Identify breast cancer subtype by gene expression profiles

白其卉
Shieh GS;Bai CH;Lee C

Abstract

Abstract: Support vector machines (SVMs), with linear, polynomial and radial kernels, were applied to classify subtypes of breast cancer by gene expression profiles of tissues samples. Using the top 500 genes ranked by between-group to within-group sum of squares, SVMs with linear kernel had an average accuracy rate about 97% when applied to a balanced dataset; this accuracy rate was significantly higher than that of the original data. After imputation, the smallest subsample of the balanced dataset was comparable to the other subsamples' (containing more than 10 samples). In biomedical sciences, it is of interest to identify genes that can be used to classify subtypes of breast cancer well. Using SVMs, we identified 500 genes and looked up the functions of 297 genes from databases. Furthermore, about 65% of these 297 genes were known to be related to breast cancer, and this confirms the consistency of our results with existing biomedical knowledge. Those 203 genes may also be investigated further to see if they are involved in breast cancer; any novel findings will be important.