinely in translational clinical studies. In the field of oncology, The Cancer Genome Atlas (TCGA) provided funding support to researchers to generate different types of omics data on a common set of biospecimens with accompanying clinical data and has made the data available for the research community to mine. One important application, and the focus of this manuscript, is to build predictive models for prognostic outcomes based on HDOD. To complement prevailing regression-based approaches, we propose to use an object-oriented regression (OOR) methodology to identify exemplars specified by HDOD patterns and to assess their associations with prognostic outcome. Through computing patient's similarities to these exemplars, the OOR-based predictive model produces a risk estimate using a patient's HDOD. The primary advantages of OOR are twofold: reducing the penalty of high dimensionality and retaining the interpretability to clinical practitioners. To illustrate its utility, we apply OOR to gene expression data from non-small cell lung cancer patients in TCGA and build a predictive model for prognostic survivorship among stage I patients, i.e., we stratify these patients by their prognostic survival risks beyond histological classifications. Identification of these high-risk patients helps oncologists to develop effective treatment protocols and post-treatment disease management plans. Using the TCGA data, the total sample is divided into training and validation data sets. After building up a predictive model in the training set, we compute risk scores from the predictive model, and validate associations of risk scores with prognostic outcome in the validation data (P-value = 0.015).

1. Introduction

The advent of next generation sequencing technologies [1,2] enables clinical researchers to routinely process hundreds of biospecimen samples collected from patients, assessing, e.g., genome wide expression levels [3], methylation levels [4], or somatic mutations [5], referred to here as high dimensional omics data (HDOD). Despite the usually limited available sizes of clinical samples, the numbers of observed variables on each sample can be in the thousands or millions. The affordability of these technologies has moved the bottleneck of clinical research from sample acquisitions to data management and data analytics. While there are numerous analytic objectives contemplated by biomedical informatics researchers, one of them, the focus of this manuscript, is to build predictive models for specific clinical outcomes, utilizing HDOD along with other clinical variables.

Building predictive models has been a long-standing research interest shared by quantitative researchers in several disciplines. Computer scientists have been actively developing predictive models with large data sets from databases [6,7]. Methods include support vector machines [8], genetic algorithms [9], and many other machine learning algorithms [10,11]. Additionally, taking full advantage of their intimate familiarity with database technologies and visualization tools, computer scientists have been effective in organizing HDOD, scaling up computing power to analyze HDOD, and presenting HDOD-derived results visually so that biomedical researchers can interact with HDOD and can intuitively comprehend results. Recent successes with these applications in biomedical research partially contribute to the growth of bioinformatics.

Building predictive models has been a long-standing interest for statisticians. A literature review is not attempted here. It suffices to note several major milestones in this area. Given the nature of

* Corresponding author at: Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109, United States.
E-mail address: lzhao@fredhutch.org (L.P. Zhao).

http://dx.doi.org/10.1016/j.jbi.2016.03.001
1532-0464/

predicting an outcome with multiple variables, regression-based predictive models are commonly built, and most are special cases within generalized linear models (GLM) [12]. Relaxing the parametric assumption, Hastie and Tibshirani described a generalized additive model (GAM), synthesizing results from decades of research on nonparametric regression methods [13]. In recent years, statisticians have been developing penalized likelihood techniques to automate the covariate selections from HDOD [14], including LASSO [15,16], GBM [17], Elastic-Net [18], Ridge regression [19] and Radom Forests [20]. These methods are commonly used tools for analyzing HDOD in translational research.

While there is some crossbreeding of methods between computer sciences and statistics, one fundamental difference in our opinion is that computer scientists often explore patterns with multiple variables from a systemic perspective, while statisticians tend to identify a few covariates following the parsimony principle. A major challenge facing statisticians is how to control the overly inflated false positive error rate in selecting predictors from HDOD, so that discoveries are reproducible in independent samples. In contrast, computer scientists or bioinformaticians, with primary interest in patterns of HDOD, often desire to quantify observed patterns in a robust manner, in hope that discovered patterns are reproducible on independent data sets.

To frame the "big picture", consider what would be a clinician's intuition in dealing with complex medical information. Clinicians typically gather multifaceted information from medical records, from physical examinations, and from diagnostic laboratory tests, a version of HDOD, and then make a clinical judgement based on the evidence plus their experiences of past cases. Mentally, an experienced clinician would compare the new patient with previously treated patients or those typical cases in textbooks or in literature, and would reduce the mental comparison to an intuitive clinical judgement with a sample size of one. In essence, the clinician's assessment is holistic by comparing individual's HDOD with those HDOD profiles of known subjects, like exemplars.

