Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Pattern Recognition Letters 27 (2006) 1884–1892

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Gene subset selection in kernel-induced feature space

Satoshi Niijima [a,*], Satoru Kuhara [b]

[a] *Department of Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan*

[b] *Faculty of Agriculture, Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan*

Received 22 June 2005; received in revised form 14 February 2006
Available online 27 June 2006

Communicated by F. Roli

Abstract

This paper proposes a new filter approach to gene subset selection for kernel-based classifiers. We derive kernel forms of several well-known class separability criteria, and gene subset selection based on the kernelized criteria is applied to microarray cancer classification problems. The performance of our proposed strategy is compared in experiments with those of the conventional filter approach as well as gene ranking methods.

Keywords: Microarray; Gene selection; Cancer classification; Kernel methods

1. Introduction

Microarray technology allows us to measure the expression levels of thousands of genes simultaneously, producing a vast amount of data. Microarrays pose a great challenge on the data analysis, because the number of genes often exceeds tens of thousands, whereas the number of samples available is at most a few hundred.

In microarray data analysis, *gene selection* has been a central issue in recent years (see Stolovitzky (2003) and Cuperlovic-Culf et al. (2005) for reviews). Gene selection is often used to identify genes most relevant to a specific classification task, for example, those differentiate between normal and cancerous tissue samples. Gene selection plays several important roles in classification tasks (Krishnapuram et al., 2004; Hochreiter and Obermayer, 2004). It improves the prediction accuracy of classifiers by using only discriminative genes. It also saves computational costs by reducing dimensionality. More importantly, if it is possible to identify a small subset of biologically relevant genes, it may provide insights into understanding the underlying mechanism of a certain biological phenomenon. Also, such information can be useful for designing less expensive experiments by targeting only a small number of genes.

The most common gene selection approach used in practice is so-called *gene ranking*. It is a univariate approach in the sense that each gene is evaluated individually with respect to a certain criterion that represents class discrimination ability, and is ranked according to the assigned score. The criteria frequently used are *t*-statistics, the signal-to-noise ratio (Golub et al., 1999), the Fisher's criterion (Dudoit et al., 2002), information gain and χ^2 statistics (Liu et al., 2002) among others. Usually, top-ranked genes are selected for subsequent classification tasks. Although they are simple to implement, those genes that function in combination with other genes can be lowly ranked, if a single gene is not discriminative by itself. On the other hand, *gene subset selection* evaluates the discrimination ability of a subset of genes, hence a multivariate approach.

In general, gene selection can be performed in different manners: the *filter* and *wrapper* approaches. The filter approach employs some separability measure as an evaluation criterion, while the wrapper approach evaluates a subset of genes based on the performance of a specific

[*] Corresponding author. Tel.: +81 92 642 3043; fax: +81 92 642 3043.
E-mail addresses: niijima@grt.kyushu-u.ac.jp (S. Niijima), kuhara@grt.kyushu-u.ac.jp (S. Kuhara).

0167-8655/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.04.017

classifier. Although the wrapper approach often gives satisfactory classification accuracy (Inza et al., 2004; Xiong et al., 2001), evaluating the classification performance for each subset of genes demands expensive computational costs. Moreover, the selected genes may overfit the classifier used, hence there is no guarantee that they also give high performance for other classifiers.

Among many classification methods previously proposed, recent studies have demonstrated the effectiveness of kernel methods such as support vector machines (SVMs) in various classification problems of bioinformatics (Schölkopf et al., 2004). In this paper, we propose a new filter approach to gene subset selection for kernel methods. Our major goal is to enhance the performance of kernel-based classifiers without resorting to the wrapper approach. This can be realized by directly performing gene subset selection in a kernel-induced feature space, and also performing classification in the same feature space. It is different from the conventional filter approach combined with a kernel-based classifier, where gene subset selection is performed in input space.

We show that several well-known class separability criteria can be kernelized. Gene subset selection is performed based on the kernelized criteria. The performance of our strategy is assessed by combining it with three kernel-based classifiers, from simple to advanced ones. We apply it to cancer classification problems on acute lymphoblastic leukemia (ALL) (Yeoh et al., 2002) and compare the performance of our proposed strategy with those of gene ranking methods and the conventional filter approach to gene subset selection.

2. Gene subset selection

In classification problems using microarray data, we are given n samples consisting of gene expression levels of p genes, i.e. $\boldsymbol{x}_i \in \mathbb{R}^p (i = 1, \ldots, n)$, and the known class labels $y_i \in \{1, \ldots, C\}$ $(i = 1, \ldots, n)$, where C is the number of classes. The problem is to predict the class label of a given test sample with unknown class.

2.1. Kernelization of class separability criteria

In order to select an appropriate subset of genes for classification, an evaluation criterion that measures class separability is needed. In our study, we employ several common measures based on the within-class and total scatter matrices (Fukunaga, 1990; Theodoridis and Koutroumbas, 1999).

