Best subset selection algorithm for classification with buit-in cross-validation

Moreno V, Guinó E, López-Doriga A, Berenguer A, Solé X

Biostatistics and Bioinformatics Unit, Catalan Institute of Oncology, Hospitalet (Barcelona), Spain

A common problem in the analysis of microarray data is the identification of the best subset of genes that provides an accurate classification of samples according to some characteristic. The excessive dimensionality of the genes relative to samples usually generates highly variable solutions depending on the statistical procedure used to solve the problem and low external validity, often anticipated from poor classification performance under bootstrapping or cross validation tests.

We propose a classification procedure based on genetic algorithm ideas. A sub-sample of cases is randomly selected for training and a set of genes are also sampled as proposal predictors. Then a classification prediction rule is learned form the proposal on the training sample and validated on the remaining samples. If classification accuracy is within a correct range, new proposal predictors are sampled and tested. Otherwise, the candidates are rejected. New train and test samples are selected at each cycle and the algorithm iterates until no new predictors are incorporated. If the number of candidates grows, some may be also randomly dropped and accuracy tested for deterioration.

An example of application to the identification of genes differentially expressed in colorectal cancer samples will be presented, together with an evaluation of the performance of the procedure an comparison with other techniques.