Applied Mathematics and Computation 238 (2014) 132–140

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Variable selection for Fisher linear discriminant analysis using the modified sequential backward selection algorithm for the microarray data

Hong-Yi Peng [a], Chun-Fu Jiang [b,*], Xiang Fang [c], Jin-Shan Liu [a]

[a] College of Science, South China Agricultural University, Guangzhou 510642, PR China
[b] College of Mathematics and Computational Science, Shenzhen University, Shenzhen 518060, PR China
[c] College of Food Science, South China Agricultural University, Guangzhou 510642, PR China

ARTICLE INFO

Keywords:
Variable selection
Fisher linear discriminant analysis
Modified SBS
Weighted Mahalanobis distance
Microarray data

ABSTRACT

One of the major challenges is small sample size as compared to large features number for microarray data. Variable selection is an important step for improving diagnostics of cancer or the classification according to the phenotypes via gene expression data. In this study, we propose a modified sequential backward selection (SBS) algorithm to deal with the case where the covariance matrix is singular. Then we propose a variable selection algorithm based on the weighted Mahalanobis distance and modified SBS methods. Furthermore, based on the proposed variable selection algorithm, a Fisher linear discriminant method is proposed to improve the accuracy of tumor classification through simultaneously taking into account genes' joint discriminatory power. To validate the efficiency, we apply the proposed discriminant method to two different DNA microarray data sets for experiment investigation. The empirical results show that our method for tumor classification can obtain better classification effectiveness than Markov random field method and independent variable group analysis I methods, which demonstrates that the proposed variable selection method can obtain more correct and informative gene subset if taking into account the joint discriminatory power of genes for tumor classification.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

The development of microarray technology increases the possibility of cancer classification and diagnosis at the gene expression level, see the literatures [1,2]. Thus, classification and clustering can be used to analyze and interpret the gene expression data, e.g., Nanni et al. [3], Zheng et al. [4] and Yeung and Ruzzo [5]. The analysis of gene expression data is motivated by the problem of distinguishing between cancer classes or identifying and discovering various subclasses of cancers. However, many factors may affect the outcome of the analysis. One of the major challenges is the so-called "the curse of dimensionality" mainly due to small sample size as compared to large number of features. Jain et al. [6] point out that there are too many features, which may be irrelevant to analysis actually, degrade the generalizable performance of a classifier. Thus, selecting discriminatory genes (i.e. variables) is critical to improve the accuracy and speed of prediction systems.

* Corresponding author.
E-mail addresses: phyzsu@scau.edu.cn (H.-Y. Peng), jiangcf@szu.edu.cn (C.-F. Jiang), fxiang@scau.edu.cn (X. Fang), liujs58@scau.edu.cn (J.-S. Liu).

http://dx.doi.org/10.1016/j.amc.2014.03.141
0096-3003/© 2014 Elsevier Inc. All rights reserved.

There are a vast amount of literatures reported, which focused on how to use gene selection method for tumor classification, see Golub et al. [7], Dudiot et al. [8] for example, and Li et al. [9], Bae and Mallick [10], Lee et al. [11], Li et al. [12] among others. Golub et al. [7] used signal-to-noise ratio (SNR) to select informative genes for the two-class prediction problem of distinguishing acute lymphoblastic leukemia (ALL) from acute myeloid leukemia (AML). Dudoit et al. [8] extend the two-class SNR method for multiple classes using the ratio of their between-group to within-group sums of squares (BWR). Some gene selection techniques are based on classical statistical methods, such as Bayesian variable selection in Bae and Mallick [10], logistic regression in Li et al. [12], Shevade and Keerthi [14], and analysis of variance in Draghici et al. [13]. Although being useful in practice, all these methods fail to take into account their joint discriminatory power because they select important genes based on individual gene discriminatory power.

Independent variable group analysis (IVGA, hereafter, see [15]) provides a principle for grouping variables that are mutually dependent together so that independent or only weakly dependent variables are placed to different groups. Alhoniemi et al. [16] argued that IVGA can also be used as a dimensionality reduction or feature selection method. Zheng et al. [17] proposed a new method called IVGA_I for gene selection based on IVGA, in which IVGA was firstly used to group the genes, i.e. clustering the genes using IVGA algorithm, then selecting the gene with the best discriminatory power for classification from each group. The number of clusters is equal to the number of key genes selected since the IVGA_I method selects one gene from each cluster. In general, IVGA_I method has two disadvantages. Firstly, the number of key genes was somehow chosen empirically. Secondly, since the key genes selected often have certain relativity, thus IVGA_I method doesn't succeed in taking into account genes' joint discriminatory power.

