What does your gut tell you? Disease detection from Metagenomic data

Bersilin & Yuva Sai Tej

Indian Institute of Technology, Madras

CS20B013 & CS20B002

May 2, 2024

Overview

1. Introduction
   - Metagenomic data
   - Human gut microbiota
   - Related work
   - Problem statement
   - Dataset
2. Methods and Inference
   - Single distribution learning
     - Top features / Potential biomarkers
   - Mixture of distributions learning
     - Feature extraction
   - Multi-class classification

Metagenomic data

- **Metagenomics** - investigate the composition and roles of microbial communities through genomic material analysis
- Obtained directly from the environment, bypassing the time-consuming culturing process.
- More comprehensive depiction compared to phylogenetic surveys
- Target microbiomes - human gut, skin, oral cavity, soil, water, animal tissues
- Shotgun sequencing genomic data $\rightarrow$ species-level abundance information [MetaPhlAn2, 2015]

Human gut microbiota

- Complex ecosystem of microorganisms like bacteria, viruses, fungi, archaea, etc
- Functions: nutrient metabolism, immune modulation, protection against pathogens
- Gastrointestinal activity - clinical applications: disease diagnosis, monitoring
- **Dysbiosis** - imbalance or disruption in the composition and function of the gut microbiota - associated with IBD, IBS, obesity, diabetes and even neurological conditions like Parkinson's disease and depression
- Information on microbiota: biomarker discovery, metagenomic profiling, diagnostic signatures, predictive models, therapeutic targets

Related work

- Paper: Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights: (`MetAML`)
- `MetAML` introduces and analyses classification models for disease prediction from metagenomic data.
- Specifically, six metagenomic datasets linked to five diseases
  - Liver cirrhosis
  - Colorectal cancer
  - Inflammatory bowel disease (IBD)
  - Obesity
  - Type 2 diabetes (drawn from two independent studies)
- Reference: [MetaML, 2016]

Related work contd.

- `MetAML` uses cross-study validation
  - train with one study (dataset) to learn the distribution of healthy and unhealthy samples
  - test on some other dataset to classify samples
- Steps involved in cross-study validation
  - data collection & data preprocessing
  - partitioning datasets
  - model training & model evaluation
  - cross-study & comparison

Related work contd.

- Dataset used in `MetAML`

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: Information about the different datasets used in `MetAML`

Problem statement

- Identification of diseases from species-level abundance information using multi-layered perceptrons (`MLP`) / feed-forward neural networks (`FFNN`)
- Multiple levels to the problem:
  - single distribution/disease identification
  - combined binary classification
  - combined multi-class classification
- Finding discriminatory species/biomarkers

Dataset

- Collection of processed metagenomic sequencing data - species level abundance information
- Total of 3600+ samples, each with 210 metadata features and 3300 abundance features
- Collection of data from 17 sequencing datasets: let's call them subsets (of the dataset)
- Some contain only healthy samples, and some contain multiple different but similar disease samples
- Some terms:
  - species-level abundance: the relative occurrence of a particular microorganism in the sample
  - positive and negative: the presence of the disease

Table of datasets

Dataset name	Positive	Negative	Total	Classes
Zeller Fecal Colorectal Cancer	87	47	134	4
WT2D	102	43	145	3
Vertical Transmission Pilot	0	17	17	1
T2D 1 (Long)	35	38	73	2
T2D 2 (Short)	135	155	290	3
Obesity	16	35	51	4
Human Skin	0	291	291	1
Quin Gut Liver Cirrhosis	118	114	232	2
Psoriasis	36	34	70	2
IBD 1	148	234	382	4
IBD 2	25	85	110	3
E-Coli Stec2 1	9	0	9	1
E-Coli Stec2 2	43	1	44	2
Hmpii	0	219	219	1
Hmp	0	762	762	1
Chatelier Gut Obesity	164	114	278	3
Candela Africa	0	38	38	1

Table: Information about the dataset

Single distribution learning

- Modified the dataset subsets to two classes - positive and negative
- Identified species with the maximum absolute difference in relative abundance for each subset
- Implemented an `FFNN` using `PyTorch` in `Python` to classify the samples
- Used `CrossEntropyLoss` to calculate loss and `Adam` optimizer to modify weights
- Performed a simple hyperparameter tuning: layers, learning rate, batch size and epochs (180 models for each dataset)

Top features / Potential biomarkers

The features with a maximum absolute difference in the abundance levels:

