Gene Selection for Tissue Classification of Microarray Data

Chen-An Tsai

Biostatistics and Bioinformatics Unit
Comprehensive Cancer Center
University of Alabama at Birmingham

Abstract

DNA microarray technology provides useful tools for profiling global gene expression patterns in different cell/tissue samples. A challenge in predicting diagnostic classes using microarray data is that the number of genes is significantly larger than the number of samples, and only a subset of potential genes could provide distinguishing different tissue types. Our approaches are based on statistical discriminative measures to rank genes with respect to differential expressions between tissues. We use the significance testing approach and information of family-wise error rate and false discovery rate to identify gene sets needed for prediction. Example data sets are used to illustrate how to select marker genes for sample classification. In addition, two well established classification algorithms, k nearest neighbors (kNN) and support vector machine (SVM), are used to evaluate the selected gene sets.