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

Running title: Machine learning methods in microbiome studies

Begüm D. Topçuoğlu[1], Jenna Wiens[2], 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

Advances in sequencing technology and decreasing costs of generating 16S rRNA gene sequences have allowed rapid exploration and taxonomic characterization of the human associated microbiome. Currently, the microbiome field is growing at an unprecedented rate and as a result, there is an ever-increasing demand for reproducible methods for identifying associations between members of the microbiome and human health. However, this is an undertaking as human associated microbial communities are remarkably uneven. It is unlikely that a single species can explain a disease state comprehensively. It is more likely that subsets of those communities, in relation to one another and to their host, account for different disease states. Thus, researchers have started to explore the utility of machine learning (ML) techniques to develop predictive models that use microbiota associated biomarkers to identify healthy and diseased states. However, currently the field's use of ML lacks clarity and consistency on which methods are used, how these methods are implemented and if these methods are robust and reproducible. Moreover, there is a lack of consideration on why a particular ML model is utilized. There is a wide range of complexity and interpretibility in models used in microbiome field (1–5); recently with many that are black boxes which cannot be interpreted by humans (6–8). ML model choice should reflect the goal of the study; whether it is to develop the best predictive model or to understand the ecology of microbiota associated diseases.

One application of machine learning to microbiome data has been to classify patients as having colorectal tumors based on microbiota-associated biomarkers. Colorectal cancer (CRC) is one of the leading causes of death in the US, therefore developing a predictive model from a stool microbiome assay as a non-invasive screening tool is a remarkable innovation. However, the efforts to combine ML methods with relative abundances of bacterial poplations in stool to predict CRC tumors have been afflicted by flawed ML methods and by the lack of discussion on modeling choices. In previous studies, ... There is a need for implementing consistent and transparent machine learning practices to generate reproducible and replicable CRC biomarker models. This study aims to shed light on how much task definition and model choices can affect the results. Here we performed an empirical analysis comparing several different modeling pipelines using the same dataset.

As our dataset, we used a previously published CRC study (3) which had fecal 16S rRNA gene sequences and immunochemical test (FIT) results from 490 patients. We built ML models using 16S abundances and FIT to predict patients with normal or screen relevant neoplasias (SRN) disease status. The study had 261 normal and 229 SRN samples. We established modeling pipelines 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 models increase in complexity while they decrease in interpretability. Our ML pipeline performed 100 data-splits and utilized held-out test data to evaluate generalization and predicton performance of each ML model. The mean AUROC varied from 0.67 (std ± 0.05) to 0.82 (std ± 0.04). Random Forest and XGBoost had the highest mean AUROC for detecting SRN. In terms of computational efficiency, L2 Logistic Regression trained the fastest (0.098 hours, std ± 0.008), while XGBoost took the longest (3.601 hours, std ± 0.524). We found that mean cross-validation and testing AUROC could vary by as much as 0.062, which highlights the importance of a seperate held-out test set for evaluation. Aside from evaluating generalization and classification performances for each of these models, this study established standards for modeling pipelines of microbiome-associated machine learning models.

Results and Discussion

The prediction and generalization performance of classifiers during cross-validation and when applied to the held-out test data.

We evaluated the prediction performance of seven binary classification models when applied to held-out test data using AUROC as our metric. Random Forest and XGBoost had the highest mean AUROC for detecting SRN, 0.818 (std ± 0.036) and 0.814 (std ± 0.038) respectively [Figure 2]. L1 linear SVM and decision tree had significantly lower AUROC values, 0.748 (std ± 0.043) and 0.741 (std ± 0.038) [Figure 2]. However, they had significanly higher performances than L2 linear SVM, RBF SVM and L2 logistic regression which had mean AUROC values of 0.67 (std ± 0.054), 0.683 (std ± 0.048) and 0.677 (std ± 0.053) respectively[Figure 2]. We also evaluated the generalization performance of each classifier by comparing their mean cross-validation AUROC and mean testing AUROC. We found a significant difference for classifiers L2 support vector machines (SVM) with

linear and radial basis function kernels and L2 logistic regression. These differences were 0.062, 0.047 and 0.061, respectively [Figure 2] .

The prediction performance of classifiers for different hyperparameter settings.

The interpretation of classifiers.

Conclusions

Materials and Methods

Data collection

The data used for this analysis are stool bacterial abundances, stool hemoglobin levels and clinical information of the patients recruited by Great Lakes-New England Early Detection Research Network study. These data were obtained from Sze et al (9). The stool samples were provided by recruited adult participants who were undergoing scheduled screening or surveillance colonoscopy. 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). Patients' colonic disease status was defined by colonoscopy with adequate preparation and tissue histopathology of all resected lesions. 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. Study had 172 patients with normal colonoscopies, 198 with adenomas and 120 with carcinomas. Of the 198 adenomas, 109 were identified as advanced adenomas. Stool provided by the patients was used for Fecal Immunological Tests (FIT) which measure human hemoglobin concentrations and for 16S rRNA gene sequencing to measure bacterial population abundances. The bacterial abundance data was generated by Sze et al, by processing 16S rRNA sequences in Mothur (v1.39.3) using the default quality filtering methods, identifying and removing chimeric sequences using VSEARCH and assigning to OTUs at 97% similarity using the OptiClust algorithm (10–12).

