Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry 29 (2005) 204–211
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Borrowing information from relevant microarray studies for sample classification using weighted partial least squares

Xiaohong Huang[a,1], Wei Pan[a,*], Xinqiang Han[b], Yingjie Chen[b], Leslie W. Miller[b], Jennifer Hall[b]

[a] *Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building (MMC 303), Minneapolis, MN 55455-0378, USA*
[b] *Cardiovascular Division, Department of Medicine, Medical School, University of Minnesota, USA*

Received 24 December 2004; accepted 5 April 2005

Abstract

With an increasing number of publicly available microarray datasets, it becomes attractive to borrow information from other relevant studies to have more reliable and powerful analysis of a given dataset. We do not assume that subjects in the current study and other relevant studies are drawn from the same population as assumed by meta-analysis. In particular, the set of parameters in the current study may be different from that of the other studies. We consider sample classification based on gene expression profiles in this context. We propose two new methods, a weighted partial least squares (WPLS) method and a weighted penalized partial least squares (WPPLS) method, to build a classifier by a combined use of multiple datasets. The methods can weight the individual datasets depending on their relevance to the current study. A more standard approach is first to build a classifier using each of the individual datasets, then to combine the outputs of the multiple classifiers using a weighted voting. Using two quite different datasets on human heart failure, we show first that WPLS/WPPLS, by borrowing information from the other dataset, can improve the performance of PLS/PPLS built on only a single dataset. Second, WPLS/WPPLS performs better than the standard approach of combining multiple classifiers. Third, WPPLS can improve over WPLS, just as PPLS does over PLS for a single dataset.

Keywords: Meta-analysis; Partial least squares; Penalized partial least squares; Gradient directed path; Squared error loss

1. Introduction

DNA microarray technologies allow the measurement of expression levels of thousands of genes simultaneously. Microarray experiments are more and more widely used in classification of tumor samples, prediction of clinical outcomes, or detecting differential gene expressions. With a rapidly increasing number of publicly available microarray datasets addressing various biological questions for various organisms, there is potential to gain more information by a combined analysis of multiple studies. For example, it has become popular to take a meta-analysis approach to combining data from multiple studies to detect differential gene expression (Rhodes et al., 2002; Xin et al., 2003; Choi et al., 2004; Ghosh et al., 2003; Wang et al., 2004a) or for sample classification (Jiang et al., 2004; Parmigiani et al., 2004; Shen et al., 2004). A technical issue is how to combine microarray data measured using different microarray techniques or platforms, such as cDNA versus Affymetrix arrays, or different versions of Affymetrix arrays, because of possibly different gene identities and possibly incomparable expression measurements across different platforms (e.g. Morris et al., 2003; Robb et al., 2003; Hu et al., 2003; Lin et al., 2003).

Here we consider a related but different problem. Our goal is to analyze a given dataset drawn from a current study. To increase the statistical power, we would like to borrow information from other relevant studies. A key difference from meta-analysis is that we do not assume that the current study shares a common set of parameters with other studies. For

* Corresponding author. Tel.: +1 612 626 2705; fax: +1 612 626 0660.
[1] Present address: DC 6166, Zyprexa/Symbyax Product Team, Eli Lilly and Company, Indianapolis, IN 46285, USA.
E-mail address: weip@biostat.umn.edu (W. Pan).

