Autism and The Human Microbiome

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

1 Introduction

Autism Spectrum Disorder (ASD) is a heterogeneous developmental disorder that affects 1 in 68 children. The current behavioral diagnostics are only applicable late in development; development of an accurate, early-stage, non-behavioral classifier could circumvent the timing challenges of diagnosis, allowing for early and therefore more effective interventions. Our goal is to create an accurate machine learning classifiers to predict the autism phenotype from gut microbiome composition. We have 16S sequencing data depicting the gut flora composition of 52 children with Autism Spectrum Disorder(ASD) and that of their 52 age-matched siblings without ASD. Samples are represented as a vector of abundances of taxa.

From the predictive power of our models, we cannot claim causative relationships, however, we can infer association between gut microbiome and autism which will provide specific avenues for further mechanism of action experiments and function as a potential screening diagnostic. Additionally, unsupervised approaches enable the discovery of latent variable structure which can be used to infer relationships between taxa as well as to further inform supervised methods.

2 Related Work

There is a significant compilation of work done on discovering the effect of the gut microbiome on neurological function and vice-versa; this connection is known as the gut brain axis[12]. In the case of autism, Hsiao et. al. found that feeding ASD phenotype mice commensal Bacteroides fragilis ameliorates their autistic symptoms[2]. Research also shows that more than 50% of children with autism also experience GI dysfunction[10]. Work on the connection between gut microbiome and autism in humans has historically suffered from incredibly small sample sizes and lack of environmentally matched controls, both deficits that our data seek to address[3,11]. Additionally, ML approaches have rarely been applied to these datasets, with researchers mostly presenting more conservative statistical methods.

3 Dataset and Features

Our dataset, obtained from the Wall lab at Stanford University, was created to investigate the link between the gut microbiome and autism. The dataset includes 108 samples from 54 families. Each family has one child with an autism diagnosis and one without, each within 2-7 years of age, and within 2 years of age of each other. These constraints were placed to modulate environmental variation (young children will live together, eat mostly the same food, have the same pet exposure, etc.). Samples represent a wide-spread geographic area, from California to New York to Canada,

and a variety of landscapes (urban, rural, etc.).

Each participant had a stool sample sequenced to obtained counts of bacterial taxa present in their gut. After sequencing, reads were cleaned for errors using the Dada2 pipeline[14] and aligned to microbial database GreenGenes[15] to identify their species of origin. To account for batch effects and differences in sequencing depth, data was normalized using Cumulative Sum Scoring [13]

4 Models, Algorithms, & Diagnostics

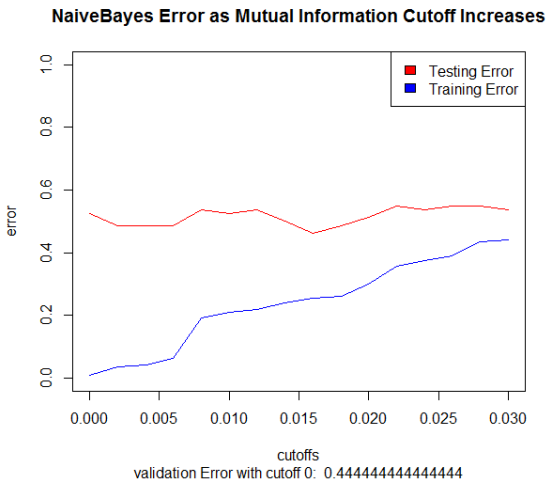
4.1 Naive Bayes

We used Naive Bayes, in spite of its strong independence assumptions, as a simple first-pass model for autism classification using microbiome data.

$$p(y = "Aut"|taxa) = \frac{\prod_{i=1}^{n} p(taxa_i|y = "Aut")p(y = "Aut")}{\prod_{i=1}^{n} p(taxa_i|y = "Aut")p(y = "Aut") + \prod_{i=1}^{n} p(taxa_i|y = "Control")p(y = "Control")}$$

We separated the data into a train and validation sets with an 80/20 split. We trained the model on a subset of taxa with high mutual information (MI) with the autism phenotype within the train set. We performed LOOCV on the train set over increasing MI cutoffs (higher cutoff equates to fewer features). We picked the cutoff that gave us the best test error during LOOCV, trained another classifier using this cutoff on the aforementioned train set, and then evaluated our model on the validation set.

