Machine Learning Manuscript Outline

Introduction

General Context of the work

- As the microbiome field continues to grow, there is an ever-increasing demand for reproducible methods for identifying associations between members of the microbiome and human health.

- Most microbial communities are pretty patchy and the likelihood of a single species that explains the differences in health is pretty small. It is likely that subsets of those communities in relation to one another are needed to account for the patchiness as well as the context dependency of the features.

- Thus, researchers have started to explore the utility of machine learning (ML) techniques.

- Currently, the field's use of machine learning lacks clarity and consistency on

  1. Which methods are used?
  2. Why those methods are used? Can these methods explain if there are linear or non-linear relationships between phenotype and features?
  3. How those methods are used?
  4. Reproducibilty

- ML methods have a trade-off for interpretibility, susceptibility to overfitting and actionability versus prediction performance. It is important to prioritize and define task accordingly.

Narrower research area and statement of its importance

- Among cancers, colorectal cancer is one of the leading causes of death in the US. While colonoscopy is an effective screening tool, it is invasive and as a result has a low rate of patient

adherence. In contrast, microbiome analysis of stool is non-invasive. Machine learning (ML) techniques are used to identify patients with colorectal tumors based on microbiota-associated biomarkers.

- However, the discriminative performance of these models varies greatly, with areas under the receiver operating characteristic curve (AUROC) of 0.7-0.9.

- Previously:

  - Differences in task definition (what is it that we want to predict?)
  - Use of logistic regression without discussion of how it explains non-linearity of microbiome & CRC.
  - Use of random forest only without discussion of interpretibility.
  - Use of random forest without proper pipeline.

- To shed light on how much differences in modeling can affect the results, we performed an empirical analysis comparing several different modeling pipelines.

Summary of appraoch and findings

- Modeling pipelines were established for L2-regularized logistic regression, L1 and L2 support vector machines (SVM) with linear and radial basis function kernels, a decision tree, random forest and XGBoost.

- These methods increase in complexity while decrease in interpretibility.

- We established ML pipeline with held-out test data and performed 100 data-splits to evaluate generalization and predicton performance of the ML method.

- Applied to held-out test data, the mean AUROC varied from 0.68 (std ± 0.04) to 0.82 (std ± 0.04).

- Random Forest had the highest mean AUROC for detecting SRN and was less susceptible to overfitting compared to other methods.

- Despite the lower mean AUROC value, the L1-regularized linear kernel SVM offered the greatest interpretability and stability.

- In terms of computational efficiency, x trained the fastest, while y took the longest.

- We found that cross-validation and testing AUROC could vary by as much as 0.06, highlighting the importance of a separate held-out test set for evaluation.

- Aside from evaluating generalization and classification performance for each of these models, this study established standards for modeling pipelines of microbiome-associated machine learning models.

Results

AUROC results of 7 modeling approaches. (FIT + OTUs)

Comparisons among the 7 modeling approaches. (FIT + OTUs)

- Compare prediction performance, generalization performance and susceptibility to overfitting.

Hyper-parameter tuning budgets and corresponding AUROC values.

- This will show that we have used the right budgets to allow the model to pick the right hyper-parameter.

AUC results of models with just (FIT)

- This could be in supplemental. We want to make sure that using just FIT as a feature does worse than FIT+OTUs.

Discussion

Interpretation of modeling results in terms of reproducibility, robustness, actionability, interpretibility and susceptibility

- What are the metrics to talk about these concepts?

- Emphasize trade-off results we observe in this study.

Consideration of possible weaknesses for each model

- The interactions between the biomarkers may be nonlinear. Obviously, the linear models will not incorporate this because they are linear. Tools like linear models (e.g. metastats, lefse, wilcoxon, etc) are likely worthless. How many OTUs would you find with these methods?

Consideration of possible weaknesses for our approach and chosen dataset

- Why did we use only one dataset? Becuase we are comparing several methods on 1 dataset.

Relationship of results to previous literature and broader implications of this work

Prospects of future progress

Methods

Brief explanation of study design/patient sampling and 16S rRNA gene sequencing/curation

- Baxter et al, 2016

Analysis of data

- What are the features and what are the labels?

  1. Features: Fecal hemoglobin concentration and 16S rRNA gene sequences from stool samples

  2. Labels: 490 patients as having advanced tumors (advanced adenoma or carinoma) or not (non-advanced adenoma or normal colon).

- What is the data (temporal or not)? Assumptions we make when we use a dataset. What will the future data look like?

- Pre-proccessing of the data

- Machine Learning pipeline backbone. How do I split, train, validate and test?

  – A diagram to explain the modeling pipeline.

- Which methods are linear/non-linear? Talk about interpretibility.

- Cross-validation and hyper-parameter tuning methods for each modeling method

- Programming languages and packages/modules utilized

- Statistical methods for comparison of model performance

- Code/data availability