CS6024 Mini-Project Proposal

Metagenomics for disease identification

Bersilin C (CS20B013), Yuva Sai Tej (CS20B002)

Jan-May 2024

1 Introduction

A diverse array of microbiomes exists independently or in symbiosis throughout our surroundings. Identifying the various microbes within a microbiome is crucial for comprehending, altering, or addressing the microbiome. Analyzing the soil microbiome allowed us to identify biomarkers for plant diseases and has facilitated the development of medications for prevention and treatment. The **human gut microbiota** represents a complex consortium of microorganisms cohabiting symbiotically within the human gastrointestinal tract. The varying prevalence of numerous species within it aids in discerning metabolic disparities between healthy individuals and those affected.

Metagenomic shotgun sequencing is a method used to analyze large-scale microbiomes. This technique provides genomic data for all microorganisms within the microbiota. MetaPhlAn[1] is a computational tool designed for species-level microbial profiling from metagenomic shotgun sequencing data for different bacteria, archaea, eukaryotes, and viruses. With tools like MetaPhlAn, we can extract species-level abundance information. This information can be utilized to identify biomarkers and develop models for identifying diseases. In this work, we aim to enhance the accuracy of MetaML [2], a set of analyses and models. We plan to achieve this by incorporating newer machine learning and deep learning techniques and rigorously testing their effectiveness. This approach could aid in pinpointing the underlying causes of diseases or uncovering potential medications to alleviate their effects.

2 Paper Critique

The MetaML [2] paper explores the microbial profiling data derived from processing metagenomic data using the MetaPhlAn [1] tool, developed by the same research group (Segata Lab). The dataset encompasses **six metagenomic datasets** linked to **five diseases**: liver cirrhosis, colorectal cancer, inflammatory bowel disease (IBD), obesity, and type 2 diabetes (drawn from two independent studies). They employ a technique known as **cross-study validation** to evaluate various models' capability to learn the distribution of diseased and control samples across different subsets of the dataset. Table(1) provides pertinent details about the datasets utilized.

2.1 Cross-study Validation

In a typical machine learning model training scenario, we partition the training data into separate sets for training and validation to assess the model's performance at different stages. Here, the trained model is validated on datasets from different studies in this study. This approach allows observation of the distribution of species abundance across the entirety of the data, irrespective of the study origin of the samples. The aim is to generalize the model's performance beyond specific

Dataset name	Body site	Disease	#case samples	#control sample
Cirrhosis	Gut	Liver Cirrhosis	118	114
Colorectal	Gut	Colorectal Cancer	48	73
HMP	Several	None	-	981
IBD	Gut	Inflammatory Bowel Diseases	25	85
Obesity	Gut	Obesity	164	89
Skin	Skin	None	-	287
T2D	Gut	Type 2 Diabetes	170	174
WT2D	Gut	Type 2 Diabetes	53	43

Table 1: Information about the different datasets [2]

study distributions and uncover patterns that lead to improved performance across datasets. Additionally, they conduct tests by training a model with all datasets except one excluded from the training data and validating the samples from that dataset.

2.2 Possible Improvements

We believe that the following improvements could be made to the existing work.

1. We examined the prospective future work outlined in the paper, which suggests exploring machine learning techniques beyond SVMs, Random Forests, Lasso, and Elastic Nets.
2. The paper solely focuses on training binary classification models. Enhancements could involve experimenting with other supervised methods such as multi-class classification and unsupervised methods like clustering.

3 Problem Statement

We intend to utilize the dataset provided in the paper titled *Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights* [2] (Dataset Link). We plan to,

1. Reproduce the findings outlined in the paper and explore additional **Exploratory Data Analysis (EDA)** methods.
2. Utilize machine learning techniques like Bayesian Logistic Regression, Ensemble methods (boosting, bagging), Neural Networks, etc., to perform **Binary Classification of diseases** within individual datasets. Evaluation using metrics like accuracy, precision, recall, and F1 scores.
3. Perform **Multi-class Classification** on the entire dataset to assess its overall quality and examine the presence of a learnable distribution in the relative abundances of species. Evaluate using metrics such as Top-k accuracy and precision scores.
4. Investigate the efficacy of **Clustering techniques** on the metagenomic data for their capability to distinguish between samples of different diseases.

4 Potential Challenges

After a preliminary examination of the dataset, we anticipate encountering several challenges, including but not limited to:

1. The dataset has high dimensionality, with each sample containing approximately 3300 features. Directly using all these features for model training may hinder the model's learning

ability and increase training time. Therefore, we intend to conduct initial data cleaning and analysis to compress certain columns and eliminate those with very low correlation with the target.

2. The dataset exhibits sparsity, with numerous zeroes present in its data. This issue is closely linked with the dimensionality problem. We may have to consider methods not impacted by the abundance of zeroes in the dataset.

3. While we can build models for analyzing datasets without domain knowledge, we do not understand certain biological terms used in the data and paper. We aim to learn some of these terms by the end of the project and be able to provide some meaningful insights from the output of the models.

5 Tentative Timeline

The tentative timeline for the project is as follows:

1. Dataset cleaning and initial exploratory data analysis (EDA). [Week 1, 2]
   - Explore methods to address dimensionality and sparsity issues.
   - Prepare the dataset for training and testing, including handling NaN values through zeroing or local averaging techniques.
2. Replicate the charts and findings outlined in the paper. [Week 1, 2]
   - Implement the models used in the paper and understand the biological interpretations.
   - Generate graphs similar to the paper and improve on them.
3. Model fitting and assessment. [Week 2, 3]
   - Experiment with various models specified in the problem statement, aiming for reproducible outcomes.
   - Conduct hyperparameter tuning if required and utilize diverse metrics to assess performance.
4. Results analysis. [Week 4]
   - Arrange and interpret various scores to extract various biological insights, like specific biomarkers, types of clusters, etc.
   - Prepare presentation and report.

References

[1] Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods* (2015).

[2] Waldron, L., Edoardo, T., Pasolli, D. T., Malik, F., and Segata, N. Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLOS Computational Biology 12*, 7 (07 2016), 1–26.