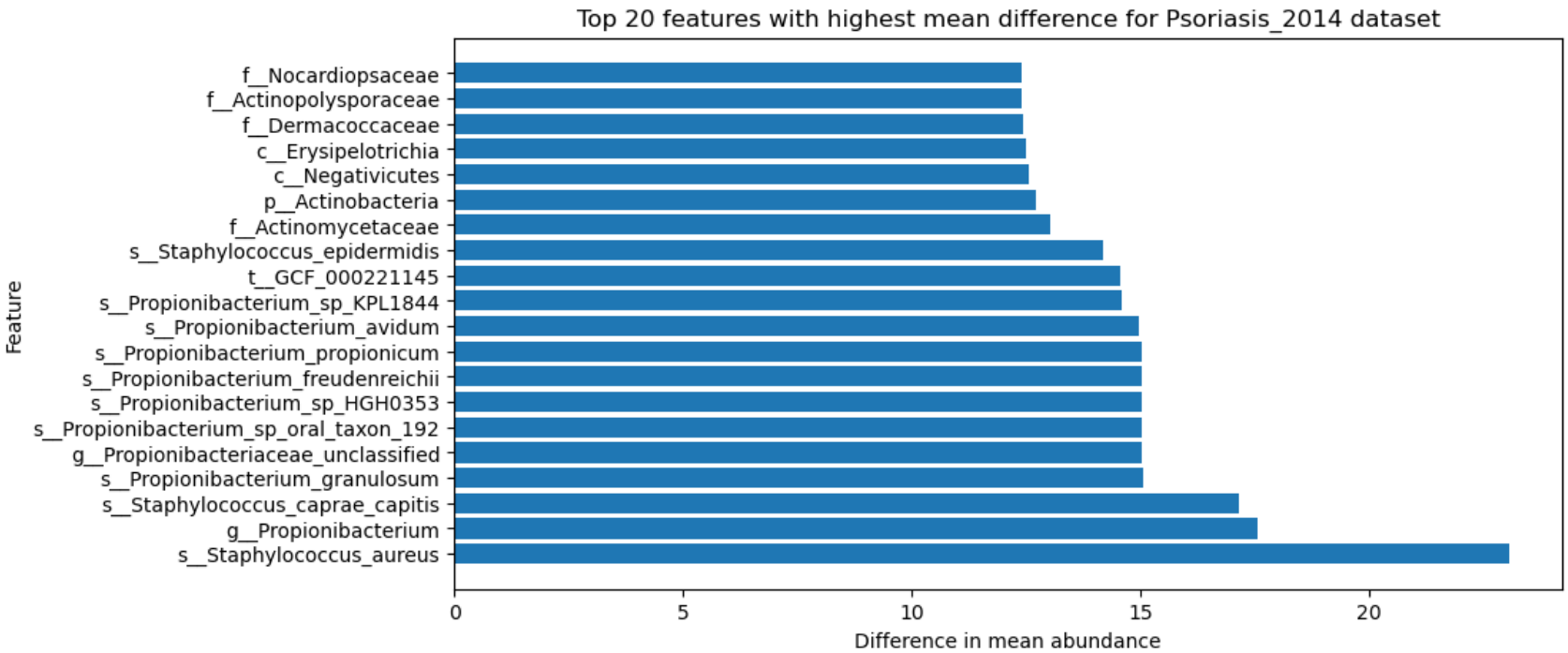

Figure: Psoriasis dataset

Top features / Potential biomarkers contd.

The features with a maximum absolute difference in the abundance levels:

