Introduction

Prediction of phenotype (ASD vs. Control) utilizing gut microbiome composition would help characterize biological associations, direct further work in the field, and may ultimately lead to an early diagnostic tool or therapeutic. We used supervised learning approaches to make predictions, unsupervised clustering to reduce the high-dimensionality variance, and next aim to use factor analysis to identify latent variables in microbial composition across samples so as to characterize and classify phenotype.

Data

The dataset was provided by Wall Lab at Stanford University. 168 sequencing on each sample tells us which taxa are present and their quantitative abundance. Every case sample has one (or two in a few cases) age-matched, environmentally-matched sibling control sample(s). After QC, we have 109 samples and 1007 bacterial taxa.

Boosting and Bias/Variance Diagnostics

We elected to fit a boosted ensemble of decision trees because of this model’s robustness to outliers and monotone transformations of the inputs, and because of its ability to straitly the feature space with nonlinear boundaries. We allowed each weak learner (each tree) to grow up to five splits in order to capture interaction effects between bacterial taxa, and we used a shrinkage factor of 0.001 and subsampling of a 0.5 fraction of the training data at each iteration of boosting in order to mitigate overfitting due to high variance.

We used 10-fold cross-validation over the boosted model on the full dataset and determined that the optimal test error was achieved when the model included 21 trees. A plot of the training and 10-fold cross-validation error indicated that boosting did not seem to generally reduce cross-validation error as additional trees were added to the ensemble; in fact, the model seems to start overfitting soon after the start of the boosting algorithm (See Figure 3).

Naive Bayes as a First Approach

We used k-means clustering on the taxa and collapsing the taxa down to cluster centroids to reduce dimensionality and attempt to capture latent relationships between taxa. We determined that using between 4 and 7 clusters resulted in some improvement to overall model performance. Thus, we preprocessed the data by running k-means with k =7 over the taxa and then collapsing the sample vectors from approximately 1000 taxa measurements down to 7 taxa centroids computed using the cluster labels.

The optimal test error was achieved with 310 trees. Although the boosting algorithm is now able to fit more trees before the onset of over-fitting, the overall improvement to the model is marginal as the minimum Bernoulli deviance achieved is not much lower than it was previously (See Figure 5).

In order to assess the models performance with respect to bias and variance, we trained the model over a range of proportions of the data, testing each time on the left-over-hold-out data. We then plotted how the training and error tests varied with the size of the training set. These diagnostics were performed for both the gbm model on the full dataset and the gbm model on the reduced dataset.

Results

We used k-means clustering on the taxa and collapsing the taxa down to cluster centroids to reduce dimensionality and attempt to capture latent relationships between taxa. We determined that using between 4 and 7 clusters resulted in some improvement to overall model performance. Thus, we preprocessed the data by running k-means with k =7 over the taxa and then collapsing the sample vectors from approximately 1000 taxa measurements down to 7 taxa centroids computed using the cluster labels.

The optimal test error was achieved with 310 trees. Although the boosting algorithm is now able to fit more trees before the onset of over-fitting, the overall improvement to the model is marginal as the minimum Bernoulli deviance achieved is not much lower than it was previously (See Figure 5).

In order to assess the models performance with respect to bias and variance, we trained the model over a range of proportions of the data, testing each time on the left-over-hold-out data. We then plotted how the training and error tests varied with the size of the training set. These diagnostics were performed for both the gbm model on the full dataset and the gbm model on the reduced dataset.

Future Work

Given that dimensionality reduction will be needed to improve performance, we would like to look more deeply into this area. A model capable of capturing more nuanced interactions between tax distributions, like corenseeding, may produce more interesting dimensionality reduction. Additionally, using domain knowledge to aggregate those taxa that are known to occupy the same niche or produce the same metabolites may prove very useful. It would also be interesting to work with matched pair machine learning, to leverage our dataset to its fullest potential. A high-dimensional factor analysis may prove useful in searching for latent factors which explain the variation within the autistic and non-autistic groups. The factor loadings can then be examined in detail in order to discover relationships between informative taxa.

References