Evaluation of machine learning methods that identify colorectal lesions with microbiota-associated biomarkers

Begüm D. Topçuoğlu, Jenna Wiens, Mack Ruffin, Patrick D. Schloss

As gut microbiome field continues to grow, there is an ever-increasing demand for reproducible machine learning methods to determine associations between the microbiome and a phenotype of interest. Currently, the use of machine learning in microbiome research lacks clarity and consistency over the modeling pipeline (training, validation and testing steps). There is a need for guidance on how to implement good machine learning practices to generate reproducible and robust models.

Recently, there has been growing interest in using machine learning to identify colorectal lesions that are precursors of coloractal cancer, with microbiota-associated biomarkers. Colorectal cancer is one of the leading cause of death among cancers in the United States. Colonoscopy as a screening tool is effective, however it is invasive, expensive and have a low rate of patient adherence. Previous studies have shown that bacterial population abundances in the stool can predict screen relevant lesions in the colon and can be used as a non-invasive screening tool. However, the prediction performance of these models vary greatly, with areas under the receiver operating characteristic curve (AUC) of 0.7-0.9 (1–4). The variation in prediction performance is based in part on differences in the study populations, and in part on the differences in modeling pipelines.

In this study, hemoglobin levels and 16S rRNA gene sequences in the stool were used to identify colorectal lesions of 490 patients. The colorectal disease stage was defined as showing screen-relevant lesions or not. Modeling pipelines were established for L2-regularized Logistic Regression, L1 and L2 Linear Support Vector Machines (SVM), Radial Basis Function SVM, Decision Tree, Random Forest and XGBoost binary classification models. The mean AUCs of these models were 0.68 ± 0.04, 0.76 ± 0.05, 0.68 ± 0.05, 0.69 ± 0.05, 0.71 ± 0.04, 0.82 ± 0.04, and 0.76 ± 0.04, respectively. Tree-based methods, namely Decision Tree, Random Forest and XGBoost were less susceptible to overfitting and in general had higher sensitivity and specificity for colonic screen-relevant lesions. Aside from evaluating generalization and classification performance of each model, this study established standards for modeling pipeline of the microbiome-associated machine learning models.

References

1. **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.

2. **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.

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. **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.