Boston University Statistics Seminar Series

Applications of Relative Expression Reversal Concept in Cancer Classification

Aik Choon Tan
The Institute for Computational Medicine,
Johns Hopkins University School of Medicine and Whiting School of Engineering,
Baltimore, MD 21218, USA

Thursday, October 12, 2006, 4:00-5:00pm
Mathematics and Computer Science (MCS) Building, Room 149
111 Cummington Street, Boston

Tea and Cookies at 3:30pm in MCS 153

Abstract: Various studies have shown that cancer tissue samples can be successfully detected and classified by their gene expression patterns using machine learning approaches. One of the challenges in applying these techniques for classifying gene expression data is to extract accurate, readily interpretable rules providing biological insight as to how classification is performed. Current methods generate classifiers that are accurate but difficult to interpret. This is the trade-off between credibility and comprehensibility of the classifiers. Here, we introduce a new classifier in order to address these problems. It is referred to as k-TSP (k-Top Scoring Pairs) and is based on the concept of 'relative expression reversals'. This method generates simple and accurate decision rules that only involve a small number of gene-to-gene expression comparisons, thereby facilitating follow-up studies. We have compared our approach to other machine learning techniques for class prediction in 19 binary and multi-class gene expression datasets involving human cancers. The k-TSP classifier performs as efficiently as Prediction Analysis of Microarray and support vector machine, and outperforms other learning methods (decision trees, k-nearest neighbour and nave Bayes). Our approach is easy to interpret as the classifier involves only a small number of informative genes. For these reasons, we consider the k-TSP method to be a useful tool for cancer classification from microarray gene expression data. Furthermore, we show that our method can be applied to cross-platform analysis. Our results show that the classifier built from the marker gene pair, which simply compares relative expression values, achieves high accuracy, sensitivity and specificity on independent datasets generated using various array platforms. Our findings suggest a new model for the discovery of marker genes from accumulated microarray data and demonstrate how the great wealth of microarray data can be exploited to increase the power of statistical analysis.

For directions and maps, please see *http://math.bu.edu/research/statistics/statseminar.html.*