Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

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

Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data

Yan Cui [a], Chun-Hou Zheng [b], Jian Yang [a,*], Wen Sha [b]

[a] School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, Jiangsu, China
[b] College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui, China

ARTICLE INFO

Article history:
Received 20 December 2011
Accepted 26 April 2013

Keywords:
Sparse maximum margin
Feature extraction
Gene selection
Gene expression data

ABSTRACT

Dimensionality reduction is necessary for gene expression data classification. In this paper, we propose a new method for reducing the dimensionality of gene expression data. First, based on a sparse representation, we developed a new criterion for characterizing the margin, which is called sparse maximum margin discriminant analysis (SMMDA); this approach can be used to find an optimal transform matrix such that the sparse margin is maximal in the transformed space. Second, using SMMDA, we present a new feature extraction method for gene expression data. Third, based on SMMDA, we propose a new discriminant gene selection method. During gene selection, we first found the one-dimensional projection of the gene expression data in the most separable direction using SMMDA. Then, we applied the sparse representation technique to regress the projection, and we obtained the relevance vector for the gene set. Discriminant genes were then selected according to this vector. Compared with the conventional method of maximum margin discriminant analysis, the proposed SMMDA method successfully avoids the difficulty of parameter selection. Extensive experiments using publicly available gene expression datasets showed that SMMDA is efficient for feature extraction and gene selection.

1. Introduction

The rapid development of microarray technologies, which can simultaneously assess the expression level of thousands of genes, makes the precise, objective, and systematic analysis and diagnosis of human cancers possible. By monitoring the expression levels of thousands of genes in cells simultaneously, microarray experiments could lead to a complete observation of the molecular variations among tumors and hence result in a reliable classification.

Gene expression data from DNA microarrays can be characterized by many variables (genes), but with only a few observations (experiments). Mathematically, the data can be expressed as a matrix $X=(x_{ij})_{m\times n}$, where each row represents a gene and each column represents a sample or a patient for tumor diagnosis. The numerical value of x_{ij} denotes the expression level of a specific gene $i(i=1,2,\ldots,m)$ in a specific sample $j(j=1,2,\ldots,n)$. Statistically, the fact that there is a very large number of variables (genes) with only a small number of observations (samples) makes most of the classical data analysis methods infeasible, including classification. Fortunately, this problem can be avoided by selecting only the relevant features or extracting the essential features from the original data, where the former methodology belongs to feature selection or subset selection and the latter is called feature extraction. For feature selection, *i.e.*, selecting a subset of genes from the original data, many related studies on tumor classification have been reported [1–4,11,12,21–23]. Moreover, a comparative study of discrimination methods based on selected sets of genes can be found in the literature [5].

Feature extraction is another type of widely-used method for tumor classification. Instead of selecting key genes from expression data, feature extraction methods aim to extract the most representative features with low dimensions from the original data. The popular feature extraction methods involved in gene data analysis include principal component analysis (PCA), linear discriminant analysis (LDA), complete PCA plus LDA [13], and partial least squares (PLS) [6]. A systematic benchmarking of these methods is reported in the literature [7]. These methods have good performance on tumor classification; however, they do not work well for non-Gaussian data sets [8]. To overcome this problem, Fukunaga and Mantock [9] presented a method called nonparametric discriminant analysis (NDA). This method is a classic margin-based discriminant analysis. In recent years, many other nonparametric discriminant analysis methods have been developed, such as nonparametric feature analysis (NFA) [8] and maximum margin criterion (MMC) [10].

Maximum margin criterion for robust feature extraction can avoid the small sample size (3s) problem, *i.e.*, the size of the samples is very small compared with the dimension of the

[*] Corresponding author. Tel.: +86 25 8431 7297.
E-mail address: csjyang@mail.njust.edu.cn (J. Yang).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.04.018

samples. In a geometrical sense, MMC projects input patterns onto the subspaces spanned by the normals of a set of pairwise orthogonal margin maximizing hyperplanes. It aims to maximize the average marginal distances between different classes, and the corresponding features can enhance the separability better than PCA and LDA (Fig. 1). Experimental results on face recognition [14–16] indicate that the method can be efficiently used for discriminant feature extraction. Similar to face recognition, the microarray data typically contains thousands of genes on each chip, and the number of the collected tumor samples is much smaller than that of the genes [17,18]. Therefore, gene expression data analysis also needs an effective discriminant analysis method for feature extraction, which aims to find the optimal projection such that the maximum margin criterion is maximized after the projection of the samples.

MMC is regarded as a nonparametric extension of LDA [19]. It measures between-class scatter based on marginal information, using the K-nearest neighbor technique. The nonparametric discriminant analysis method, however, encounters the problem of how to choose the optimal K. In addition, the weighting function used to deemphasize the samples far from the classification margin is too complicated. To solve these problems, in this paper, we propose a new maximum margin characterization method by virtue of sparse representation because it has a discriminative nature for classification [20]. The presented method, called sparse maximum margin discriminant analysis (SMMDA), can successfully avoid the difficulty of parameter selection and does not need the weighting function.

