Evaluation of machine learning 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 of human associated microbiome and its health implications. Currently, the human microbiome field is growing at an unprecedented rate and as a result, there is an increasing demand for methods that identify associations between members of the microbiome and human health. However, this is difficult as human associated microbial communities are remarkably complex and uneven. It is unlikely that a single species can explain a disease. Instead, subsets of those communities, in relation to one another and to their host, account for the differences in health outcomes. Machine learning (ML) methods are effective at recognizing and highlighting patterns in complex microbial datasets. Therefore, researchers have started to explore the utility of ML models that use microbiota associated biomarkers to predict human health and to understand the microbial ecology of diseases such as liver cirrhosis, colorectal cancer, inflammatory bowel diseases (IBD), obesity, and type 2 diabetes (1–11). However, currently the field's use of ML lacks clarity and consistency on which methods are used and how these methods are implemented. Most notably, there are misleading practices of using "leaky" ML pipelines where models are tested on training data or reporting the best outcome of different iterations of cross-validation. Moreover, there is a lack of discussion on why a particular ML model is utilized. Recently, there is a trend towards using more complex ML models such as random forest, extreme gradient boosting and neural networks without a discussion on if and how much model interpretibility is necessary for the study (11–14). The lack of transparency on modeling methodology and model selection negatively impact model reproducibility and reliability. We need to strive toward better machine learning practices by (1) implementing consistent and correct machine learning pipelines; (2) selecting ML models that reflect the goal of the study as it will inform our expectations of model accuracy, complexity, interpretibility and computational efficiency.

To showcase reliable ML methodologies and to shed light on how much ML model selection can affect modeling results, we performed an empirical analysis comparing several different ML models using the same dataset and a robust ML pipeline. We used a previously published colorectal cancer (CRC) study (3) which had fecal 16S rRNA gene sequences from 490 patients. We built ML

models using fecal 16S rRNA gene sequences to predict patients with normal colons or patients with colonic tumors which are called screen relevant neoplasias (SRN). 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 which are ML models that increase in complexity respectively. Our ML pipeline utilized held-out test data to evaluate generalization and prediction performance of each ML model. The mean test AUROC varied from 0.598 (std ± 0.044) to 0.697 (std ± 0.037). Random forest had the highest mean AUROC for detecting SRN. Despite the simplicity, the L2-regularized logistic regression followed random forest in performance. In terms of computational efficiency, L2 logistic regression trained the fastest (0.202 hours, std ± 0.019), while XGBoost took the longest (162.843 hours, std ± 10.018). We found that mean cross-validation and testing AUROC varied only 0.01, which highlights the importance of a seperate held-out test set and consistent preprocessing of the data prior to 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

Model selection and construction

The data has 6920 features (6920 OTUs) and a two-class label that defines the colonic health of the patients (SRN or normal). We established modeling pipelines for binary classification with 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 to emphasize the differences in model accuracy, complexity, interpretibility and computational efficiency due to model selection. We randomly split the data into stratified training/validation and test sets 100 times with a 80/20 proportion [Figure 1]. For each data-split, training/validation set consisted of 393 patients (209 SRN), while the test set was composed of 97 patients (52 SRN). The training/validation data was used for training purposes and validation of hyperparameter selection, and the test set was used for evaluation purposes. Validation of hyperparameter selection was performed by randomly splitting

the training/validation set into stratified training and validation sets 100 times with a 80/20 proportion [Figure 1]. For L2-regularized logistic regression, L1 and L2 support vector machines (SVM) with linear and radial basis function kernels we tuned the **cost** hyperparameter which determines the regularization strength where smaller values specify stronger regularization. For SVM with radial basis function kernel we also tuned **sigma** hyperparameter which determines the reach of a single training instance where for a high value of sigma, the SVM decision boundary will be dependent on the points that are closest to the decision boundary. For the decision tree model, we tuned the **depth of the tree** where deeper the tree, the more splits it has. For random forest, we tuned the **number of features** to consider when looking for the best tree split. For xgboost, we tuned for **learning rate** and the **fraction of samples** to be used for fitting the individual base learners. We validated the prediction performances of each hyperparameter setting (Figure S1].

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 over 100 data-splits using AUROC as classification performance metric. Random Forest had the highest mean AUROC for detecting SRN, 0.697 (std ± 0.037) [Figure 2]. L2 logistic regression and XGBoost had significantly lower AUROC values, 0.676 (std ± 0.042) and 0.678 (std ± 0.04) [Figure 2]. Random forest also had significanly higher performances than L2 linear SVM, RBF SVM and decision tree which had mean AUROC values of 0.671 (std ± 0.045), 0.663 (std ± 0.044), and 0.598 (std ± 0.044), 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 that mean cross-validation and testing AUROC difference for each model was below 0.01.