We see evidence of overfitting which is mitigated as we apply increasingly strict MI cutoffs. As we decrease feature space, train error rises as expected, but test error does not improve, which suggests high bias. Next, we sought a model that would provide us with more capacity to capture the patterns of our data, without the imposed independence assumptions of naive bayes.



4.2 Boosted Decision Trees

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 non-linear boundaries. See Gred Ridgway's guide to generalized boosted models[16] for a detailed specification of the model and software used. Using Bernoulli loss, we allowed each weak learner (each tree) to grow up to five splits in order to capture interaction effects between OTUs, 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 1). This suggests that additional trees are generally picking up noise. This may be due to the high dimensionality and general sparsity of the data.

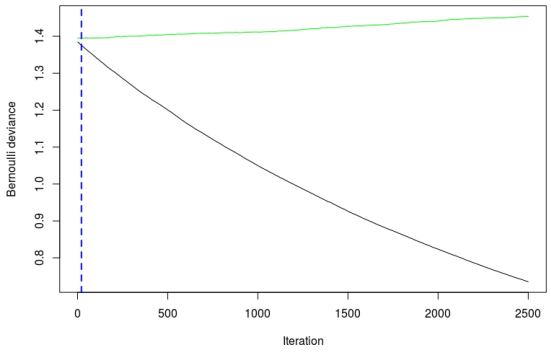


Figure 1: GBM: Full Dataset

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

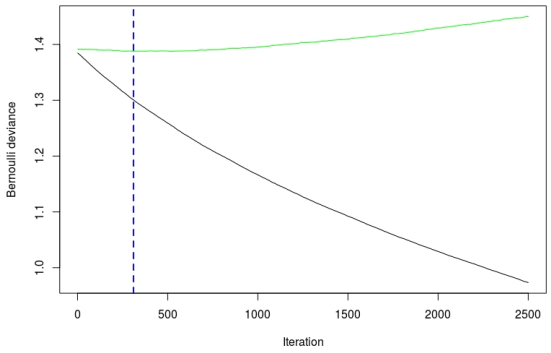


Figure 2: GBM: Reduced Dataset

The optimal test error was achieved with 310 trees. Although the boosting algorithm is now able to fit more trees before the onset of overfitting, the overall improvement to the model is marginal as the minimum Bernoulli deviance

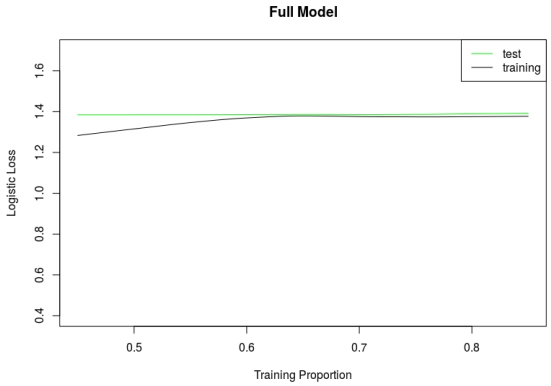


Figure 3: Diagnostic: Full Dataset

achieved is not much lower than it was previously (See Figure 2).

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.

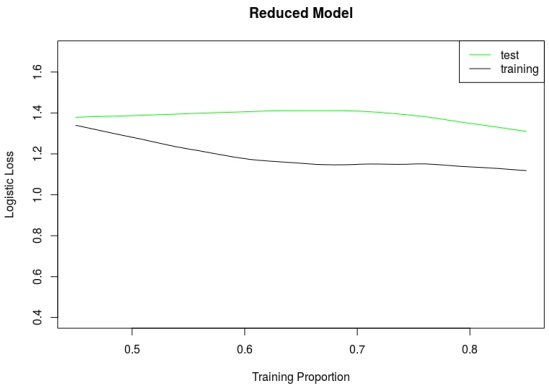


Figure 4: Diagnostic: Reduced Dataset

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 3). 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 4). It is possible that k-means over the OTUs was able to capture latent relationships between OTUs, allowing the model capture some signal 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.