Being motivated by this clinician's intuition, we propose a hybrid approach of integrating data pattern discovery and regression analytics, to retain desired features of both analytic approaches. This approach has two steps. At the first step, the goal is to identify a group of "exemplars" that are representative of subjects' HDOD patterns, typically observed through clustering analysis of unsupervised learning [14,21,22]. To have cluster patterns represented, one could choose centroids of clusters as exemplars. To represent those samples under-represented by clusters, one could choose singletons to be exemplars. In essence, a HDOD pattern characterizes an exemplar. The number of exemplars (q) is generally smaller than the sample size (n), unless exemplars are derived externally (see discussion below). With reference to each exemplar, one can compute a similarity measurement with each subject, resulting in a matrix of similarity measurements with the dimension ($n \times q$). Typically, $p \gg n > q$. Effectively, this step transforms high dimension and sparse HDOD ($n \times p$) into a "dense data matrix" ($n \times q$). Then, at the second step, we use penalized likelihood methods to select those exemplars that are predictive of the outcome. Because of the substantially reduced dimensionality from p to q, the penalized likelihood can readily picks up informative exemplars, at much reduced penalty. The dual step procedure relies on exemplars from "unsupervised learning" and then selects informative exemplars with their associations with outcome via "supervised learning". Because of regressing outcome on exemplar-specific similarities, this method is referred to as "object-oriented regression" or OOR for short. In contrast, most of regression-based methods mentioned above are known as covariate-specific regression methods (CSR).

2. Methodology

2.1. Motivation

The Statement of Problem: Consider a sample of n subjects ($i = 1, 2, \ldots, n$) in a clinical follow-up database. On each ith subject, we observe a set of high dimensional and sparse covariates, denoted as $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, where the number of covariates is typically much greater than the sample size ($p \gg n$), typical of HDOD. Also observed on each ith subject is time-to-event outcome variable of interest $Y_i = (\delta_i, t_i)$, in which binary indicator δ_i is for, e.g., alive or death, at the observed time t_i. The likelihood of all observed data may be written as

$$L(Y_i, X_i, \forall i) = -\sum_i \log f(Y_i|X_i) - \sum_i \log f(X_i), \tag{1}$$

where the summation is over n subjects, $f(Y_i|X_i)$ the conditional density of Y_i given covariates X_i, and $f(X_i)$ is the multivariate distribution of covariates [23]. To capture association of the time-to-event outcome with covariates, it is a common practice to model a hazard function [24], which may be written as

$$\lambda(t|X_i, \theta) = \lambda_0(t) \exp[h(X_i, \theta)], \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard function independent of covariates, and $h(X_i, \theta)$ is an arbitrary function indexed by a vector of unknown parameters θ to be estimated from a data set. Correspondingly, the distribution function $f(Y_i|X_i)$ is specified by the hazard function via

$$f(Y_i|X_i) = [\lambda(t_i|X_i, \theta)]^{\delta_i} \exp\left[-\int_0^{t_i} \lambda(u|X_i, \theta)du\right]. \tag{3}$$

The analytic objective is to establish the outcome (Y_i) association with covariates (X_i) via modeling the arbitrary function $h(X_i, \theta)$.

The Representer Theorem: When the covariate function is unknown and is left unspecified, Kimeldorf and Wahba [25] have shown that given the observed samples ($X_1, X_2, \ldots, X_n$), the above arbitrary function $h(X_i, \theta)$ in Eq. (2) can be generally represented by

$$h(X, \theta) = \sum_{k=1}^{n} \theta_k K(X, X_k), \tag{4}$$

where θ_k is a sample-specific and unknown parameter, and $K(X, X_i)$ is known as the kernel function and needs to be semi-positive definite [25]. One class of kernel function is the similarity measure that quantifies the similarity of X with X_k. For an observation X identical to X_k, the corresponding term is $\theta_k K(X, X_k) = \theta_k$. If X is completely different from X_k, $\theta_k K(X, X_k) = 0$. Further, if X_k and $X_{k'}$ are identical or nearly identical, corresponding terms can be merged as $\theta_k K(X, X_k) + \theta_{k'} K(X, X_{k'}) \approx (\theta_k + \theta_{k'}) K(X, X_k) = \alpha_k K(X, X_k)$. Lastly, one expects that the coefficient θ_k, quantifying outcome association with similarity measure $K(X, X_k)$ with the kth individual, is likely to equal zero, if the covariate profile of the kth individual is not associated with the corresponding outcome. Zhu and Hastie used some of these observations to describe an import vector machine approach by grouping some $K(X, X_k)$ terms [26].

The Representer theorem, together with above observations, forms the theoretical foundation for us to propose OOR by modeling this arbitrary function via

$$h(X_i, \alpha, \beta's) = \alpha + \sum_{k=1}^{q} \beta_k s_k(X_i), \tag{5}$$

where $s_k(X_i) = K(X_i, Z_k)$ is the similarity measurement of X with the kth unique HDOD Z_k, and (α, β_k) are unknown regression coefficients to be estimated. Formally, HDOD vector Z_k represents a pattern of HDOD or HDOD profile, and is referred to as an exemplar.