SMMDA, as a nonparametric discriminant analysis method for feature extraction, does not need to select the parameter K. Moreover, it obtains the number of margin samples flexibly by using a sparse representation technique. In addition, SMMDA can be used for gene selection. Gene selection has the following biological explanation: most of the abnormalities in cell behavior are due to irregular gene activities, and thus, it is critical to highlight these specific genes [18]. SMMDA could find the one-dimensional projection of gene expression data in the most separable direction; thus, we can use the sparse representation technique to regress the projection to obtain the relevance vector for the gene set and to select the genes according to the vector.

The remainder of this paper is organized as follows. Section 2 describes the method proposed in this paper. The maximum margin criterion is first presented, and the algorithms of sparse maximum margin discriminant analysis for feature extraction and gene selection are consequently given. Section 3 presents the numerical experiments. Section 4 concludes the paper and outlines directions for future work.

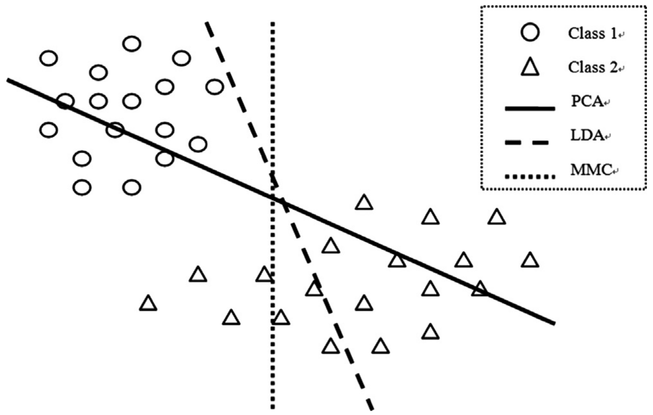

Fig. 1. An illustration of the behavior of PCA, LDA and MMC for a binary classification problem.

2. Methods

2.1. Maximum margin criterion

We are given $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \in \mathbb{R}^m \times \{C_1, ..., C_l\}$, where sample x_i is an m-dimensional vector and y_i is the corresponding class label for sample $x_i (i = 1, 2, ..., n)$. The maximum margin criterion aims to maximize the distances between classes after the transformation, and the criterion is [10]

$$J = \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} p_i p_j d(C_i, C_j) \tag{1}$$

We can use the distance between the mean vectors as the distance between classes, *i.e.*,

$$d(C_i, C_j) = d(\mathbf{m}_i, \mathbf{m}_j) \tag{2}$$

where $\mathbf{m}_i$ and $\mathbf{m}_j$ are the mean vectors of class C_i and class C_j, respectively. The variables p_i and p_j are a priori probabilities of class C_i and class C_j, respectively. Eq. (2) does not take the scatter of the classes into account; thus, it is not suitable for classification. Even if the distance between the mean vectors is large, it is not easy to separate two classes that have the large spread and that overlap with each other. In statistics, we usually use the overall variance $tr(\mathbf{S}_i)$ to measure the scatter of the data, where $\mathbf{S}_i$ is the covariance matrix of class C_i.

Then, we define the interclass distance as:

$$d(C_i, C_j) = d(\mathbf{m}_i, \mathbf{m}_j) - tr(\mathbf{S}_i) - tr(\mathbf{S}_j) \tag{3}$$

With Eq. (3), we can decompose Eq. (1) into two parts

$$J = \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} p_i p_j d(\mathbf{m}_i, \mathbf{m}_j) - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} p_i p_j (tr(\mathbf{S}_i) + tr(\mathbf{S}_j)) \tag{4}$$

The second part equals $tr(\mathbf{S}_w)$. By employing the Euclidean distance, the first part can be simplified to $tr(\mathbf{S}_b)$, which measures the overall variance of the class mean vectors.

Then, we obtain

$$J = tr(\mathbf{S}_b - \mathbf{S}_w) \tag{5}$$

A large J indicates that the class mean vectors scatter in a large space and that each class has a small spread. Additionally, a large J means that the samples in the same class are close to each other, while they are far from each other if the samples are in different classes. More details of the method can be found in [10].

Li et al. [10] proposed to use the maximum margin criterion to find projection vectors. Now, the criterion can be defined as

$$J = tr(W^T (\mathbf{S}_b - \mathbf{S}_w) W) \tag{6}$$

The projection vectors W that maximize Eq. (6) can be found as the eigenvectors of $S_b - S_w$ that correspond to the largest eigenvalues. The advantage of using the maximum margin criterion is that we need not compute the inverse of S_w; hence, the singularity problem can be avoided.

2.2. Sparse maximum margin discriminant analysis for feature extraction

When performing dimensionality reduction, we want to find a mapping from the measurement space to some feature space such that J is maximized after the transformation. However, the maximum margin criterion (MMC) is nonparametric discriminant analysis. It measures between-class scatter based on marginal information, using the K-nearest neighbor technique.