4.3 High Dimensional Factor Analysis

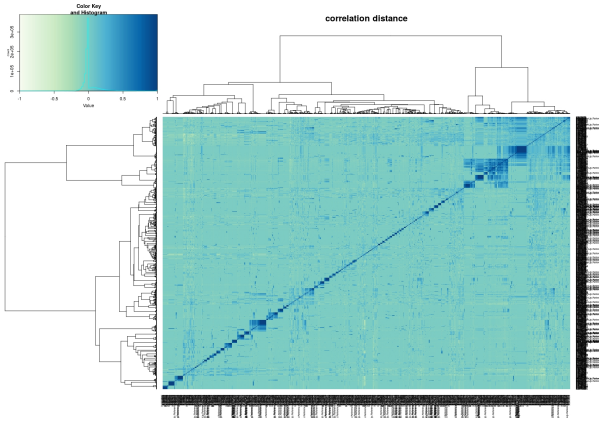


Figure 5: Hierarchical Correlation Heatmap

One of the unique characteristics of the dataset used for this project is it consists of well-controlled paired samples where one carries a positive label and the other a negative label. While this complicates the application of supervised learning algorithms on this dataset by virtue of the correlations that exist between pairs, it also presents an opportunity to analyze the structure of the covariates of the positively labeled samples and negatively labeled samples independently of each other. We accomplished this by fitting a factor analysis model on the full dataset and then examining the resulting factor loadings for the positively labeled samples and negatively labeled samples separately. See Bai & Li[17] for a detailed specification of the factor analysis model used.

Summarizing the original correlation matrix using a hierarchical correlation heatmap reveals that, although the majority of the taxa are uncorrelated, there exists some correlation/covariance structure which is illustrated along the diagonal and the corners of the heatmap (See Figure 5). We fit a factor model to the microbiome data using a total of 20 factors. In order to measure the goodness of fit of the factor model, we examined how well it reproduced the original correlation matrix. A correlation statistic of 0.76 was achieved when comparing the off-diagonal values of the original correlation matrix to that of the fitted factors. This result is impressive

when considering that the original data consists of 988 covariates. This suggests that the factor model is effectively capturing non-trivial covariance structure via the 20 latent factors. The following table summarizes the each of the factor loadings across all positively labeled samples:

F1	F2	F3	F4	F5	F6	F7
Min. :-0.29560	Min. :-1.34202	Min. :-8.86628	Min. :-0.37860	Min. :-0.46071	Min. :-3.70942	Min. :-0.87489
1st Qu.:-0.18225	1st Qu.: 0.02694	1st Qu.:-0.09012	1st Qu.:-0.19794	1st Qu.:-0.29455	1st Qu.:-0.65573	1st Qu.:-0.38218
Median :-0.12655	Median : 0.14940	Median : 0.20055	Median :-0.12350	Median :-0.19536	Median :-0.01126	Median :-0.14626
Mean : 0.09257	Mean : 0.04835	Mean :-0.14088	Mean :-0.11071	Mean : 0.11960	Mean : 0.11461	Mean : 0.04067
3rd Qu.:-0.02074	3rd Qu.: 0.20741	3rd Qu.: 0.30462	3rd Qu.:-0.05667	3rd Qu.:-0.05035	3rd Qu.: 0.78265	3rd Qu.: 0.19251
Max. :10.17115	Max. : 0.46477	Max. : 0.46756	Max. : 0.27360	Max. : 9.54197	Max. : 3.06091	Max. : 9.28002

F8	F9	F10	F11	F12	F13	F14
Min. :-5.2243	Min. :-0.52450	Min. :-0.70977	Min. :-1.75249	Min. :-0.6940	Min. :-0.88826	Min. :-0.89357
1st Qu.:-0.1428	1st Qu.:-0.27273	1st Qu.:-0.27168	1st Qu.:-0.42802	1st Qu.:-0.3247	1st Qu.:-0.02088	1st Qu.:-0.27391
Median : 0.2114	Median :-0.10835	Median :-0.10047	Median :-0.19927	Median :-0.1558	Median : 0.12299	Median :-0.09558
Mean :-0.1155	Mean :-0.06960	Mean :-0.01129	Mean : 0.04623	Mean : 0.1042	Mean : 0.08754	Mean :-0.01747
3rd Qu.: 0.5894	3rd Qu.: 0.05218	3rd Qu.: 0.08131	3rd Qu.: 0.18991	3rd Qu.: 0.1778	3rd Qu.: 0.23750	3rd Qu.: 0.21849
Max. : 1.1881	Max. : 1.85924	Max. : 3.38254	Max. : 7.76009	Max. : 9.3008	Max. : 0.90035	Max. : 1.74655