Recently, Stingo and Vannucci [18] captured the gene-gene network information via a Markov random field (MRF). They performed posterior inference by concentrating on the posterior distribution and implement a stochastic search variable selection (SSVS) algorithm used in the variable selection literatures, see Madigan and York [19] for graphical models, Brown et al. [20] for linear regression models, Sha et al. [21] for probit models and Tadesse et al. [22] for clustering, among others. Thus the variable selection method with the MRF does not succeed in taking into account genes' joint discriminatory power.

Feature selection has been widely used to alleviate the curse of dimensionality. Optimal selection methods such as exhaustive search or the Branch-and-Bound method in Narendra and Fukunaga [23] are not practical for very high-dimensional problems, for example, the case of including gene expression profiling studies. Therefore, alternative suboptimal methods, such as sequential search methods known as sequential backward selection (SBS) in Marill and Green [24] and its counterpart sequential forward selection (SFS) in Whitney [25], are often considered. An effectively known suboptimal methods are the sequential floating search methods, see Pudil et al. [26]. But the SBS method costs much less computation expense than the sequential floating search method. Compared with the SFS method, the SBF is not easy to leave out significance variables. If the Mahalanobis distance of two vectors is used as a criterion of feature set effectiveness, the SBS methods cannot deal with the case of singular covariance matrix. However, when the number of samples is less than the number of genes as in many gene expression profiling studies, the matrix is likely to be singular, hence resulting in a numerical problem in calculating the inverse matrix. Therefore, we proposed a modified SBS algorithm to deal with such case in this paper. Furthermore, in order to overcome the two defects of the IVGA_I and MRF methods, we also propose a new method called VSBW to select variable for tumor classification based on the modified SBS and weighted Mahalanobis distance (WMD) as the effectiveness criterion of feature set.

The remainder of the paper is organized as follows. Section 2 presents the Fisher linear discriminant analysis method. Section 3 presents the modified SBS method. Section 4 proposes a new method called VSBW to select the variable subset for tumor classification. Section 5, an experiment is carried out to demonstrate the proposed VSBW method. Finally, the conclusion is provided in Section 6.

2. Fisher linear discriminant analysis method

2.1. Preliminaries

The gene expression data on p genes for n mRNA samples can be summarized by an matrix $X = (x_{ij})_{n\times p}$, where x_{ij} denotes the expression level of the jth gene (variable) in the mRNA sample i, namely the ith observation. The expression levels might be either absolute or relative with respect to the expression levels of a common reference sample defined suitably. When the mRNA samples belong to known classes, the data for each observation consists of a gene expression profile $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and a class label y_i, i.e. of predictor variables $\boldsymbol{x}_i$ and response y_i. If we have K classes, the class labels y_i are defined to be integers ranging from 1 to K. Let n_k be the number of observations belonging to class k.

A predictor or classifier for K tumor classes partitions the space $\mathcal{X}$ of gene expression profiles into K disjoint subsets $A_1, A_2, \ldots, A_K$, such that for a sample with expression profile $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, the predicted class of $\boldsymbol{x}$ is $y = k$ if $\boldsymbol{x} \in A_k$. Predictors are built from past experience, i.e. from observations which are known to belong to certain classes. Such observations comprise the learning set denoted by $\mathcal{L} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$, in which each sample is known to belong to certain classes. For each observation $\boldsymbol{x}$ in the testing set, in which each sample in unknown to belong to certain classes, one need predict its class y. In the event that the observation $\boldsymbol{x}$ is predicted, namely y is known, then the estimation error rate of predictor can be obtained by comparing the predicted classes and the true classes. Below the classifier obtained from a learning set $\mathcal{L}$ is denoted by $C(\cdot, \mathcal{L})$. The predicted class for an observation $\boldsymbol{x}$ is accordingly denoted by $C(\boldsymbol{x}, \mathcal{L})$.

Download English Version:

https://daneshyari.com/en/article/4627944

Download Persian Version:

https://daneshyari.com/article/4627944

Daneshyari.com