Vinnie Ramesh

[vramesh@mit.edu](mailto:vramesh@mit.edu)

**Integrating Genomic Data Sets to Improve Disease Prediction Ability of Bayesian Networks**

The purpose of this research is to investigate the behavior and effects of heterogeneous data sets and Bayesian network prediction ability. The ultimate objectives are threefold:

* Identify if and when combining (heterogeneous) experiment data in bayes nets result in better predictors than bayes nets made from a single homogeneous experiment data source
* Investigate this behavior under a variety of different diseases, amount of training data, and number of features
* Explain the underlying mechanisms by which we can develop better predictors

This will be investigated in the context of gene expression data. Data sources (experiments) measure, given a particular phenotype, gene expressions of patient populations. Bayes networks classify whether a particular patient with a certain gene expression profile will display a particular phenotype (obesity, diabetes, etc). Initial data are standard GDS experiments from ncbi.nlm.nih.gov/pub/geo/DATA/SOFT/GDS\_full

We are interested in understanding the circumstances under which Bayesian networks become better predictors when adding additional training data from different GEO data sets related to the same disease. We are also interested in investigating how and why predictors made from heterogeneous sources of data are better than those made from homogenous sources of data. To clarify, a heterogeneous data source includes several different GEO data sets for a disease, whereas a homogeneous data source might only consider a single experiment from GEO. Previous research suggests that combining several different (heterogeneous) experiments may result in better overall predictors than combining homogeneous experiments.

The first goal of this project is to investigate different ways of combining heterogeneous experiments and observe their effect on prediction ability of Bayesian networks. We are particularly interested in looking for thresholds where improvements in prediction ability are statistically significant over bayes nets made from homogenous experiment data.

This will be done as follows:

* Given some number of experiments (experiments test gene expression for a particular phenotype), rank the significant genes expressed in each experiment
* Determine if the expression of particular genes are significant across control and non-control populations for each experiment
* Compare the top n% of significant genes across experiments
* Use these common significant genes to construct a bayes network
* Test how different amounts and combinations of data (experiments – both homogenous and heterogeneous) affect the prediction power of these bayes nets
* Use a feature selection algorithm to calculate the top x features. See how prediction ability varies as the number of features varies.
* Report statistically significant improvements in prediction power (and the conditions under which they occur)

After showing the existence of thresholds where increases in prediction ability for heterogeneous experiment data occur, and where these thresholds occur, the next steps are to investigate why these jumps occur

The goal is to complete these 3 objectives for UAP. In order to get some good results, these have to be completed for as many GEO diseases as possible. If time permits, it would be useful to run similar analysis on other types of data sets (TCGA, etc).