F15	F16	F17	F18	F19	F20
Min. :-0.644783	Min. :-1.45006	Min. :-1.47366	Min. :-1.17678	Min. :-7.147395	Min. :-3.583538
1st Qu.:-0.309740	1st Qu.:-0.04659	1st Qu.:-0.35021	1st Qu.:-0.29268	1st Qu.: 0.001193	1st Qu.:-0.000948
Median :-0.165370	Median : 0.12834	Median :-0.06774	Median :-0.13968	Median : 0.131359	Median : 0.205799
Mean :-0.087435	Mean : 0.07216	Mean :-0.01530	Mean :-0.03319	Mean : 0.040279	Mean : 0.074000
3rd Qu.:-0.007577	3rd Qu.: 0.26617	3rd Qu.: 0.10372	3rd Qu.: 0.13714	3rd Qu.: 0.376445	3rd Qu.: 0.356296
Max. : 3.127732	Max. : 0.46133	Max. : 4.82231	Max. : 3.45039	Max. : 1.250024	Max. : 1.450077

In contrast, following table contains the same summary statistics computed across all negatively labeled samples:

F1	F2	F3	F4	F5	F6	F7
Min. :-0.29055	Min. :-9.97116	Min. :-2.8644	Min. :-0.401003	Min. :-0.54733	Min. :-2.4817	Min. :-0.93301
1st Qu.:-0.17890	1st Qu.: 0.09473	1st Qu.: 0.1697	1st Qu.:-0.168626	1st Qu.:-0.29027	1st Qu.:-0.7029	1st Qu.:-0.27764
Median :-0.14420	Median : 0.15954	Median : 0.2424	Median :-0.104237	Median :-0.18904	Median :-0.2352	Median :-0.01769
Mean :-0.09970	Mean :-0.05207	Mean : 0.1517	Mean : 0.119225	Mean :-0.12879	Mean :-0.1234	Mean :-0.04380
3rd Qu.:-0.04567	3rd Qu.: 0.20953	3rd Qu.: 0.3120	3rd Qu.:-0.008959	3rd Qu.:-0.08799	3rd Qu.: 0.4025	3rd Qu.: 0.18053
Max. : 0.91189	Max. : 0.40510	Max. : 0.4714	Max. :10.187418	Max. : 1.19787	Max. : 2.0928	Max. : 1.78348

F8	F9	F10	F11	F12	F13	F14
Min. :-2.2199	Min. :-0.59420	Min. :-0.67216	Min. :-1.17218	Min. :-0.87974	Min. :-9.77051	Min. :-0.745673
1st Qu.:-0.1393	1st Qu.:-0.27978	1st Qu.:-0.33722	1st Qu.:-0.41703	1st Qu.:-0.34633	1st Qu.:-0.07975	1st Qu.:-0.349287
Median : 0.1995	Median :-0.08150	Median :-0.19640	Median :-0.20041	Median :-0.15712	Median : 0.11955	Median :-0.219507
Mean : 0.1244	Mean : 0.07495	Mean : 0.01216	Mean :-0.04979	Mean :-0.11225	Mean :-0.09428	Mean : 0.018815
3rd Qu.: 0.5578	3rd Qu.: 0.04666	3rd Qu.:-0.05654	3rd Qu.: 0.23064	3rd Qu.: 0.01863	3rd Qu.: 0.26479	3rd Qu.:-0.007028
Max. : 1.6191	Max. : 9.77208	Max. : 9.09681	Max. : 3.05046	Max. : 1.16574	Max. : 1.00107	Max. : 9.274429

F15	F16	F17	F18	F19	F20
Min. :-0.65103	Min. :-9.80345	Min. :-1.18173	Min. :-0.67072	Min. :-5.90842	Min. :-8.19866
1st Qu.:-0.27688	1st Qu.:-0.02021	1st Qu.:-0.29120	1st Qu.:-0.31571	1st Qu.:-0.14309	1st Qu.:-0.07496
Median :-0.10543	Median : 0.15117	Median :-0.15408	Median :-0.20804	Median : 0.12917	Median : 0.15330
Mean : 0.09416	Mean :-0.07771	Mean : 0.01648	Mean : 0.03574	Mean :-0.04338	Mean :-0.07969
3rd Qu.: 0.03996	3rd Qu.: 0.25451	3rd Qu.: 0.05485	3rd Qu.: 0.02771	3rd Qu.: 0.32963	3rd Qu.: 0.27410
Max. : 9.14875	Max. : 0.61177	Max. : 7.85350	Max. : 8.87667	Max. : 0.89329	Max. : 1.13246

