ABSTRACT

Differential expression analysis and feature selection is central to gene expression microarray data analysis. Standard approaches are flawed with the arbitrary assignment of cut-off parameters and the inability to adapt to the particular data set under analysis. Presented in this thesis are three novel approaches to microarray data feature selection and differential expression analysis based on various machine learning and soft computing paradigms. The first approach uses a Separability Index to select ranked genes, making gene selection less arbitrary and more data intrinsic. The second approach is a novel gene ranking system, the Fuzzy Gene Filter, which provides a more holistic and adaptive approach to ranking genes. The third approach is based on a Stochastic Search paradigm and uses the Population Based Incremental Learning algorithm to identify an optimal gene set with maximum inter-class distinction.

All three approaches were implemented and tested on a number of data sets and the results compared to those of standard approaches. The Separability Index approach attained a K-Nearest Neighbour classification accuracy of 92%, outperforming the standard approach which attained an accuracy of 89.6%. The gene list identified also displayed significant functional enrichment. The Fuzzy Gene Filter also outperformed standard approaches, attaining significantly higher accuracies for all of the classifiers tested, on both data sets ($p < 0.0231$ for the prostate data set and $p < 0.1888$ for the lymphoma data set). Population Based Incremental Learning outperformed Genetic Algorithm, identifying a maximum Separability Index of 97.04% (as opposed to 96.39%).

Future developments include incorporating biological knowledge when ranking genes using the Fuzzy Gene Filter as well as incorporating a functional enrichment assessment in the fitness function of the Population Based Incremental Learning algorithm.