1476-9271/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.04.002

example, we might be interested in identifying genes associated with ventilator-associated lung injury (VALI) based on a human study. On the other hand, there are studies on the same subject using animal models, such as rat, mouse and dog (Grigoryev et al., 2004). We would like to borrow information from these animal studies to address the scientific question, identifying the human genes associated with VALI. This is different from a meta-analysis (Grigoryev et al., 2004) with its goal to identify the genes associated with VALI that are conserved across the species over the evolutionary history, perhaps only a subset of the human genes of interest. In other words, meta-analysis has to assume that we have a common set of parameters across the human and animal studies, on which a statistical inference is to be drawn. In our analysis, we would not assume such a common set of parameters; rather, we are only interested in inference on a set of parameters specific for humans. Although the human-specific set of parameters is in general different from that of animal models, it is reasonable to assume a priori that they are likely to be close. Hence, we may borrow information from the animal studies to improve the estimation on the parameters for humans; the animal studies will be called secondary or auxiliary as compared to the primary human study. The statistical motivation of our proposal is similar to that of the weighted likelihood theory, which seeks to reduce the variance of an estimator (with a possible price of increasing its bias) and thus to result in a smaller mean squared error or prediction error (Newton and Raftery, 1994; Rao, 1991; Hu and Zidek, 2002; Wang et al., 2004b; Ghosh et al., 2004, and references therein). As a concept-of-proof, we consider two studies on human heart failure, the LVAD study and the PGA study; more details on the studies are presented later. Our goal is to use gene expression profiles to distinguish etiologies of heart failure for LVAD patients, and we treat the PGA data as secondary or auxiliary. To be specific, we consider comparing ischemic (IS) group with idiopathic (ID) group. There are 10/13 IS/ID samples and 11/13 IS/ID samples in the LVAD and PGA data, respectively. Intuitively, due to the small sample size and the relevance of the two studies, we would like to borrow information from the PGA data to build a model for LVAD patients. Although both the LVAD and PGA studies are on humans, they own some features shared by other more typical and less relevant studies for which and from which we would like to borrow information: due to the population heterogeneity (i.e. unobserved differences in patient characteristics), different study protocols and different microarray platforms, the data from the two studies are quite different. In particular, it is much harder to distinguish IS/ID patients using gene expression profiles in the LVAD study than in the PGA study. For example, using the penalized partial least squares (PPLS) method with varying numbers of genes and of components in a starting partial least squares (PLS) model (Huang and Pan, 2003), (i) the leave-one-out-cross-validation (LOOCV) misclassification errors range from 5 to 11 for the LVAD data; (ii) the LOOCV errors range only from 1 to 3 for the PGA data; (iii) the minimum test error on the LVAD data with the PPLS model built using the PGA data is 8; see Huang et al. (2004b), for a more detailed analysis. These results highlight some existing differences underlying the two datasets; in particular, appropriate models for the LVAD data and the PGA data may be different. Nevertheless, it is desirable to take account possible differences between the two datasets while borrowing information from the PGA to build a better classifier for the LVAD study.

We propose a weighted partial least squares (WPLS) method and a weighted penalized partial least squares (WPPLS) method, which account for possibly different relevances of the studies by assigning them possibly different weights. PLS method is considered especially useful for constructing linear models when there are many covariates and a relatively small sample size, as is typical with microarray data. There has been an increasing application of PLS/PPLS to microarray data (e.g. among others, Nguyen and Rocke, 2002; Hawkins et al., 2003; Huang and Pan, 2003; Huang et al., 2004; Tan et al., 2004; Boulesteix, 2004). WPLS is an extension of the standard PLS method by giving samples different weights (based on their relevance to the current study); it is facilitated by formulating PLS solutions as a gradient directed path in minimizing a loss function. The WPPLS method, which penalizes or regularizes the coefficients of a WPLS model, aims to facilitate model interpretation and further reduce noise effects of microarray data on the model and thus to improve the performance over the WPLS method.

2. Methods

Suppose that x_{ij} is the expression level of gene i in sample (array) j, and the random variable y_j is the response variable for sample j, where $i = 1, \ldots, p$ and $j = 1, \ldots, n$. In the current context, $y_j = 1$ or -1, representing one of the two types of the sample. Denote column vectors $x_i = (x_{i1}, \ldots, x_{in})^{\mathrm{T}}$ and $y = (y_1, \ldots, y_n)^{\mathrm{T}}$. Given the data, the goal is to construct a linear model for the sample type Y; that is, in the linear model

$$F(X, \mathbf{a}) = a_0 + \sum_{i=1}^{p} a_i x_{ij}$$

we are to estimate the parameters $\mathbf{a} = (a_0, a_1, \ldots, a_p)$. Ideally, we would like to minimize the expected loss/risk

$$R(\mathbf{a}) = E_Y L(Y, F(X, \mathbf{a})),$$

where $L(Y, F(X, \mathbf{a}))$ is the loss of predicting the response variable Y by its predicted value of $F(X, \mathbf{a})$. The optimal values of $\mathbf{a}$ are those that minimize the expected loss. Since in practice the distribution of Y is unknown, we estimate the expected loss by the empirical loss based on the observed