Classification of Acute Leukemia Based on Gene Expression from DNA Microarrays Using Partial Least Squares

Authors: Danh V. Nguyen and David M. Rocke

Contact Address: Department of Statistics
University of California
1 Shields Ave.
Davis, CA 95616
Phone: (530) 752-5854

Analysts of microarray data when presented with raw gene expression intensity data often take two main steps when analyzing the data: (1) preprocess the data by rescaling and standardizing so that overall intensities for each array are equivalent and then (2) apply some statistical methodologies to answer scientific questions of interest. In this paper step 2 involves statistical classification and dimension reduction methodologies. For the data preprocessing step, we present a thresholding algorithm for applying to each array. The algorithm may be used in conjunction with some current data preprocessing and thresholding. The algorithm converges to a ``cutoff point'' for p gene expressions on a given array. The analyst can then decide to analyze genes with expression measurements above this cut off point, or use the information from the algorithm for array rescaling. The thresholding algorithm have two parameters: (a) the percentage (q) of the smallest expression values in the array to form the initial set and (b) the number of median absolute deviations (MAD) above the median, c, to determine the cutoff point. The algorithm is robust to outlying observations and is not sensitive to the first parameter q. As an example, the algorithm is applied to a leukemia dataset (high-density oligonucleotide arrays of Golub et al. (Oct. 1999)) and we also outline the procedure for use with cDNA arrays. After applying the thresholding algorithm, we used the method of partial least squares (PLS) and principal componenet analysis (PCA) to reduce the dimension of microarray data by constructing a few components of the gene expressions. For classification problems, we illustrate the use of PLS components in quadratic discriminant analysis (QDA) and logistic discrimination (LD). The classification results of QDA based on PLS components are favorable, compared to the original results of Golub et al. on the same dataset. However, LD classification using PLS components completely separates the samples in both training and test datasets. To investigate whether this result is coincidental we randomize the samples into training and test datasets 100 times and repeated the classifcation procedures. This empirical result add further evidence (consistent with Press and Wilson (1978)) and explaination as to why LD performed better than QDA, even under dimension reduction context. Theoretical and detailed simulation results of the methodologies proposed here are given elsewhere (Nguyen and Rocke, 2000, 2000b, 2000c). However, we briefly describe the results of a simulation study comparing classification using PLS components and principal components (PCs) which shed favorable light on the use of PLS over ordinary PC for microarray data. A discussion of other potential uses of PLS in analyzing gene expression data including (1) prediction of a target gene based on remaining genes and (2) assessing survival experience based on gene expression as covariates are also described.