Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mathematical Biosciences 193 (2005) 79–100
Mathematical Biosciences
www.elsevier.com/locate/mbs

Multi-class clustering and prediction in the analysis of microarray data

Chen-An Tsai [a], Te-Chang Lee [b,c], I-Ching Ho [c], Ueng-Cheng Yang [d], Chun-Houh Chen [e], James J. Chen [a,*]

[a] *Division of Biometry and Risk Assessment, National Center for Toxicological Research, Food and Drug Administration NCTR/FDA/HFT-20 Jefferson, AR 72079, USA*
[b] *Institute of Biopharmaceutical Sciences, National Yang-Ming University, Taipei, 112 Taiwan*
[c] *Institute of Biomedical Sciences, Academia Sinica, Taipei, 115 Taiwan*
[d] *Institute of Biochemistry, National Yang-Ming University, Taipei, 112 Taiwan*
[e] *Institute of Statistical Science, Academia Sinica, Taipei, 115 Taiwan*

Received 24 November 2003; received in revised form 7 June 2004; accepted 27 July 2004
Available online 28 December 2004

Abstract

DNA microarray technology provides tools for studying the expression profiles of a large number of distinct genes simultaneously. This technology has been applied to sample clustering and sample prediction. Because of a large number of genes measured, many of the genes in the original data set are irrelevant to the analysis. Selection of discriminatory genes is critical to the accuracy of clustering and prediction. This paper considers statistical significance testing approach to selecting discriminatory gene sets for multi-class clustering and prediction of experimental samples. A toxicogenomic data set with nine treatments (a control and eight metals, As, Cd, Ni, Cr, Sb, Pb, Cu, and AsV with a total of 55 samples) is used to illustrate a general framework of the approach. Among four selected gene sets, a gene set Ω_I formed by the intersection of the F-test and the set of the union of one-versus-all t-tests performs the best in terms of clustering as well as prediction. Hierarchical and two modified partition (k-means) methods all show that the set Ω_I is able to group the 55 samples into seven clusters reasonably well, in which the As and AsV samples are considered as one cluster (the same group) as are the Cd and Cu samples. With respect to prediction, the overall

* Corresponding author. Tel.: +1 870 543 7007; fax: +1 870 543 7662.
E-mail address: jchen@nctr.fda.gov (J.J. Chen).

0025-5564/$ - see front matter
doi:10.1016/j.mbs.2004.07.002

accuracy for the gene set Ω_I using the nearest neighbors algorithm to predict 55 samples into one of the nine treatments is 85%.

Keywords: Bagged clustering; Bagging fuzzy clustering; Gene selection; *k*-nn classification; Rand statistic; Shaded similarity matrix plot

1. Introduction

DNA microarray technology provides tools for studying the expression profiles of a large number of distinct genes simultaneously. Common objectives in microarray experiments are gene identification (class comparison), class discovery, and class prediction. Gene identification is to select overexpressed or under-expressed genes from studying expression profiles between different samples (e.g., with or without exposure to a specific drug or toxic compound). Class discovery usually refers to identifying previously unknown sample subtypes from the study of gene expression profiles. Class prediction is to predict the class membership of a new sample based on a gene-expression prediction function.

Cluster analysis techniques have been applied to organize gene expression data by grouping genes or samples with similar patterns of expression; refer to [7,10]. In clustering samples overexpression levels of multiple genes, the expression patterns of the samples in the same group are more homogeneous as compared to the expression patterns in the other group. Clustering samples is used to characterize similar/distinct samples or to discover new sample classes. Hierarchical and *k*-means are two commonly used cluster analysis for class discovery. The hierarchical clustering algorithm forms clusters in a hierarchical fashion resulting in a tree-like dendrogram. In the *k*-means clustering procedure, genes are divided into *k* partitions or groups with each partition representing a cluster of genes. Therefore, as opposed to the hierarchical clustering, the number of clusters must be known (decided) a priori. Cluster analysis is considered as an unsupervised method of analysis because no information about sample grouping is used.

Parallel to unsupervised clustering algorithms for class discovery, class prediction uses supervised discriminant algorithms to classify samples into known groups. The goal of class prediction is to develop a decision rule that accurately predicts the class membership of a new sample based on the expression profiles of some key genes. Several supervised discriminant algorithms have been adopted for classification of cancer subtypes or gene functions. Discrimination analysis methods include Fisher's linear discriminant function, classification tree, nearest-neighbor classifiers, and support vector machines; refer to [6,16]. Recently, Dudoit et al. [6] compared the performance of different discrimination algorithms for tumor classification. They concluded that simple classification methods such as linear discriminant and nearest neighbors methods tend to perform well compared to other more sophisticated methods.

Microarray gene expression data are characterized by the number of variables (genes) far exceeding the number of samples. This presents challenges for supervised discriminant algorithms, which are generally designed with large number of samples over few variables. A common problem is overfitting the data [19]. That is, the predicted model can fit the original data well but may predict poorly for new data. Supervised discriminant algorithms, typically, involve a training