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\textsuperscript{12}. In the case of autism, Hsiao et. al. found that feeding ASD phenotype mice commensal Bacteroides fragilis ameliorates their autistic symptoms\textsuperscript{2}. Research also shows that more than 50% of children with autism also experience GI dysfunction\textsuperscript{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\textsuperscript{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 and aligned to microbial database GreenGenes to identify their species of origin. To account for batch effects and differences in sequencing depth, data was normalized using Cumulative Sum Scoring.

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.

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

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\(^\text{16}\) 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.

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.

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
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 plotted 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.

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

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\textsuperscript{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:

<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min: -0.29569</td>
<td>Min: 1.34202</td>
<td>Min: -0.86628</td>
<td>Min: -0.37860</td>
<td>Min: -0.46671</td>
<td>Min: 1.70942</td>
<td>Min: -0.87439</td>
</tr>
<tr>
<td>1st Qu.: -0.18225</td>
<td>1st Qu.: 0.06294</td>
<td>1st Qu.: -0.09812</td>
<td>1st Qu.: -0.19794</td>
<td>1st Qu.: -0.29455</td>
<td>1st Qu.: -0.65573</td>
<td>1st Qu.: -0.38218</td>
</tr>
<tr>
<td>Median: -0.12653</td>
<td>Median: 0.12940</td>
<td>Median: 0.20053</td>
<td>Median: -0.12256</td>
<td>Median: -0.19536</td>
<td>Median: -0.01126</td>
<td>Median: 0.14626</td>
</tr>
<tr>
<td>Mean: 0.82572</td>
<td>Mean: 0.08435</td>
<td>Mean: -0.14988</td>
<td>Mean: -0.11671</td>
<td>Mean: 0.11980</td>
<td>Mean: 0.11461</td>
<td>Mean: -0.30467</td>
</tr>
<tr>
<td>3rd Qu.: 0.20974</td>
<td>3rd Qu.: 0.30741</td>
<td>3rd Qu.: 0.36062</td>
<td>3rd Qu.: 0.05667</td>
<td>3rd Qu.: -0.00635</td>
<td>3rd Qu.: 0.78203</td>
<td>3rd Qu.: 0.19221</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F8</th>
<th>F9</th>
<th>F10</th>
<th>F11</th>
<th>F12</th>
<th>F13</th>
<th>F14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min: -0.22437</td>
<td>Min: -0.52400</td>
<td>Min: -0.70977</td>
<td>Min: -1.75249</td>
<td>Min: -0.6908</td>
<td>Min: -0.88626</td>
<td>Min: -0.93357</td>
</tr>
<tr>
<td>1st Qu.: -0.4428</td>
<td>1st Qu.: -0.27273</td>
<td>1st Qu.: -0.27164</td>
<td>1st Qu.: -0.24912</td>
<td>1st Qu.: -0.3247</td>
<td>1st Qu.: -0.28088</td>
<td>1st Qu.: -0.27391</td>
</tr>
<tr>
<td>Median: -0.2114</td>
<td>Median: -0.10835</td>
<td>Median: -0.19047</td>
<td>Median: -0.19297</td>
<td>Median: -0.1550</td>
<td>Median: -0.12299</td>
<td>Median: -0.09558</td>
</tr>
<tr>
<td>Mean: -0.11550</td>
<td>Mean: -0.06966</td>
<td>Mean: -0.08112</td>
<td>Mean: 0.84623</td>
<td>Mean: 0.1042</td>
<td>Mean: 0.08764</td>
<td>Mean: -0.01747</td>
</tr>
<tr>
<td>3rd Qu.: 0.85904</td>
<td>3rd Qu.: 0.085218</td>
<td>3rd Qu.: 0.28813</td>
<td>3rd Qu.: 0.18991</td>
<td>3rd Qu.: 0.1778</td>
<td>3rd Qu.: 0.25750</td>
<td>3rd Qu.: 0.21849</td>
</tr>
<tr>
<td>Max: 1.1199</td>
<td>Max: 1.6524</td>
<td>Max: 1.38284</td>
<td>Max: 7.76099</td>
<td>Max: 9.2000</td>
<td>Max: 0.90335</td>
<td>Max: 1.74655</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>F6</th>
<th>F7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min: -0.64473</td>
<td>Min: -1.45066</td>
<td>Min: -1.47366</td>
<td>Min: -1.17678</td>
<td>Min: 0.147395</td>
<td>Min: 3.58338</td>
<td></td>
</tr>
<tr>
<td>1st Qu.: -0.39974</td>
<td>1st Qu.: -0.84659</td>
<td>1st Qu.: -0.35921</td>
<td>1st Qu.: -0.29268</td>
<td>1st Qu.: -0.00193</td>
<td>1st Qu.: -0.590948</td>
<td></td>
</tr>
<tr>
<td>Median: -0.16537</td>
<td>Median: 0.12308</td>
<td>Median: -0.66774</td>
<td>Median: -0.16498</td>
<td>Median: 0.123357</td>
<td>Median: -0.207999</td>
<td></td>
</tr>
<tr>
<td>Mean: -0.057425</td>
<td>Mean: 0.07216</td>
<td>Mean: -0.01256</td>
<td>Mean: -0.01339</td>
<td>Mean: 0.009279</td>
<td>Mean: 0.074809</td>
<td></td>
</tr>
<tr>
<td>3rd Qu.: 0.687577</td>
<td>3rd Qu.: 0.26617</td>
<td>3rd Qu.: 0.18372</td>
<td>3rd Qu.: 0.13714</td>
<td>3rd Qu.: 0.37445</td>
<td>3rd Qu.: 0.356296</td>
<td></td>
</tr>
<tr>
<td>Max: 3.127732</td>
<td>Max: 0.46133</td>
<td>Max: 4.82231</td>
<td>Max: 3.45939</td>
<td>Max: 1.250024</td>
<td>Max: 1.450677</td>
<td></td>
</tr>
</tbody>
</table>

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 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.

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


13. Paulson J. Robust methods for differential abundance analysis in
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4010126/


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
