Investigating Autism and the Human Microbiome

Christine A. Tataru, Michael D. Salerno, Filip M. Zivkovic

CS229 Final Project, Autumn 2016

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. 16S sequencing on each sample tells us which taxa are present and their quantitative abundance. Every case sample has one (or two in fewer cases) age-matched, environmentally-matched sibling control sample(s). After QC, we have 109 samples and 1007 bacterial taxa.

Naive Bayes as a First Approach

Figure 1: K-Fold CV for Each Sibling Pair Using NB with Laplace Smoothing

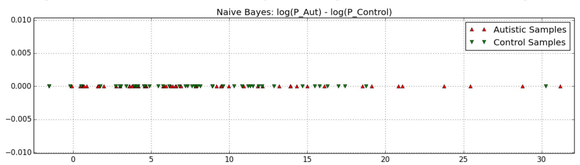
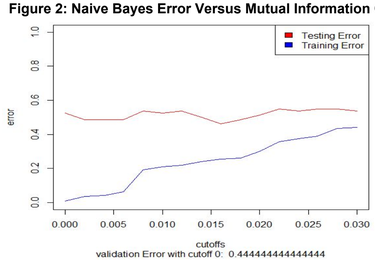


Figure 2: Naive Bayes Error Versus Mutual Information Cutoff



validation Error with cutoff 0: 0.444444444444444

Naive Bayes with Laplace Smoothing is a suitable first approach, providing insight into the data. Figure 1 was generated using k-fold CV, individually holding out each sibling pair. The plot has visible clustering, is highly biased to predict autistic classification, and most of the data is inseparable. However, at the max and min values (where the most confident predictions are found), the accuracy is seen to improve. Each sample is abundant with bacterial taxa; however, adding a firm Mutual Information cut-off fails to reduce testing error. Naive Bayes suffers from too much bias to be a useful classifier on this dataset, indicating linear boundaries cannot separate this data.

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 stratify 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).

Figure 3: GBM, Full Dataset

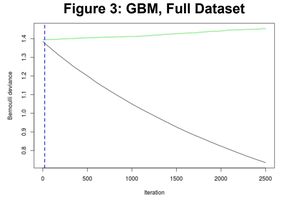
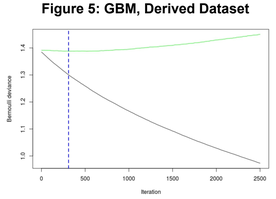


Figure 5: GBM, Derived Dataset



K-Means for Dimensionality Reduction

The overfitting suggests that additional trees are generally picking up noise. This may be due to the high dimensionality and general sparsity of the data.

We use kmeans 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 ap- proximately 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 previ ously (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 ploted how the training and test errors 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.

Boosting Continued

On the full dataset, the test error and training error flatten out at high values and with a small gap between each other as training set size increases, suggesting that model bias is an issue here (See Figure 4). On the reduced dataset, we now observe that both the test error and training error appear to be decreasing with increasing training set size at the right cut-off (See figure 6). It is possible that k-means over the feature space was able to capture latent relationships between taxa, allowing the model to capture more information in spite of the small sample size. However, both training and test error are still quite high, indicating that we still have a bias problem.

Figure 4: Diagnostic, Full Dataset

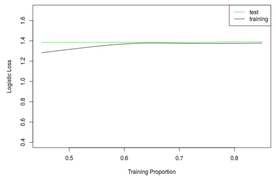
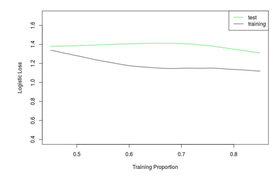
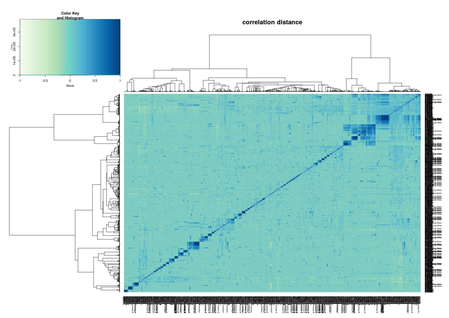


Figure 6: Diagnostic, Derived Dataset



Correlation Heatmap

Figure 7: Latent Groups Among the Taxa



Results

Using 20% holdout CV, we have generated the following result.

Model	Train Error	Test Error
Naive Bayes	.0125	.4444
GBM	.341	.485
GBM with Kmeans reduction	.381	.530

Discussion

Although we did not accurately predict phenotype from gut microbiome, this does not mean there is not a connection. Firstly, our problem was prone to very high variance, with small sample size in the face of a large feature space. While K-means reduction was meant to mitigate this variance, more informed dimensionality reduction may be more fruitful. Secondly, most ML algorithms assume independence of samples, while our dataset had carefully controlled criteria of sibling pairs. We mandated that our train and test sets never split sibling pairs, however, our algorithms did not adequately exploit this structure. Lastly, there is a significant portion of autism that can be explained by genetics. It is possible that our dataset is not complete enough to reveal interesting correlations, but that combining it with the genetic data modality will result in greater predictive power.

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 taxa distributions, like autoencoding, 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 order to discover relationships between informative taxa.

References

1. Pinto-Martin JA, Young LM, Mandell DS, Poghosyan L, Giarelli E, Levy SE. Screening strategies for autism spectrum disorders in pediatric primary care. J Dev Behav Pediatr 2008; 29: 345350.

2. Hsiao, Elaine Y et al. "Microbiota modulate behavioral and physiological abnormali- ties associated with neurodevelopmental disorders." Cell 155.7 (2013): 1451-1463.

3. Parracho, Helena MRT et al. "Differences between the gut microflora of children with autistic spectrum disorders and that of healthy children." Journal of medical microbiology 54.10 (2005): 987-991.

4. Caporaso et al., 2010; Edgar, 2010 Caporaso JG, Kuczynski J, Stombaugh J, Bit- tinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods. 2010 May;7(5):3356. PMCID: PMC3156573

5. Wall DP et al. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiatry. 2012;2:e100. PMCID: PMC3337074

6. Duda M et al. Testing the accuracy of an observation-based classifier for rapid detec- tion of autism risk. Transl Psychiatry. 2014;4:e424. PMCID: PMC4150240

7. La Rosa PS et al. Hypothesis testing and power calculations for taxonomic-based hu- man microbiome data. PLoS ONE. 2012;7(12):e52078. PMCID: PMC3527355

8. Lozupone, C, Knight, R. UniFrac: a New Phylogenetic Method for Comparing Mi- crobial Communities. American Society for Microbiology. 2005; 71(12): 9. Arumugam, M. et al. Enterotypes of the Human Gut Microbiome. Nature. U.S. National Library of Medicine, 12 May 2011. Web. 20 Nov. 2016.