NAME OF THIS STUDY

T

t

Method	Parameter
Logistic Regression	C
L1 SVM Linear Kernel	C
L2 SVM Linear Kernel	C
SVM RBF Kernel	C, gamma
Decision Tree	max_depth, min_samples_split
Random Forest	n_estimators, max_features
XGBoost	n_estimators, colsample_bytree, learning_rate, subsample, max_depth, min_child_weig

Running title: INSERT RUNNING TITLE HERE

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 microbiome sequence read count data and to determine association with a continuous or categorical phenotype of interest.

Colorectal cancer is one of the leading cause of death among cancers in the United States. Early diagnosis increases the chance of survival. However the current diagnostic methods are expensive and invasive. As a less invasive tool, numerous studies use relative abundances of the gut bacteria populations to predict disease progression. Most microbial communities are pretty patchy and the likelihood of a single feature that explains the differences in health is pretty small. It is likely that many biomarkers are needed to account for the patchiness as well as the context dependency of the features.

ML use in microbiome literature is a bit like the wild west with lack of clarity over methods, testing, validation, etc. There is a need for guidance on how to properly implement these different methods. We need to emphasize good machine learning practices and pipelines and discuss the reproducibility, robustness and actionability of models.

We established a non-leaky pipeline. We performed L1 and L2-regularized logistic regression, Linear SVM, Non-Linear SVM, Decision tree, Random forest, XGBoost and Feed Forward Neural Net classification models. We evaluated the classification performance of different machine learning methods. We also want to discuss the reproducibility, robustness, actionability, interpretibility and susceptibility to overfitting of each method.

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

Insert figure legends with the first sentence in bold, for example:

Figure 1. Number of OTUs sampled among bacterial and archaeal 16S rRNA gene sequences for different OTU definitions and level of sequencing effort. Rarefaction curves for different OTU definitions of Bacteria (A) and Archaea (B). Rarefaction curves for the coarse environments in Table 1 for Bacteria (C) and Archaea (D).

References