Journal of Biomedical Informatics 46 (2013) 594–601

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Selecting significant genes by randomization test for cancer classification using gene expression data

Zhiyi Mao, Wensheng Cai, Xueguang Shao *

State Key Laboratory of Medicinal Chemical Biology, and Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin 300071, China

ARTICLE INFO

Article history:
Received 3 July 2012
Accepted 28 March 2013
Available online 6 April 2013

Keywords:
Gene expression data
Randomization test
Partial least squares discriminant analysis
Gene selection
Cancer classification

ABSTRACT

Gene selection is an important task in bioinformatics studies, because the accuracy of cancer classification generally depends upon the genes that have biological relevance to the classifying problems. In this work, randomization test (RT) is used as a gene selection method for dealing with gene expression data. In the method, a statistic derived from the statistics of the regression coefficients in a series of partial least squares discriminant analysis (PLSDA) models is used to evaluate the significance of the genes. Informative genes are selected for classifying the four gene expression datasets of prostate cancer, lung cancer, leukemia and non-small cell lung cancer (NSCLC) and the rationality of the results is validated by multiple linear regression (MLR) modeling and principal component analysis (PCA). With the selected genes, satisfactory results can be obtained.

1. Introduction

Cancer classification based on microarray has become a popular research topic in bioinformatics, which can be used to detect subtypes of cancers and produce therapies. A great many of studies have appeared for cancer classification [1–3]. These methods include principal component analysis (PCA) [4,5], *k*-nearest neighbor (*k*-NN) [6], hierarchical clustering analysis (HCA) [7], support vector machine (SVM) [8], Bayesian method [9], partial least squares discriminant analysis (PLSDA) [10], ensemble methods [11], etc. Among these methods, PLSDA has been the most commonly used one for cancer classification due to its simplicity [12–14]. Moreover, as a dimension reduction technique, PLS has been used in gene expression data analysis even in the case where the number of genes exceeds the number of samples.

Except for a few classification methods using full genes [15], classification is generally performed based on selecting significant genes for constructing accurate prediction models. Furthermore, gene selection may provide insights into understanding the underlying mechanism of a specific biological phenomenon. Also, such information can be useful for designing less expensive experiments by targeting only a handful of genes [16]. However, how to effectively select significant biomarker genes from thousands or even ten thousands of genes is a difficult problem. A comprehensive review of feature selection methods has been described by Saeys et al. [17]. Depending on how the genes interact with the construction of the classification model, feature selection techniques can be characterized into three classes: filter, wrapper and embedded methods. Filter methods [18] assess the relevance of features by looking only at the intrinsic properties of the data, and thus they are computationally simple and fast. Wrapper methods [19] employ a selection strategy in the space of all possible feature subsets, guided by the predictive performance of a classification model. Advantage of these methods includes the interaction between gene subset search and model selection. However, they may have a higher risk of over-fitting than filter methods and may be computationally intensive. Embedded methods [20] make use of the internal parameters in a classification model to perform feature selection, and, therefore, the computational cost is reduced but the advantage of the interaction between the gene selection and classification model is preserved. Based on the three classes of feature selection techniques, various gene selection algorithms have been proposed and successfully used in selecting informative genes for cancer classification [21–25]. In our previous works, Monte Carlo based uninformative variable elimination (MC-UVE) [26], randomization test (RT) [27], PLS with the influential variables (IVs) [28] and latent projective graph (LPG) [29] have been proposed for selecting informative variables in near-infrared spectral analysis. Among these methods, RT has been proved to be an efficient approach to extract useful information from the spectra. The method builds a regular model and a series of random models, and then evaluates the importance of the variables based on the significance test of coefficients in regular and random models. The variables with high significance can be selected as the informative ones.

In this study, RT coupled with PLSDA was employed to seek the significant genes for cancer classification. A set of PLSDA models

* Corresponding author. Fax: +86 22 23502458.
E-mail address: xshao@nankai.edu.cn (X. Shao).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.03.009

are built by permutation, and the significance of genes are evaluated by a statistic. To validate the performance and applicability of the method, four gene expression datasets were investigated. The results show that the method can select significant genes for cancer classification.

2. Methods

2.1. Partial least squares discriminant analysis (PLSDA)

Partial least squares (PLS) regression is a well-known method to find the relationship between predictor variables **X** and dependent variables **y**. In a PLS model, not only the variance of **X**, but also the covariance between **X** and **y** is taken into account. Therefore, the central point of PLS is to find latent variables in the feature space that have a maximum covariance with **y**. PLSDA is a variant of PLS to improve the separation between classes using a categorical response variable **y**. In this study, **X** is the matrix of gene expression values and the values of **y** are given as 1 and -1 for positive and negative class, respectively. Each row of **X** matrix represents the gene expression values of all the genes for each sample, and each column corresponds to the gene expression values of all samples for a gene. PLSDA is used for modeling the genes expression data (**X**) and the response variable (**y**) using the training set. In the calculations, the optimal latent variable (LV) number used in the modeling is determined by Monte Carlo cross validation (MCCV). In the prediction, the samples with predicted values above zero are ascribed to positive class, otherwise to negative class. The parameters of accuracy (Acc), precision (P), recall (R) and *F*-measure (*F*) are used to evaluate the classification effect.

