Combining machine learning and network analysis pipelines: the case of microbiome and metabolomics data in colorectal cancer

Eleni NTZIONI[a,1] and Ioanna CHOUVARDA[a]

[a] *Laboratory of Computing, Medical Informatics and Biomedical Imaging Technologies, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece*

Abstract. This study analyzes samples of intestinal microbiome and metabolites, from healthy individuals (HE) and patients with adenomas (AD) or colorectal carcinomas (CRC). A network analysis (NetAn) method was applied to the data, to identify the metabolites and microbial genera associated with the 3 classes and then 7 classification models were used. The models were initially trained with classic feature selection vs features resulting from NetAn. The distinction of HE and AD is successful, while CRC distinction presented lower success.

Keywords. Machine learning, network analysis, metabolomics, microbiome

1. Introduction

The primary aim of this study is to examine the correlations of specific metabolites and microbes with the occurrence of cancer via NetAn methods, so as to select informative features for a predictive model. Consequently, predictive models are explored using machine learning methods. The results obtained with a classic feature selection method are compared with the features selected by the NetAn method.

2. Methodology

The analysis of microbiome (MI) and metabolomic (ME) data in the present work is based on the study carried out by Kim et al. [1], which included 102 HE samples, 102 AD samples and 36 CRC samples, normalized and openly available. ME samples include data for 462 metabolites and MI samples include data for 70 genera. NetAn was performed with the R package NetCoMi [2], employing Pearson correlation after normalization and zero-handling. NetAn was applied to each of the 3 classes separately, and then network comparisons of the classes took place. Based on this process, 30 features were selected for MI and 65 for ME data. The machine learning methodology used for prediction is based on the work of Topçuoğlu et al. [3]. The open-source script was adjusted as needed and 7 multi-class classification models were applied to the 2

[1] Corresponding Author, Ioanna Chouvarda; E-mail: ioannach@auth.gr

datasets (MI/ME). Initially, the 7 models were applied to the complete dataset and permutation analysis was used for feature selection. In a second iteration, the 7 models were applied to the features selected via NetAn, and in a third iteration they were applied to the subset produced with permutation analysis. All methods were compared to determine whether the feature selection with NetAn produced better prediction results.

3. Results

The comparisons indicated that by a large majority, the models produced better results in the analysis of the subsets compared to the complete set of features. The distinction of HE and AD is successful (average values MI: HE~77%, AD~78% and ME: HE~80%, AD~81%), but the distinction of CRC presented lower success regardless of the model used, the dataset or the subset applied. In both the MI and ME datasets, the prediction using features from NetAn showed better results in distinguishing HE from AD or CRC.

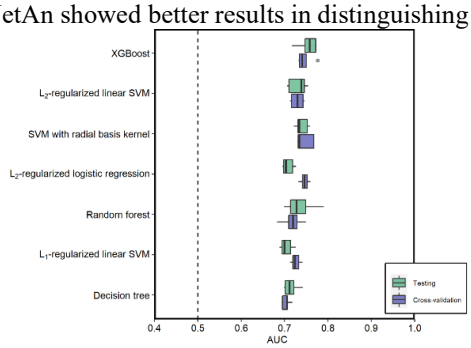

Figure 1. AUC values of the 7 models for the ME subset with the 30 features from NetAn.

4. Conclusions

The results show that the fewer samples of the carcinoma class had a strong negative effect on the performance of the models. For a set of samples where the number of samples will be equal to the other classes, we consider that the features selected with NetAn may lead to better predictions than a classic feature selection method.

Acknowledgement

This work has received funding from H2020 programme under grant agreement 952179.

References

[1] Kim M, Vogtmann E, Ahlquist D, Devens M, Kisiel J, Taylor W et al. Fecal Metabolomic Signatures in Colorectal Adenoma Patients Are Associated with Gut Microbiota and Early Events of Colorectal Cancer Pathogenesis. mBio. 2020;11(1).

[2] Peschel S, Müller C, von Mutius E, Boulesteix A, Depner M. NetCoMi: network construction and comparison for microbiome data in R. Briefings in Bioinformatics. 2020;22(4).

[3] Topçuoğlu B, Lesniak N, Ruffin M, Wiens J, Schloss P. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. mBio. 2020;11(3).