For simplicity, we treat binary cases, i.e. $C = 2$. As mentioned later, however, the proposed strategy can be used also in multiclass cases. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be a sample matrix containing $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ as rows. Then, the total scatter matrix $\boldsymbol{S}_{\mathrm{T}}$ can be expressed as

$$\boldsymbol{S}_{\mathrm{T}} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{X}^{\mathrm{T}}\vec{1}_n\vec{1}_n^{\mathrm{T}}\boldsymbol{X}, \tag{1}$$

where $(\cdot)^{\mathrm{T}}$ is the transpose operator, and $\vec{1}_n$ denotes an n-dimensional vector with all the components equal to $1/n$. We shall further denote a sample matrix for class k $(k = 1,2)$ by $\boldsymbol{X}_k \in \mathbb{R}^{n_k \times p}$, where n_k is the number of samples for class k. Then, the pooled within-class scatter matrix $\boldsymbol{S}_{\mathrm{W}}$ can be expressed as

$$\boldsymbol{S}_{\mathrm{W}} = \frac{n_1}{n}\boldsymbol{S}_1 + \frac{n_2}{n}\boldsymbol{S}_2, \tag{2}$$

where

$$\boldsymbol{S}_k = \frac{1}{n_k}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k - \boldsymbol{X}_k^{\mathrm{T}}\vec{1}_{n_k}\vec{1}_{n_k}^{\mathrm{T}}\boldsymbol{X}_k.$$

For obtaining good class separability, larger between-class scatter and smaller within-class scatter are preferable. Such criteria used in our study are

$$J_1 = \frac{\mathrm{Tr}(\boldsymbol{S}_{\mathrm{T}})}{\mathrm{Tr}(\boldsymbol{S}_{\mathrm{W}})}, \tag{3}$$

$$J_2 = \mathrm{Tr}(\boldsymbol{S}_{\mathrm{W}}^{-1}\boldsymbol{S}_{\mathrm{T}}), \tag{4}$$

and the Mahalanobis distance defined as

$$J_3 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{5}$$

where $\mathrm{Tr}(\cdot)$ and $(\cdot)^{-1}$ denote the trace and inverse of a matrix, respectively. $\boldsymbol{\mu}_k$ is a sample mean vector for class k, which can be written as

$$\boldsymbol{\mu}_k = \boldsymbol{X}_k^{\mathrm{T}}\vec{1}_{n_k}.$$

We may find gene subsets that maximize these criteria.

In terms of kernel methods, selecting a subset of genes based on the criteria corresponds to performing gene subset selection in *input space*. It turns out that these criteria can be rewritten by using the symmetric matrix

$$\boldsymbol{K} = \{\boldsymbol{x}_i \cdot \boldsymbol{x}_j\}_{i,j} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}, \tag{6}$$

which represents the inner products of samples.

Following Ruiz and López-de-Teruel (2001), let us define:

$$\boldsymbol{Z} = \left(\frac{1}{n}\boldsymbol{I}_n - \vec{1}_n\vec{1}_n^{\mathrm{T}}\right)^{1/2}, \tag{7}$$

$$\boldsymbol{Z}_k = \left(\frac{1}{n_k}\boldsymbol{I}_{n_k} - \vec{1}_{n_k}\vec{1}_{n_k}^{\mathrm{T}}\right)^{1/2}, \tag{8}$$

where $\boldsymbol{I}_n \in \mathbb{R}^{n \times n}$ and $\boldsymbol{I}_{n_k} \in \mathbb{R}^{n_k \times n_k}$ denote the identity matrices. Then, we have

$$J_1 = \frac{n\mathrm{Tr}(\boldsymbol{Z}^2\boldsymbol{K})}{n_1\mathrm{Tr}(\boldsymbol{Z}_1^2\boldsymbol{K}_{11}) + n_2\mathrm{Tr}(\boldsymbol{Z}_2^2\boldsymbol{K}_{22})}, \tag{9}$$

where

$$\boldsymbol{K}_{kk} = \boldsymbol{X}_k\boldsymbol{X}_k^{\mathrm{T}}.$$

From the Sherman–Morrison–Woodbury formula (Meyer, 2000) and Corollary 2 in (Ruiz and López-de-Teruel, 2001), we also have

$$J_2 = \mathrm{Tr}(\boldsymbol{K}_1\boldsymbol{V}(\boldsymbol{I}_{n_1} - \boldsymbol{K}_{12}((\boldsymbol{W}_2^2)^{-1} + \boldsymbol{K}_{12}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{K}_{12})^{-1}\boldsymbol{K}_{12}^{\mathrm{T}}\boldsymbol{V})\boldsymbol{K}_1^{\mathrm{T}}\boldsymbol{Z}^2), \tag{10}$$