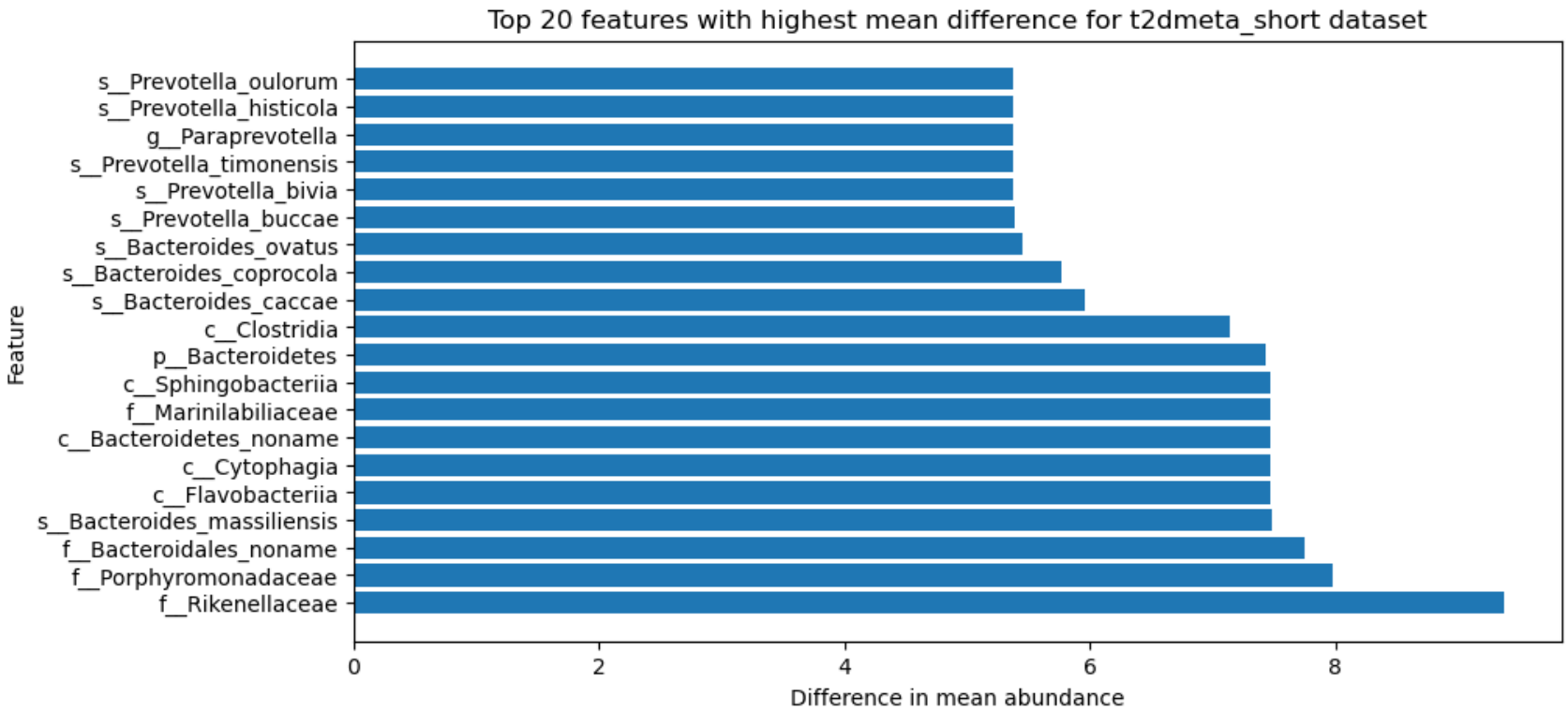

Figure: Type 2 diabetes dataset

Binary classification metrics

Dataset name	Train Accuracy	Test Accuracy	F1 Score
Zeller Fecal Colorectal Cancer	1.000	0.778	0.864
WT2D	0.983	0.828	0.884
Quin Gut Liver Cirrhosis	0.935	**0.894**	0.902
Psoriasis	0.946	**1.000**	1.000
IBD 1	1.000	**0.922**	0.909
IBD 2	1.000	**1.000**	1.000
E-Coli Stec2 2	1.000	1.000	1.000
Chatelier Gut Obesity	0.977	0.696	0.754
T2D 2 (Short)	1.000	**0.933**	0.923
T2D 1 (Long)	0.991	0.759	0.741
Obesity	1.000	0.818	0.500

Table: Performance on the subsets

Mixture of distributions learning

- Combined the positive and negative samples from all the subsets
- Total samples: 2200 negative, 1400 positive
- Used the entire feature set (3300 features) to train a model - classified most samples as negative
- Reason:
  - different distribution for positives but same for all negatives
  - very close relation between positive and negative samples of different subsets?
  - non-separable into learnable distributions?

Feature extraction

- Feature extraction methods $\rightarrow$ reduce the dimensionality of the data
- Compress the required information into fewer features to better understand the underlying structure for visualization and learning
- Principal Component Analysis (`PCA`):
  - statistical method for dimensionality reduction
  - selection of top-k eigenvectors as the principal components with maximum eigenvalues
  - projection onto this lower dimensional vector space with minimum loss in the variance of the data

