A Filtering Method of Gene Expression Data for Multi-type Disease Classification.

Tzu-Ming Chu and B.S. Weir, Department of Statistics, North Carolina State University

Gene expression array technology can be used to distinguish nearly identical disease sub-types. Here, we implement a data mining approach for multi-type disease classification. First, we describe a filtering method (BKW test) that combines the Bootstrap resampling and the Kruskal-Wallis statistics to select significant genes. Then, principal components analysis and k-means clustering are applied to data, which contains information of all significant genes, for succeeding classification. In this paper, SAS Enterprise Miner software is used to analyze leukemia data sets (http://www.genome.wi.mit.edu/MPR) for disease sub-type classifications. For two-type disease classification, 95% of training data and 97% of test data are correctly classified respectively, based on the classification rules from the training data. Similarly, for three-type leukemia classification, the correct classification rates are 92% and 97% for training data and test data.

Bioinformatics Research Center	Phone: (919) 515-3574
Department of Statistics	Fax: (919) 515-7315
North Carolina State University	URL: http://statgen.ncsu.edu
Raleigh NC 27695-7566	email: weir@stat.ncsu.edu