How Many Genes Are Needed for a Discriminant Microarray Data Analysis?

Wentian Li
Rockefeller University
Lab of Statistical Genetics,
Box 192 Rockefeller University
1230 York Avenue New York, NY 10021
USA
212-327-7977
212-327-7996
wli@linkage.rockefeller.edu
CAMDA00 Dataset 2: Leukemia

Wentian Li, Yaning Yang

The analysis of the second data set (from the MIT group) is a discriminant analysis or supervised learning. Among thousands of genes whose expression level is measured, not all are needed for discriminant analysis: a gene may either not contribute to the separation of two types of tissues/cancers, or it may be redundant because it is highly correlated with other genes. There are two theoretical frameworks in which variable selection (or gene selection in our case) can be addressed. The first is model selection, and the second is model averaging. We have carried out model selection using Akaike information criterion and Bayesian information criterion within logistic regression (discrimination, predictor, classifier) to determine the number of genes that provide the best model. These model selection criteria set an upper limit of 22-25 and 12-13 genes for the sample size of this data (38), and the best model consists of only one (no.4847, zyxin) or two genes. We have also carried out model averaging over the best single-gene logistic predictors using three different weights: maximized likelihood, prediction rate on training set, and equal weight. We have observed that the performance of these weighted predictors on the testing set is gradually reduced as more genes are included, but a clear cutoff that separates good and bad prediction performance is not found.

Keywords

discriminant analysis, variable selection, model selection, model averaging, Akaike information criterion(AIC), Bayesian information criterion (BIC)

Tools

we use the s-plus statistical package

website

http://linkage.rockefeller.edu/wli/