Handling the Challenges of Microbiome Data through Supervised Autoencoders for the Non-invasive Disease Diagnosis

Veronica Buttaro[1,*], Michelangelo Ceci[1,2,3] and Gianvito Pio[1,2]

[1]Dept. of Computer Science, University of Bari Aldo Moro, Bari, Italy
[2]Data Science Lab, CINI Consortium, Rome, Italy
[3]Dept. of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia

Keywords
Supervised Autoencoders, Microbiome, Autism Spectrum Disorder, Colorectal Cancer

1. Introduction

The analysis of the human microbiome is very important for the maintenance of human health and the possible early diagnosis of diseases. The so-called *dysbiosis* of the microbiome, that is the disruption of the state of equilibrium between "good" and "bad" bacteria, can trigger several disease conditions and disturbs [1, 2]. For example, through the microbiome-gut-brain axis, there is evidence of a correlation between alterations in the microbiome and neurodevelopmental conditions, such as Autism Spectrum Disorder (ASD) [3, 4]. Another relevant disease that has shown to be correlated with the microbiome is the Colorectal Cancer (CRC): several studies identified changes in the composition of the gut microbiome associated with CRC progression [5, 4]. In this context, the analysis of the microbiome would represent a non-invasive solution, that would complement other approaches, such as the FIT analysis [6].

The adoption of machine learning techniques can nowadays accelerate the construction of novel predictive models for the early and non-invasive diagnosis of diseases and disturbs from microbiome data. However, while the adoption of these algorithms could facilitate the identification of novel biomarkers, there are numerous challenges to be faced when working with microbiome data [7]. Among the main challenges, it is worth mentioning the high dimensionality, the data sparsity, the high variability and heterogeneity, and data compositionality.

Following this line of research, we propose a novel machine learning approach to analyze human microbiome data to build a predictive model for a non-invasive diagnosis of ASD and CRC, that is able to handle such challenges.

2. The proposed method

Microbiome data are collections of counts for a wide range of Operational Taxonomic Units (OTUs), namely, counts at a given level of detail (genus, species, families, etc.) observed in fecal samples. They are usually expressed as relative abundances, thus introducing data compositionality. In order to handle the high variability and the issues raised by data compositionality, we rely on the pseudo Centered Log Ratio normalization (CLR) [8].

On the other hand, we handle the high dimensionality and sparsity of microbiome data through a specific kind of neural network, based on Autoencoders (AEs). AEs usually exhibit a funnel-shaped structure, that aims to learn a compressed representation, such that data provided to the input layer is

Discovery Science - Late Breaking Contributions 2024
*Corresponding author.
veronica.buttaro@uniba.it (V. Buttaro); michelangelo.ceci@uniba.it (M. Ceci); gianvito.pio@uniba.it (G. Pio)
0009-0007-9910-8538 (V. Buttaro); 0000-0002-6690-7583 (M. Ceci); 0000-0003-2520-3616 (G. Pio)

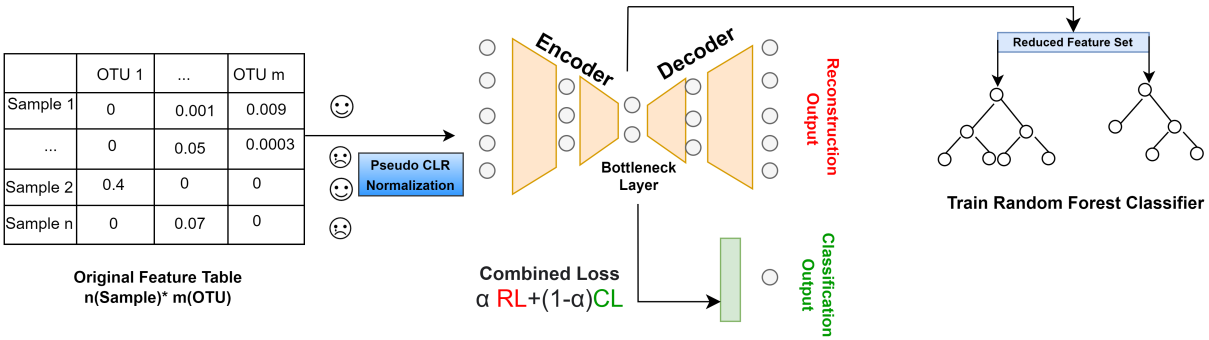

Figure 1: The proposed approach for the prediction of disease/disorder conditions from gut microbiome data, based on a Supervised Autoencoder (SAE).

accurately reconstructed in the output layer. However, standard AEs tend to discard the actual label (i.e., diagnosis, in our case) of training instances (i.e., individuals, in our case) while learning the compressed space. The novelty of our approach with respect to other works [9, 10] consists in the exploitation of the actual diagnosis of individuals during the training of a supervised autoencoder (SAE), that is performed by simultaneously optimizing a reconstruction loss (RL) and a classification loss (CL). We measure the RL through the Mean Squared Error (MSE) computed between the input and the reconstructed output, while we measure the CL through the Binary Cross Entropy. The combined loss is then computed as the linear combination of these losses, where $\alpha \in [0; 1]$ (resp., $1 - \alpha$) represents the weight provided to the reconstruction loss (resp., classification loss).