2.2. Randomization test (RT)

RT is a method for variable selection by employing the statistics of the regression coefficients in the models built with permutation of the dependent variables **y** in the training set [27]. In the calculation of RT, a regular model showing the relationship of **y** and **X** is built for reference and a number (M) of random PLSDA models are built by randomization, i.e., randomly scrambling the indices of **y** while keeping the indices of **X**. In this study, the number of the permutations is 1000, as discussed in our previous work [27]. In each of the random models, a regression coefficient can be obtained for each gene. Clearly, the regression coefficients of each gene in the random models must be due to chance. Therefore, the values of the regression coefficients can be referred to as 'noise values'.

A statistic, P, is defined as the fraction of the 'noise values' exceeding the regression coefficient in the regular PLSDA model,

$$P_j = num(|\boldsymbol{\beta}_j| > |\beta_{0j}|)/M \quad (j = 1, 2, \ldots, p) \tag{1}$$

where j is the index of the genes, and p is the number of genes. $\boldsymbol{\beta}_j$ and β_{0j} represent the 'noise values' and the regression coefficient in the regular model of the gene j, and M is the number of random models. Since the value of the regression coefficient for each gene is a reflection of its importance in the model, the informative or relevant genes generally have coefficients of large absolute values. Therefore, the 'noise values' should be significantly smaller than the coefficients of the normal model, because they are obtained by randomization, and the significance of a gene can be assessed by its P value. If a cutoff value is defined, the genes whose coefficients are smaller than the threshold should be selected as informative ones. In this study, all the genes are ranked by their P values, and thus the genes are selected according to the order from low to high P values.

2.3. RT-PLSDA method

RT-PLSDA means a combination of RT and PLSDA, in which the coefficients of PLSDA models were used to calculate P values. Four steps are included in the calculations. Among the steps, the first two steps are used for selecting the informative genes according to the P values. The third step determines the retained genes by repetition of RT procedures to make the result more reliable, and the fourth step involves the modeling and prediction with the selected genes. The calculation details can be described as follows.

(1) With the training set, a regular PLSDA model is built, and the regression coefficients for the genes are recorded in a $1 \times p$ vector β_0. With the same training set, M permutations of **y** are performed to build M PLSDA random models. The regression coefficients are recorded in an $M \times p$ matrix $\boldsymbol{\beta}$ as the 'noise values'. It should be noted that before the calculation, auto-scaling were performed to the datasets in order to eliminate the effect of intensity difference between genes and make each gene have a comparable contribution to the classification.
(2) P value of each gene is calculated by using Eq. (1), and the genes are ranked in an ascending order of P values. With a number (N) of genes with lower P values, the error of cross-validation (ECV), which is defined as the number of misclassified samples, is obtained by MCCV. In the calculation of MCCV, 50% of the samples in the training set are randomly selected to build the model and predict the remaining samples, and 1000 repetitions were performed. The ECV value is calculated by the sum of misclassified samples number in the 1000 repetitions. The number of genes with the minimum ECV value is selected.
(3) Because random permutation is involved in the calculations, the distribution of P values is not identical in different runs. A large number of runs may not be necessary considering the time consumed, 100 was used for ensuring the reliability and for investigating the repeatability of the method. A frequency number in the 100 runs is used to further describe the significance of each gene. The selected genes are ranked in a descending order with the frequency number, and with different number of the selected genes, a series of PLSDA models are built and the ECV is obtained by MCCV. The optimal number of retained genes can be therefore determined by the lowest ECV value for the training set as calculated in step (2).
(4) With the retained genes, a multiple linear regression (MLR) model for classification is built and used to predict the test set.

In RT method, the distribution of P values is plotted for determination of the variables with low value. In this study, $-\lg P$ is used in place of P to make the distribution more clear. In this case, the genes with higher values will be more significant. It should be noted that for few genes, P value may be zero when the regression coefficient in the regular model is larger than all the 'noise values'. Such genes are obviously significant ones. For these genes, $-\lg P$ is defined as 4 because the maximum value of $-\lg P$ is 3 when only one of the 'noise values' is larger than the regression coefficient in the regular model.

3. Datasets

Four gene expression datasets of prostate cancer [30], lung cancer [31], leukemia [32] and non-small cell lung cancer (NSCLC) [33] were used in this study. A summary of the four datasets are listed