Machine Learning Application on Human Gut Microbiome Data

Yiyan Zhang

Introduction

The human body (especially the gastrointestinal tract) harbors a diverse of microorganisms that includes bacteria, protozoa, archaea, viruses, and fungi (Huseyin et al., 2017). The gut microbiota has notable influence on the host health and the dysbiosis of gut microbiota has been associated with many diseases (Thursby & Juge, 2017). Thus, in this project, I would take the advantage of the metagenomic information from human gut to make the classification between healthy and disease samples with several popular machine learning models.

In this study, I leveraged gut microbiome samples from a publicly available large-scale metagenomic database — curatedMetagenomicData (CMD; version 3.6.0) which contains uniformly processed human microbiome taxonomic abundances data and phenotypic data (Pasolli et al., 2017). I selected samples based on two criteria: the host age is between 18 and 65; no antibiotic use. Samples include healthy individuals, as well individuals with one of the following diseases: Inflammatory Bowel Disease (IBD), Colorectal cancer (CRC), Impaired Glucose Tolerance (IGT), and Type 2 Diabetes (T2D). The final dataset includes 2,815 healthy samples, 768 IBD samples, 368 CRC samples, 199 IGT samples, and 164 T2D samples. I also filtered out those species whose prevalence were lower than 0.1, resulting in 194 species.

In this analysis, I will apply PCA between healthy and each disease group first, as the number of features is relatively large for microbiome datasets. This step can help to get sense whether disease samples are easy to differentiate from healthy samples based on microbiome composition. However, this step will loss the interpretability of the model, and one of the goals of metagenomic study is to find the taxonomy marker associated with different disease. Thus, I will use abundance table as the input for classification. I will employ the state-of-the-art machine learning methods which have been shown as powerful tools to do the classification in previous study (Wang & Liu, 2020), including logistic regression model with further feature selection using LASSO (Logit-Lasso) and Random Forest (RF). For each method, I will use k-fold cross-validation to select the best model and use k-fold cross-validation in higher structure to estimate the model performance. In order to better evaluate the model performance, I used accuracy, F1 score, and area under precision and recall curve (AUPRC). Considering we have much more heathy samples than the disease samples, I figured out F1-score and AUPRC might works better than accuracy under the imbalance data setting. I will also discuss this issue in detail at the results part.

Results

First, I generated PCA plots for each disease samples comparing with healthy samples (Figure 1.). From the plots for four diseases, we can only observe some differentiation in IBD at bottom left part. Thus, we might expect the classification for most disease would be relatively hard to achieve, but IBD might have better performance among all diseases.

Then, I trained logistic regression model with variable selection using LASSO as we have the binary outcome of healthy or disease.

Table of Logit RF importance plot Comparison of Accuracy Comparison of F1 Comparison of AUPRC AUPRC curve

Conclusion However, out study are limited at two-class classification. More pivotal studies that worked on multi-class classification give me further directions on the project.