BIOF 509 Final Project Report

Project Topic:

Using Machine Learning to Classify Children with Autism Spectrum Disorder Based on Gut Microbiome

Introduction:

Autism Spectrum Disorder (ASD) is a range of neurodevelopmental disorders including Asperger’s Syndrome and Autism. Psychologists usually diagnose ASD by classifying a child’s behavior based on social interactions. If the child has a difficult time forming bonds, communicating with others, and displays other abnormal social behaviors, the child can be diagnosed with ASD. While behavioral tests have been the standard for diagnosing ASD for decades, more quantitative methods are needed for scientifically diagnosing this mental disability.

ASD has been correlated with gastrointestinal problems and changes in gut microbiota. Whether this causes ASD or it is simply an effect of ASD children usually having picky diets remains unknown. However, we can use this information to build machine learning classifiers based on differential microbiota abundance between ASD children and children without ASD. 16s rRNA sequencing and shotgun meta-abundance are two powerful methods used for sequencing the microbes in the gut. With this information, we can quantify the abundance of certain microbiota in children with and without ASD and build machine learning models to classify ASD status based on the gut microbiome.

Description of Datasets:

I found my datasets on the website “Kaggle” where two preprocessed datasets in the form of csv files were given. Here is the link to the dataset: <https://www.kaggle.com/antaresnyc/human-gut-microbiome-with-asd>. This work is also published in *Gut Microbes*: <https://www.tandfonline.com/doi/full/10.1080/19490976.2020.1747329>. The researchers used gut microbiota to perform principle component analysis (PCA) on children with and without ASD. They also build a machine learning classifier in R that uses the random forest model.

In the 16s rRNA dataset, the first two columns serve as indexers, with the second column holding the information for microbiota taxonomy. In the meta-abundance dataset, the first column contains the information for microbiota taxonomy. The first row for both datasets displays each patient ID number, with the gut microbiota for each patient represented in a column for each patient. Patients which have the letter “A” before their ID number are children with ASD, while patients with the letter “B” before their ID number are children without ASD, which serve as controls.

Description of Methods:

Step 1 – Preprocessing: There are several lowly abundant microbiotas, which slow down machine learning algorithms and do not hold much if any weight on the analysis. Therefore, filtered out these lowly abundant microbiotas. Then, I normalized the microbiotas based for each patient. Lastly, I transposed both datasets so that each patient was represented as a row and every microbiota was a feature (column) in the transposed dataframes. I also made ASD status binary (with 1 representing ASD positive and 0 representing ASD negative).

Step 2 – Dimensionality Reduction: I wrote functions to plot PCA and UMAP plots from the processed dataframes. The PCA function plots both a scree plot and a PCA plot to see any large differences between patients according to ASD status. The UMAP plot functions in a similar fashion.

Step 3 – Running Supervised Learning Algorithms: I decided to use some common supervised machine learning algorithms on the processed dataframes, such as logistic regression, SVM, and random forest. The paper used random forest, but I wanted to try some extras for optimization. I importing these models from sklearn and used 10-fold K Folds cross validation on each model. I wrote a function to automate this and return the scores for each model as well as a dataframe which includes the scores. For random forest, I wrote a for loop for testing different numbers of trees and later decided on an optimal number of trees based on a combination of score accuracy and run time. I ran this on both the 16s rRNA and meta-abundance datasets to see if the method used for sequencing microbiota abundance yielded different model predictions. Lastly, I plotted the accuracy of each model with a catplot in order to visualize results.

Justification of Methods:

I decided to use 10-fold K Fold cross validation instead of test train split because of the potential to randomly sample from different portions of the dataset. These randomly selected portions were used to both train and test the selected machine learning algorithms 10-fold. Test train split can only yield one split up dataset and the results can vary by chance depending on where the training and testing portions come from. Therefore, K Fold cross validation was the more robust method to use in my opinion. It was not too computationally demanding, so I was able to run this in a reasonable amount of time. If the datasets were substantially larger, I might have resorted to test train split.

I decided to run random forest based on the paper’s results. Random forest, similar to decision trees, is good for taking lots of samples with many features and making classifications based on these methods. I had many patients with thousands of microbiotas, therefore, this supervised learning model was efficient for this task. Logistic regression and SVM can also serve as efficient classification learning algorithms. Logistic regression makes classifications based on maximum likelihood. I did not opt to use gradient descent here, but that may be a nice future direction. SVM tries to draw a hyperplane which results in maximum separation based on samples from each class. Thus, I thought it would be appropriate for separating patients based on ASD status using multiple microbiota features. I ran all three of these algorithms on my data in order to see which one performed the best.

Results:

I did not notice any obvious grouping in the PCA plots for both datasets. The UMAP plots did give some interesting patterns, especially in the 16s rRNA dataset, but the ASD status groups were not completed separated spatially. These results largely agree with the results described in the paper.

For the meta-abundance dataset, the random forest model generated the most accurate results, with 75% accuracy most of the time. The logistic regression and svm models performed with about 65% accuracy on average, about 10% lower than the random forest models. For the random forest model, lower tree counts were not the most accurate, but higher tree counts (over 50) did not improve accuracy. I settled on using 30 trees for analysis.

For the 16s rRNA dataset, again the random forest model performed the best. I also decided to use 30 trees on this dataset after testing for tree number. On this dataset, the random forest model had roughly 89% accuracy, which was usually over 20% higher than the logistic regression and svm models. I found it interesting that on this dataset one machine learning algorithm performed much better than the others. It goes to illustrate the importance of optimization.

Conclusion:

I tested multiple machine learning algorithms on gut microbiota datasets with the random forest model consistently performing the best. Future work may involve further optimization of additional machine learning algorithms or further tweaking the random forest model. Clinically speaking, these machine learning models can be used in hospital settings in order to help psychologists accurately diagnose children with ASD. In the 16s rRNA dataset, the random forest model had 89% accuracy. If further optimization and more samples are supplied, this number may increase. If the psychologist’s diagnosis is in agreement with the machine learning model, there will be greater likelihood of diagnosing ASD in patients.

Analyzing this dataset was personally very helpful for me because I am embarking on a research project which involves analyzing gut microbiota according to different diets. Our lab aims to test the impact of different diets (American diet, Mediterranean diet, supplements…) on the gut microbiome as wells as on the epigenome throughout the lifespan. Getting experience working on a microbiota dataset and with machine learning will make me more proficient when I have to analyze these datasets for my research.