Evalution of binary classification pipelines and methods for 16S rRNA gene data

Running title: Machine learning methods in microbiome studies

Begüm D. Topçuoğluˆ1, Jenna Wiensˆ2, Patrick D. Schloss[1†]

† To whom correspondence should be addressed: pschloss@umich.edu

1. Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109

2. Department of Computer Science and Engineering, University or Michigan, Ann Arbor, MI 49109

Abstract

Introduction

As gut microbiome field continues to grow, there will be an ever-increasing demand for reproducible machine learning methods to analyze 16S rRNA gene sequence data and to determine association of the microbiome with a continuous or categorical phenotype of interest. The use of machine learning in microbiome literature lack clarity over the learning pipeline which spans the problem formulation, feature selection, feature pre-processing, model learning, and output. There is a need for guidance on how to properly implement good machine learning practices to generate reproducible, robust and actionable models. We also need to generate interpretable models for biomedical researchers to adopt and use regularly. In this study we chose to focus on a binary classification model that predicts if an individual has screen-relevant tumors in their colon or not, to establish a model pipeline.

Colorectal cancer is one of the leading cause of death among cancers in the United States. Each person in the industrialized world has on average a one-in-twenty chance of developing colorectal cancer (CRC) in their lifetime and once diagnosed, more than one-third will not survive 5 years. Colonoscopy as a screening tool is very effective, however it is very invasive, expensive and have a low rate of patient adherence. Therefore, there is a need for improved non-invasive methods to screen individuals. One proposed non-invasive screening tool is using gut microbiome-based biomarkers. Patients with colorectal cancer have different stool community of microbes compared to adults with normal colons. This difference however cannot be explained by a single or a handful of features in the gut microbiome but by many of them in relation to one another. Therefore, machine learning is a great tool to investigate the differences between the gut microbiomes of CRC patients and healthy individuals. Previous studies have shown that human hemoglobin levels and bacterial population abundances in the stool help us predict screen relevant growth in the colon, however the literature for the problem of classification colorectal cancer diagnosis vary greatly, with areas under the receiver operating characteristic curve (AUC) of 0.7-0.8. The variation in classification performance is based in part on differences in the task definition, in part on differences in the study populations, and in part on the evaluation method. The highest reported AUCs were from studies of . . . Additionally, some of the reported results were not obtained from testing on held-out sets. In

this study, we have defined classification pipelines with L2-regularized logistic regression, L1 and L2 linear suppor vector machines (SVM), radial basis function SVM , decision tree, random forest and XGBoost. We evaluated the generalization and prediction performance of these methods. We also compared each method based on their reproducibility, robustness, actionability, interpretibility and susceptibility to overfitting.

Generalisation Perfomance of each model. Is there a maximum threshold of prediction with all these methods? Does an increase in model complexity improve predictibility? Synthesis statement regarding modeling 16S microbiome data

Results and Discussion

Conclusions

Materials and Methods

The data

We obtained stool OTU abundance data and metadata from the Sze et al. (1). The stool OTU abundance data and metadata comes from the Great Lakes- New England Early Detection Research Network study which collected stool samples from eligible individuals. Briefly, eligible patients for this study were aged at least 18 years and willing to collect stool samples. Colonoscopies were performed and fecal samples were collected from participants in four locations: Toronto (ON, Canada), Boston (MA, USA), Houston (TX, USA), and Ann Arbor (MI, USA). Patient diagnoses were determined by colonoscopic examination and histopathological review of any biopsies taken. Patients with an adenoma greater than 1 cm, more than three adenomas of any size, or an adenoma with villous histology were classified as advanced adenoma. Fecal material was used for Fecal Immunological Tests and these tests were used to measure human hemoglobin concentrations. This study had 172 control, 198 adenomas and 120 carcinomas. Of the 198 adenomas, 109 were advanced adenomas.

Figure 1. Generalization and classification performance of modeling methods AUC values of all cross validation and testing performances. The boxplot shows quartiles at the box ends and the statistical median as athe horizontal line in the box. The whiskers show the farthest points that are not outliers. Outliers are data points that are not within 3/2 times the interquartile ranges.