The complexity and interpretibility of each classifier.

We interpreted the feature weights of L1 and L2 SVM with linear kernel and regression coefficients of L2 logistic regression using the training data. To get a sense of the most important features in linear models, we calculated the mean weights of all the features over the 100 data-splits for each model. [Figure 4]. In the three linear models, FIT was the feature that drove the detection of SRNs,

followed by eight OTUs which belong to family *Lachnospiraceae* and *Ruminococcaceae* (four with positive signs; Otu01239, Otu00742, Otu00008, Otu00659, and four with negative signs; Otu00015, Otu000150, Otu00609, Otu00629). To get the importance of features in a decision tree, we used gini importance which allows us to use the improvement in the split-criterion is the importance measure attributed to the splitting variable

The computational efficiency of each classifier.

Linear models trained faster than non-linear models. L2 logistic Rregression and L1 and L2 SVM with linear kernel had training times of 13.2 min, (std ± 1.8), 563.4 min, (std ± 46.8) and 12.6 min, (std ± 1.8) respectively. Whereas, SVM with radial basis function kernel, decision tree, random forest and xgboost had training times of 101.32 hrs, (std ± 10.02), 162.84 hrs, (std ± 3.99), 0.2 hrs, (std ± 0.02) and 64.71 hrs, (std ± 9.85), respectively [Figure 4].

Conclusions

Materials and Methods

The classification models were built to classify normal versus adenoma, normal versus carcinoma, and high versus low SCFA concentrations. The regression models were built to classify the SCFA concentrations of acetate, butyrate, and propionate regardless of disease status. The featThe bacterial abundances 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, features are transformd by scaling each feature to a [0-1] range

To evaluate the model performances, we randomly split the data into training/validation and test sets 100 times with a 80/20 proportion. Data-splits were stratified for diagnosis or SCFA concentration levels in classification models. For each data-split, training/validation set consisted of x patients (x carcinoma, y adenoma, z normal), while the test set was composed of x patients (x carcinoma, y adenoma, z normal). The training/validation data was used for training purposes and validation

of hyperparameter selection, and the test set was used for evaluation purposes. Validation of hyperparameter selection was performed by randomly splitting the training/validation set into stratified training and validation sets 100 times with a 80/20 proportion [Figure 1].

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 (15). 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 (16–18).

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, features are transformd by scaling each feature to a [0-1] range (Table 1).

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. **Hale VL**, **Chen J**, **Johnson S**, **Harrington SC**, **Yab TC**, **Smyrk TC**, **Nelson H**, **Boardman LA**, **Druliner BR**, **Levin TR**, **Rex DK**, **Ahnen DJ**, **Lance P**, **Ahlquist DA**, **Chia N**. 2017. Shifts in the fecal microbiota associated with adenomatous polyps. Cancer Epidemiol Biomarkers Prev **26**:85–94. doi:10.1158/1055-9965.EPI-16-0337.

6. **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.

7. **Sze MA**, **Schloss PD**. 2016. Looking for a signal in the noise: Revisiting obesity and the microbiome. mBio **7**. doi:10.1128/mBio.01018-16.

8. **Walters WA**, **Xu Z**, **Knight R**. 2014. Meta-analyses of human gut microbes associated with

obesity and IBD. FEBS Lett **588**:4223–4233. doi:10.1016/j.febslet.2014.09.039.

9. **Vázquez-Baeza Y**, **Gonzalez A**, **Xu ZZ**, **Washburne A**, **Herfarth HH**, **Sartor RB**, **Knight R**. 2018. Guiding longitudinal sampling in IBD cohorts. Gut **67**:1743–1745. doi:10.1136/gutjnl-2017-315352.

10. **Qin N**, **Yang F**, **Li A**, **Prifti E**, **Chen Y**, **Shao L**, **Guo J**, **Le Chatelier E**, **Yao J**, **Wu L**, **Zhou J**, **Ni S**, **Liu L**, **Pons N**, **Batto JM**, **Kennedy SP**, **Leonard P**, **Yuan C**, **Ding W**, **Chen Y**, **Hu X**, **Zheng B**, **Qian G**, **Xu W**, **Ehrlich SD**, **Zheng S**, **Li L**. 2014. Alterations of the human gut microbiome in liver cirrhosis. Nature **513**:59–64. doi:10.1038/nature13568.

11. **Geman O**, **Chiuchisan I**, **Covasa M**, **Doloc C**, **Milici M-R**, **Milici L-D**. 2018. Deep learning tools for human microbiome big data, pp. 265–275. *In* Balas, VE, Jain, LC, Balas, MM (eds.), Soft computing applications. Springer International Publishing.

12. **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.

13. **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).

14. **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].

15. **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.

16. **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.

17. **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

18. **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.