Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 39 (2006) 2426–2438
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Pattern classification in DNA microarray data of multiple tumor types

Tsun-Chen Lin[a,*], Ru-Sheng Liu[a], Chien-Yu Chen[c], Ya-Ting Chao[a,b], Shu-Yuan Chen[a]

[a]*Department of Computer Science and Engineering, Yuan Ze University, 135 Yuan-Tung Rd., Nei-Li, Chung-Li, Taoyuan 32026, Taiwan, ROC*
[b]*Graduate School of Biotechnology and Bioinformatics, Yuan Ze University, 135 Yuan-Tung Rd., Nei-Li, Chung-Li, Taoyuan 32026, Taiwan, ROC*
[c]*Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan, ROC*

Received 30 June 2005; received in revised form 24 December 2005; accepted 9 January 2006

Abstract

In this paper, we propose a genetic algorithm with silhouette statistics as discriminant function (GASS) for gene selection and pattern recognition. The proposed method evaluates gene expression patterns for discriminating heterogeneous cancers. Distance metrics and classification rules have also been analyzed to design a GASS with high classification accuracy. Moreover, the proposed method is compared to previously published methods. Various experimental results show that our method is effective for classifying the NCI60, the GCM and the SRBCTs datasets. Moreover, GASS outperforms other existing methods in both the leave-one-out cross validations and the independent test for novel data.

Keywords: Gene expression profiling; Cancer classification; Genetic algorithm; Silhouette statistics

1. Introduction

Microarray technology allows large-scale parallel measurements for the expression of many thousands of genes and produces a very large amount of genetic data. One of the most promising applications of this technology is as a useful tool for tumor classification through gene expression patterns. Generally, the classification of microarray data from tumors can be considered as a problem consisting of two tasks: gene selection and classification. Gene selection is the recognition of important genes from thousands of highly correlated gene expression profiles to capture the informative genes for pattern recognition. Classification requires the construction of a model, which processes input patterns representing objects, and predicts the class or category associated with the objects under consideration. In the past few years, many algorithms [1–5] with rank-based gene selection schemes have been applied to 2- or 3-class classification problems based on the gene expression data, and most of them have achieved 95–100% classification accuracy.

Recently, Li et al. [6] evaluated 7 commonly used classifiers based on rank-based feature selection methods for use in classifying multiclass datasets. The results showed that all methodologies performed much better for datasets with a small number of classes, and if the tumor classification problem was expanded to multiple tumor classes, such as the 9-class NCI60 dataset and the 14-class GCM dataset, the performance of these methods would deteriorate significantly. The challenge is that the data are of high dimensionality and the sample size is small. As one of the future directions discussed in this earlier paper, Li et al. [6] suggested designing a feature selection method to consider the correlations between features. However, the classification problem for multiclass data is much more difficult than that for the 2-class data. Romualdi et al. [7] presented a simulation approach to control variations between patients and to evaluate a series of supervised statistical techniques. This simulation results showed that all methodologies have comparable performance when the number of patient samples per tumor type is greater than 50, the number of tumors is lower than 4 and the number of discriminating genes is larger than 40. In practice there might be over 100 types of cancer and potentially even more subtypes [8], and since microarray

* Corresponding author. Tel.: +886 3 8210872; fax: +886 3 8340169.
E-mail address: lintsunc@ms6.hinet.net (T.-C. Lin).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.01.004

experiments are still too costly and time consuming, thereby limiting the number of samples. Thus a dataset with a small sample size containing many classes will make variations of gene expressions within a class more accentuated and will reduce the performance of classifiers.

While the classification problem on multiclass data still remains a challenge, instead of ranking genes for feature selection, three recent approaches [9–11] based on GA are proposed to summarize the high-dimensional input spaces and to search near-optimal groups of interacting genes in chromosomes for discriminant analyses. These analyses are in turn used to measure the effectiveness of the features selected in a chromosome on the actual classification task for the separation of multiclass of cancers. Since computational models based on GA have already shown a better degree of accuracy than rank-based methods in multiclass microarray classification, we herein propose using GA to identify a set of correlated features in a chromosome and to evaluate them by silhouette statistics with different distance metrics to filter out the key features for classification.