The bottleneck layer of the trained autoencoder is then used as input to learn a classification model based on Random Forests (RF). A figure depicting the proposed method is shown in Fig. 1.

3. Results and discussion

We focus our experiments on the diagnosis performed on two public datasets about CRC and ASD . All the experiments were conducted using a stratified 5-fold cross validation, collecting precision, recall and F1-score. For comparison, we considered the results obtained: *i)* without reducing the data dimensionality; *ii)* by reducing the input space through PCA; *iii)* with a standard (unsupervised) AE. We also experimented with different values of α to assess its influence on the results of our SAE.

For both datasets, the proposed architecture proved to be able to consider the label of training instances (i.e., known diagnosis), during the identification of the optimal compression of the data. Indeed, the proposed SAE led to better results in comparison with classifiers learned from the original features, as well as to those learned after the application of the PCA or the standard AE. Specifically, we observed an improvement in terms of macro F1 score of 5.6% (with $\alpha = 0.5$) on the ASD dataset, and of 32% on the CRC dataset (with $\alpha = 0.7$) over the results obtained from the original features. While other values of α did not provide the same improvement, the obtained results were almost always higher than those achieved by training the classifier from the original features.

In the future, we will integrate an explainability component to identify which bacteria mostly contributed to making the diagnosis. We will also extend our method to work in a semi-supervised setting to also exploit unlabeled instances.

Acknowledgments

This work was partially supported by the European Union - NextGenerationEU through the Italian Ministry of University and Research, Projects: “FAIR - Future AI Research (PE00000013)”, Spoke 6 - Symbiotic AI; PRIN 2022 “BA-PHERD: Big Data Analytics Pipeline for the Identification of Heterogeneous Extracellular non-coding RNAs as Disease Biomarkers", grant n. 2022XABBMA, CUP:

H53D23003690006; CN3 RNA - “National Center for Gene Therapy and Drugs based on RNA Technology", CUP: H93C22000430007.

References

[1] S. Askarova, B. Umbayev, A.-R. Masoud, A. Kaiyrlykyzy, Y. Safarova, A. Tsoy, F. Olzhayev, A. Kushugulova, The links between the gut microbiome, aging, modern lifestyle and Alzheimer’s disease, Frontiers in cellular and infection microbiology 10 (2020) 104.

[2] Y. Chen, J. Zhou, L. Wang, Role and mechanism of gut microbiota in human disease, Frontiers in Cellular and Infection Microbiology 11 (2021) 625913.

[3] Z. Dan, X. Mao, Q. Liu, M. Guo, et al., Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder, Gut microbes 11 (2020) 1246–1267.

[4] A. Simeon, M. Radovanović, T. Lončar-Turukalo, M. Ceci, S. Brdar, G. Pio, Multi-class boosting for the analysis of multiple incomplete views on microbiome data, BMC bioinformatics 25 (2024) 188.

[5] P. Novielli, D. Romano, M. Magarelli, P. D. Bitonto, D. Diacono, A. Chiatante, G. Lopalco, D. Sabella, V. Venerito, P. Filannino, et al., Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification, Frontiers in Microbiology 15 (2024) 1348974.

[6] N. T. Baxter, M. T. Ruffin, M. A. Rogers, P. D. Schloss, Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions, Genome medicine 8 (2016) 1–10.

[7] I. Moreno-Indias, L. Lahti, M. Nedyalkova, I. Elbere, G. Roshchupkin, et al., Statistical and machine learning techniques in human microbiome studies: contemporary challenges and solutions, Frontiers in microbiology 12 (2021) 635781.

[8] D. Swift, K. Cresswell, R. Johnson, S. Stilianoudakis, X. Wei, A review of normalization and differential abundance methods for microbiome counts data, WIREs Computational Statistics 15 (2023) e1586.

[9] M. Oh, L. Zhang, Deepgeni: Deep generalized interpretable autoencoder elucidates gut microbiota for better cancer immunotherapy, Scientific Reports 13 (2023) 4599.

[10] D. Reiman, Y. Dai, Using autoencoders for predicting latent microbiome community shifts responding to dietary changes, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 1884–1891.

A. Online Resources

The source code and data are publicly available, as follows:

- GitHub (source code): https://github.com/VeronicaButtaro98/SAE-microbiome
- CRC dataset: https://hackmd.io/@laurichi13/rJt3ewZut
- ASD dataset: https://www.kaggle.com/datasets/antaresnyc/human-gut-microbiome-with-asd