Name: Xian Xu
Email: xianxu@cse.buffalo.edu, xianxu@gmail.com
Dissertation Title: INTEGRATED FEATURE SUBSET SELECTION/EXTRACTION WITH
APPLICATIONS IN BIOINFORMATICS

Committee Members: Dr. Aidong Zhang (chair)
Dr. Matthew J. Beal
Dr. Murali Ramanathan
Outside Reader: Dr. Yulan Liang

Abstract:

Feature subset selection and extraction algorithms are actively and extensively studied in machine learning literature to reduce the dimensionality of feature space, since high dimensional data sets are generally not efficiently and effectively handled by a large array of machine learning and pattern recognition algorithms. When we stride into the analysis of large scale bioinformatics data sets, such as microarray gene expression data sets, the high dimensionality of feature space compounded with the low dimensionality of sample space, creates even more problems for data analysis algorithms.

Two foremost characteristics of microarray gene expression data sets are: 1. the correlation between features (genes) and 2. the availability of domain knowledge in computable format. In this dissertation, we will study effective feature selection and extraction algorithms with applications to the analysis of the new emerging data sets in the bioinformatics domain. Microarray gene expression data set, the result of large scale RNA profiling techniques, is our primary focus in this thesis. Several novel feature (gene) selection and extraction algorithms are proposed to deal with peculiarities on microarray gene expression data set.

To address the first characteristic of the microarray gene expression data set, we first propose a general feature selection algorithm called Boost Feature Subset Selection (BFSS) based on permutation analysis to broaden the scope of selected gene set and thus improve classification performance. In BFSS, subsequent features to be selected focus on those samples where previously selected features fail. Our experiments showed the benefit of BFSS for t-score and S2N (signal to noise) based single gene scores on a variety of publicly available microarray gene expression data sets.

We then examine the correlations among features (genes) explicitly to

see if such correlations are informative for the purpose of sample classification. This results in our gene extraction algorithm called virtual gene. A virtual gene is a group of genes whose expression levels are combined linearly. The combined expression levels of a virtual gene instead of the real gene expression levels are used for sample classification. Our experiments confirm that by taking into consideration the correlations between gene pairs, we could indeed build a better sample classifier.

Microarray gene expression data set only represents one aspect of our knowledge of the underlying biological system. Currently there are lots of biological knowledge in computable format that can be accessed from Internet. Continue to address the second characteristic of the microarray gene expression data set, we investigate the integration of domain knowledge, such as those imbedded in gene ontology annotations, for the use of gene selection and extraction. GO annotation enables us to investigate correlations among bigger groups of genes in an informed way and thus expand beyond pairwise virtual gene algorithm. Correlations between biologically related genes are examined and used for sample classification. Our experiments showed considerable improvement in the term of classification accuracy. GO annotations also enable us to suppress false positives in selected gene set, which becomes an increasing problem for gene selection algorithms on microarray gene data set due to the limited number of samples.