Scripps Research

Scripps Research Translational Institute

Creating Fake Human Gut Microbiome Data? — a Negotiation Toward an Imperfect Data World to Improve Data Quality

By Guoxuan Xu
Mentors: Bailin Zhang, Chunlei Wu
The Wu Lab

Background

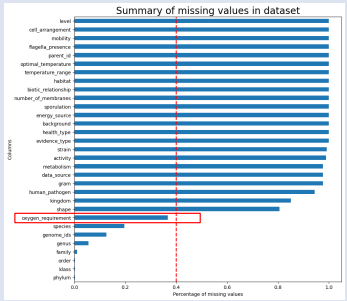
- Microbiome within human plays a critical role in health by:
  - Residing in diverse anatomical locations
  - Producing diverse metabolites interacting with human immune system that
    - Maintaining bodily homeostasis to keep individuals healthy
    - Contributing to the development of pathological diseases
- Despite the growing availability of microbiome datasets, most suffer from low quality:
  - Stored in heterogeneous, unstructured formats
  - Had formatting issue across entities of a same feature
  - Characterized by high sparsity and extensive missing data
- These issues cause:
  - Analytical challenges
  - Erroneous conclusions
  - Unsuitability for Machine Learning Applications
  - **EVERYTHING GO WASTED !!**

What is Data Imputation?

- A prevalent method in data science
- Using available data to estimate missing values
  - Imputation methods can vary depending on datasets
  - Aim: preserve the dataset's inherent patterns
- Enable effective analysis on large scale data
- Potential relationships can be pinpointed with sufficient imputed data

Objective

- **Improving human gut microbiome data quality by imputing missing values**
- Data obtained from the Human Microbial Metabolome Database maintained by the Metabolomics Innovation Center (NIH-supported)



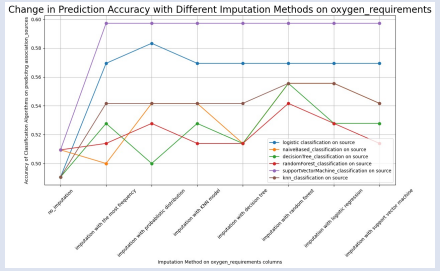
  - Focusing on features with less than 40% missing data to avoid high bias
- The type of oxygen requirements for the microbes will be imputed
  - Oxygen requirements determine microorganisms' anatomical locations
  - Different anatomical locations will result in different metabolic strategies
- Other features such as taxonomic rank of the microorganisms including super kingdom, genus, phylum, and class are used to impute oxygen requirements

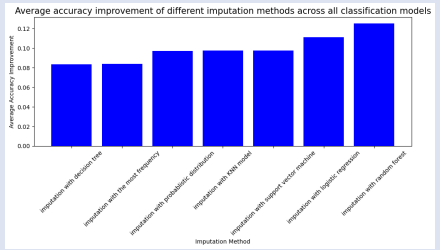
Method

- Step #1: Conducting hypothesis tests exploring the dependency of oxygen requirement on other columns
  - Using total variation distance as test statistic

Method Cont.

  - Comparing distributions of columns with and without the present of oxygen requirements
  - Concluding that the missingness of oxygen requirements depends on features: genus, phylum, class, and super kingdom
- Step #2: Running various imputation methods for oxygen requirements
  - In total, 7 imputation methods
  - Ranging from simple imputation based on the most frequent value to complex methods like probabilistic imputation and random forest imputation
- Step #3: Evaluating the performance of imputation methods
  - Using imputed data to predict the anatomical locations of the microbes



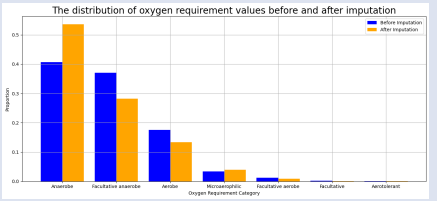


  - Calculating the average improvement in accuracy of each imputation method across different classification models

Method Cont.

  - Random forest algorithm has the highest average improvement of 12.53%
- Step #4: Conducting final imputation using random forest algorithm with hyper tunning

Results



- Imputing all missing oxygen requirements
- Preserving the distribution of non missing oxygen requirements data
- Final Model with an accuracy of 46.67% on predicting anatomical locations of the microbes

Discussion

- Observed reduction in prediction accuracy with final imputed data compare to Step #3.
  - The final imputation applied to entire dataset (2174 rows)
  - Performance evaluation conducted on a subset
- Future Work:
  - Manually check data imputation quality
  - Sequential imputation: imputing multiple columns