

Inflammatory bowel disease prediction based on metagenomics data

Andrea Mihajlović^{1*}, Katarina Mladenović², Tatjana Lončar-Turukalo², Sanja Brdar¹

¹ BioSense Institute, University of Novi Sad, dr Zorana Đinđića 16,
21000 Novi Sad, Serbia

² Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6,
21000 Novi Sad, Serbia

Abstract

Inflammatory bowel disease (IBD) is a genetic disease manifested under certain external influences. Traditional disease models aim to find a single pathogen which causes disease. However, many studies have revealed that no single pathogen causes IBD. Machine learning algorithms applied on microbiome data have huge potential in uncovering patterns and aiding diagnosis of diseases including IBD. In this study we investigate microbiome variations in stool samples, in an attempt to evaluate performance of classification algorithms in identifying IBD state.

Dataset used in this study (available from Integrative Human Microbiome Project) contains 429 samples from 27 healthy subjects, and 1209 samples from 103 IBD subjects. Microbes are grouped into 1479 operational taxonomic units (OTUs) and in this form used in our analysis. In preprocessing, data was log2-transformed. From many investigated algorithms, a random forest (RF) classifier was selected for detailed evaluation in a binary (IBD, nonIBD) classification task. The class imbalance was approached using balanced RF (BRF) which under-samples the majority class in a bootstrap process. Parameter searching was conducted using a cross-validation approach. Initial set of parameters was created at random, further evaluated through grid search and fine tuned in the vicinity of best performing parameters. Dimensionality was reduced by searching for the smallest feature subset which preserves the performance. Experiments included hand-picked taxa and/or selected k best scoring features. Training was performed for each model in 100 iterations with 10-fold cross-validation, which ensured comprehensive evaluation. Upon sample-wise binary classification, subjects were labelled as IBD based on average decision probability of their samples, by varying different thresholds. Change in classification performance as a function of the employed threshold was noted.

Best model comprised 150 trees with a maximum depth of 15, using entropy for node splitting. With the average $F1$ score of 94% our study confirms the strong connection of IBD and gastrointestinal microbiome. Retraining model using 100 most important features showed minor decrease in $F1$ score of 1% and exclusion of all strains and organisms other than bacteria showed no decrease. Further research efforts should focus on gathering more data and improved model explainability in predicting IBD state.

Keywords:

microbiome, OTU table, machine learning, features selection

*Corresponding author, e-mail: andrea.mihajlovic@biosense.rs