These factor loadings indicate how each latent factor is associated with the observable taxa. One of the interesting phenomena within the above summaries is that the corresponding mean loadings for each factor take on similar but oppositely signed values. This indicates that, on the average, corresponding factors across the two groups have the characteristic opposite relationship with observable taxa. It will be valuable to examine this phenomenon more carefully with the guidance of a domain expert in order inform the next steps in the analysis.

5 Conclusions & Future Work

For the purpose of supervised prediction of the autism phenotype, microbiome data presents several challenges. The data tends to be very sparse and high in dimension compared to the number of samples. For this reason, it is beneficial to perform some sort of dimensionality reduction prior to

training a supervised model on the taxa. Doing so appears to improve overall model performance, however, the supervised models used still suffered from what appeared to be very high bias and variance. Collecting more samples and adding informative features for future analysis may alleviate these problems and help in diagnosing what the sources of error might be.

Another challenge results from the the measures taken to control for outside factors. Because the data is a collection of siblings where one is diagnosed as autistic and the other isn't, there exist strong pairwise correlations throughout the dataset. We ultimately made use of this idiosyncrasy by fitting a factor model and examining how latent factor loadings differed between the two label groups. This revealed an, on average, approximately equal but opposite relationship between corresponding factors from the two groups and the taxa.

Lastly, the fact that the factor model was able to produce a good fit to the original covariance matrix using only 20 factors supports the notion of subgroups within the autistic group. This is also supported by the fact that k-means had some effectiveness in increasing the prediction accuracy of boosting. By examining the patterns among taxa and observations on which the factors load more heavily, we may be able uncover more evidence of such enterotypes[9].

Our next step will be to consult with a domain expert in order to determine how the factor loadings should be interpreted. The hope is that what we learn will lead us to insights which we could used to better inform a supervised model or to inform future research and future collections of microbiome data. We are also interested in the idea of using factor analysis to fit a separate density to each label group and then form a discriminant function that can be used for prediction, not unlike the ideas behind Gaussian discriminant analysis.

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 abnormalities 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, Bittinger 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 detection of autism risk. Transl Psychiatry. 2014;4:e424. PMCID: PMC4150240

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

8. Lozupone, C, Knight, R. UniFrac: a New Phylogenetic Method for Comparing Microbial 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.

10. Horvath K et. al.Gastrointestinal abnormalities in children with autistic disorder. Journal of Pediatrics, 1999 Nov. 135(5):559-63.https://www.ncbi.nlm.nih.gov/pubmed/10547242?dopt=Abstract

11. Krajmalnik-Brown R, Gut bacteria in children with autism spectrum disorders: challenges and promise of studying how a complex community influences a complex disease. Microbial Ecology in Health and Disease, 2015 Mar 12;26:26914. https://www.ncbi.nlm.nih.gov/pubmed/25769266

12. Li Q., Zhou J.M. The microbiotagutbrain axis and its potential therapeutic role in autism spectrum disorder. Neuroscience, 2 June 2016, 324;131139http://www.sciencedirect.com/science/article/pii/S0306452216002360

13. Paulson J. Robust methods for differential abundance analysis in

marker gene surveys. Nat Methods. 2013 Dec; 10(12): 12001202. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4010126/

14. Callahan et. al. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13, 581583 (2016)http://www.nature.com/nmeth/journal/v13/n7/full/nmeth.3869.html

15. DeSantis TZ. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology, 2006 Jul;72(7):5069-72 https://www.ncbi.nlm.nih.gov/pubmed/16820507

16. Greg Ridgeway, Generalized Boosted Models: A guide to the gbm package. 3 Aug. 2007; http://www.saedsayad.com/docs/gbm2.pdf

17. Bai Jushuan, Li Kunpeng, Statistical Analysis of Factor Models of High Dimension. Annals of Statistics, 2012, Vol. 40, No. 1, 436 - 465. http://www.columbia.edu/ jb3064/papers/2012_Statistical_analysis_of_factor_models_of_high_dimension.pdf