Applying Machine Learning Techniques to Analysis of Gene Expression Data: Cancer Diagnosis

Kyu-Baek Hwang, Dong-Yeon Cho, Sang-Wook Park, Sung-Dong Kim, Byoung-Tak Zhang

School of Computer Science and Engineering,
Seoul National University, Korea.
{kbhwang, dycho, swpark, sdkim, btzhang}@scai.snu.ac.kr

Bayesian networks, neural trees, and radial basis function networks are used for the analysis of CAMDA data set 2. Bayesian networks represent statistical relationships among various variables, and are useful to analyze gene expression patterns and to examine statistical properties of dependence and conditional independence in the data. The nodes in the Bayesian network correspond to each gene and the values of the nodes represent the gene expression level. In building Bayesian networks, we choose 10 genes for its nodes and convert all the values of data into categorical values to train the networks. Neural tree models represent conventional neural networks as a tree structure. They have heterogeneous neuron types in a single network, and the connectivity of the neurons is irregular and sparse. An evolutionary algorithm has been used to find the appropriate structure and weights of neural trees for classifying gene expression data from leukemia patients. In this method, essential genes for this classification are included into neural trees and less important genes are weed out automatically. Radial basis function (RBF) networks are similar to multi-layer perceptron networks, but their hidden neurons contain RBF, a statistical transformation based on a Gaussian distribution. RBF makes the influence of data be localized, so the case when a particular data significantly influences the output of the network is reduced. Experiments are performed with varying the number of input genes and the number of hidden neurons.

The training data was 100% correctly classified by all the methods. For the test data, the results depend on the parameters and the structures of the networks. The error rate for the test data is expected to decrease by refining network structure, optimizing parameters through training, and choosing more appropriate genes as input to the network.