In this paper, we will explore genes with the best discrimination ability in classifying tumor samples. Since most genes in a microarray are irrelevant to class distinction, we first follow the criterion established by Hall and Smith [12] that good predictor sets should contain genes that are strongly correlated to class distinction but each of these genes should be as uncorrelated with each other as possible, and to thereby select differentiated genes using between-group to within-group sum of squares ratios (BSS/WSS) [13] for data dimensionality reduction. Second, the proposed GASS uses a block of genes (a chromosome of genes) to find a minimal variation across the samples in the actual classification task to capture the similarity pattern of gene expressions. Finally, the GASS with one-minus-Pearson distance metric, which is invariant to monotone changes, is used to minimize the variations between samples so as to effectively find gene expression patterns for sample classification. Experimental results show that the proposed GASS method outperforms many existing methods.

2. Related works in multiclass classification using genetic algorithms

Recently, genetic algorithms have been mainly applied as gene selectors to consider the dependency or correlation within a subset of genes for microarray classification tasks [9–11,14,15]. Generally, these GA approaches embed the gene selection method into the classifier. The chromosomes, or predictors, in the population are used to construct the subset of genes in making a prediction for grouping the data into well-separated clusters, with each cluster corresponding to the same type of cancer. The fitness function or scoring function, used to evaluate the prediction capability of chromosomes, is usually built by calculating the leave-one-out cross-validation (LOOCV). Since sizes of most sample sets are small, crossover operator and mutation operator of the GA are involved to extend the search space so as to minimize the training errors of classifiers. Some recently published techniques based on GA as gene selection methods for multiclass classification problems are listed below for comparison with the GASS method.

- *GA/MLHD* (maximum likelihood classifier) [9]: the main concept of this approach is to use the maximum likelihood classifier that quantitatively evaluates the correlation among covariates (a chromosome of selected genes) into a multi-dimensional structure of data. The truncated NCI and GCM datasets of 1000 genes with highest standard deviation of gene expression among samples were used throughout the experiments. The fitness of chromosomes was computed by building the function of LOOCV error rate on training data, and adding the test error rate on testing data so that classification accuracies of 95% with 13 predictive genes and 86% with 32 predictive genes on testing data were obtained for NCI data and GCM data, respectively. Although those authors claimed successful results, the error rate estimation method used for performance determination was different from the common error rate estimators. Therefore, those authors also claimed a modified fitness function that did not consider the independent test error rate in training process. The results showed a significant drop of accuracy from previous reports. (See the author's website http://www.omniarray.com/bioinformatics/GA/supplement.pdf)
- *GA/SVM* (all paired support vector machines) [10]: the GA/SVM was used for multiclass cancer identification. A preliminary selection of the top 1000 genes with the highest standard deviation of gene expression level among samples were applied on the original NCI and GCM datasets to obtain the truncated dataset for use as the analysis data set throughout this approach. The approach utilized the same toolbox based on the traditional genetic operators as GA/MLHD used to compare the performance between these two techniques. The fitness function was built by calculating the prediction error rate of the LOOCV test using all samples in the dataset. Classification accuracies of 88.52% with 40 predictive genes and 80.99% with 40 predictive genes were achieved for NCI data and GCM data, respectively.
- *GESSES* (genetic evolution of sub-sets of expressed sequences) [11]: this approach marked useful genes to build a large number of different subsets of genes so as in order to obtain a population of the highest scoring predictors by random mutations and deletions of genes, according to how the mutations affect the scoring function estimated by the LOOCV test. The best predictor is the one giving the fewest mistakes on test data. Some gene preselection methods were involved