Abstract

Statistical Learning of Gene Annotations in Saccharomyces Cerevisiae

Miles Trochesset

Master of Science

Department of Computer Science

University of Toronto

2004

This work presents an investigation of machine learning techniques which can be applied to quantitative data obtained from experiments on yeast. Our goal is to predict the function of genes, more specifically for those which currently have no known function.

We investigate data-preprocessing methods, and develop two algorithms for filling missing values in the datasets. This is a necessary step before some statistical methods can be used to predict gene function.

We present two standard machine learning algorithms, one used in a non-standard way, for predicting the biological functions of genes in a systematic and comprehensive manner. Determining gene function is simplified to a series of binary classifications and one of the challenges of this learning task lies in the extremely small number of positives, compared with large amounts of negatives samples. We develop a method based on hierarchical clustering used with labeled data to search for regions of high positive concentrations and make predictions for the unlabeled genes. We investigate logistic regression as a baseline for comparing to our technique. Both of these methods are based on different views of the data and we found that depending on the biological processes, one or the other of these approaches performs better, although our method makes more confident predictions for more biological processes.

The outcomes of the research are threefold: first we present two algorithms for missing value estimation. Second we build a new biological data mining method based on existing

machine learning tools that are readily accepted in the biological community. Third we make biological predictions of gene functions, each associated with a level of confidence and all above 50% precision.