Data definitions and pre-processing

The colonic disease status is re-defined as two encompassing classes; Normal or Screen Relevant Neoplasias (SRNs). Normal class includes patients with non-advanced adenomas or normal colons whereas SRN class includes patients with advanced adenomas or carcinomas. Colonic disease status is the label predicted with each classifier. The bacterial abundances and FIT results are the features used to predict colonic disease status. Bacterial abundances are discrete data in the form of Operational Taxonomic Unit (OTU) counts. There are 6920 OTUs for each sample. FIT levels are continuous data present for each sample. Because the data are in different scales, Python programming language v, module scikit-learn v is used to transform features by scaling each feature to a [0-1] range (Table 1) (13).

Learning the Classifier

To train and validate our model, labeled data is randomly split 80/20 into a training set and testing set. Then, seven binary class classifiers, L2 logistic regression, L1 and L2 linear suppor vector machines (SVM), radial basis function SVM, decision tree, random forest and XGBoost, are learned. The training set is used for training purposes and validation of hyperparameter selection, and the test set is used for evaluation purposes. Hyperparameters are selected using 5-fold cross-validation with 100-repeats on the training set. Since the colonic disease status are not uniformly represented in the data, 5-fold splits are stratified to maintain the overall label distribution on the training set.

Classifier Performance

The classification performance of learned classifier is evaluated on the labeled held-out testing set. The optimal classifier with optimal hyperparameters selected in the cross-validation step is used to produce a prediction for the testing set. The performance of this prediction is measured in terms of the sensitivity and specificity, in addition to Area Under the Curve (AUC) metrics. This process of splitting the data, learning a classifier with cross-validation, and testing the classifier is repeated on 100 different splits. In the end cross-validation AUC and testing AUC averaged over the 100

different training/test splits are reported. Hyperparameter budget and performance for each split is also reported.

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 the 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.

References

1. **Zeller G**, **Tap J**, **Voigt AY**, **Sunagawa S**, **Kultima JR**, **Costea PI**, **Amiot A**, **Böhm J**, **Brunetti F**, **Habermann N**, **Hercog R**, **Koch M**, **Luciani A**, **Mende DR**, **Schneider MA**, **Schrotz-King P**, **Tournigand C**, **Tran Van Nhieu J**, **Yamada T**, **Zimmermann J**, **Benes V**, **Kloor M**, **Ulrich CM**, **Knebel Doeberitz M von**, **Sobhani I**, **Bork P**. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol **10**. doi:10.15252/msb.20145645.

2. **Zackular JP**, **Rogers MAM**, **Ruffin MT**, **Schloss PD**. 2014. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res **7**:1112–1121. doi:10.1158/1940-6207.CAPR-14-0129.

3. **Baxter NT**, **Koumpouras CC**, **Rogers MAM**, **Ruffin MT**, **Schloss PD**. 2016. DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. Microbiome **4**. doi:10.1186/s40168-016-0205-y.

4. **Baxter NT**, **Ruffin MT**, **Rogers MAM**, **Schloss PD**. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine **8**:37. doi:10.1186/s13073-016-0290-3.

5. **Pasolli E**, **Truong DT**, **Malik F**, **Waldron L**, **Segata N**. 2016. Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. PLoS Comput Biol **12**. doi:10.1371/journal.pcbi.1004977.

6. **Galkin F**, **Aliper A**, **Putin E**, **Kuznetsov I**, **Gladyshev VN**, **Zhavoronkov A**. 2018. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. bioRxiv. doi:10.1101/507780.

7. **Reiman D**, **Metwally A**, **Dai Y**. 2017. Using convolutional neural networks to explore the microbiome, pp. 4269–4272. *In* 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC).

8. **Fioravanti D**, **Giarratano Y**, **Maggio V**, **Agostinelli C**, **Chierici M**, **Jurman G**, **Furlanello C**.

2017. Phylogenetic convolutional neural networks in metagenomics. arXiv:170902268 [cs, q-bio].

9. **Sze MA**, **Schloss PD**. 2018. Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. mBio **9**:e00630–18. doi:10.1128/mBio.00630-18.

10. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Van Horn DJ**, **Weber CF**. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. ApplEnvironMicrobiol **75**:7537–7541.

11. **Westcott SL**, **Schloss PD**. 2017. OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere **2**. doi:10.1128/mSphereDirect.00073-1

12. **Rognes T**, **Flouri T**, **Nichols B**, **Quince C**, **Mahé F**. 2016. VSEARCH: A versatile open source tool for metagenomics. PeerJ **4**:e2584. doi:10.7717/peerj.2584.

13. **Pedregosa F**, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V**, **Vanderplas J**, **Passos A**, **Cournapeau D**, **Brucher M**, **Perrot M**, **Duchesnay E**. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**:2825–2830.