

Statistical methods for sample classification and prediction with microarray gene expression data

Xiaohong Huang, Ph.D.
University of Minnesota

Abstract:

Using gene expression data to classify sample types or patient survivals has received much research attention recently. To accommodate special features of gene expression data, several new methods have been proposed, including a weighted voting scheme of Golub et al (1999), a compound covariate method of Hedenfalk et al (2001) (originally proposed by Tukey (1993)), and a shrunken centroids method of Tibshirani et al (2002). These methods look different and are more or less ad hoc. Here we point out a close connection of the three methods with a linear regression model and partial least squares (PLS). Under the general framework of PLS, we propose a penalized PLS (PPLS) method that can handle both categorical (for classification) and continuous (e.g. survival times) responses. Using real data, we show the competitive performance of our proposal when compared with other methods.

This is a joint work with Wei Pan (Biostatistics, U of Minnesota) and Jennifer Hall (Medicine, U of Minnesota).