Retained Variance

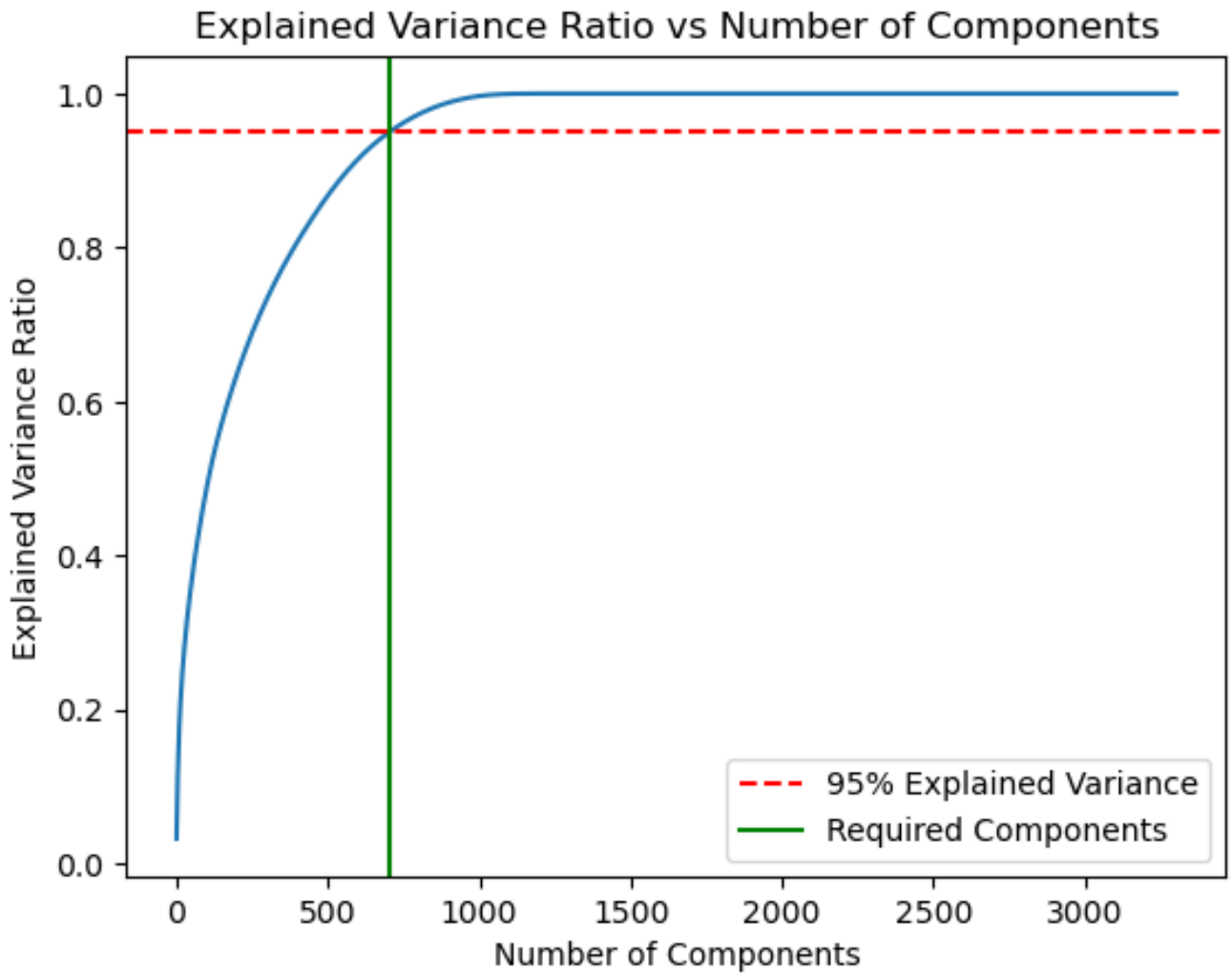

Figure: The wasted potential

Mixture of distributions learning + PCA

- 705 (23%) of features contain 95% of the variance of the 3300 features
- Training and hyperparameter tuning: layers, batch size, learning rate, epochs
- Total models trained: 350+
- Best performance: 83% on the test data (20% of the entire data)

Best hyperparameters & Inference

- The best set of hyperparameters to classify the samples into positive and negative is:
  - Input Dimension: 705
  - Layers: [256, 128, 64, 32]
  - Batch size: 64
  - Learning Rate: 5e-05 (0.00005)
  - Epochs: 200
- Shallow & narrower network $>$ deep & wider network
- Plateau is reached at almost 200 epochs
- Deeper and wider models tend to overfit even at very low learning rates

Multi-class classification + PCA

- Used the same 705 features to classify multiclass across 15 classes, one being "healthy".
- Training and hyperparameter tuning: layers, batch size, learning rate, epochs
- Total models trained: 180+
- Best performance: 80% on the test data (20% of the overall data)

Best hyperparameters & Inference

- The best set of hyperparameters to classify the samples into positive and negative is:
  - Input Dimension: 705
  - Layers: [1024, 512, 256, 128]
  - Batch size: 32
  - Learning Rate: 5e-05 (0.00005)
  - Epochs: 500 (maximum tried)
- Shallow networks $>$ deeper networks
- Seems to improve even after 500 epochs
- Narrower networks could improve performance

Feature extraction contd.

- One more method to extract features for classification task
- Linear Discriminant Analysis:
  - Supervised classification technique (requires labelled data)
  - Finds linear combination of features to separate classes while preserving discriminatory information
  - Aims to maximize the ratio of between-class variance to within-class variance
- Training a simple neural network gave 86% accuracy in test data for multi-class classification

Ongoing work

- Second level of hyperparameter tuning (more finer adjustments, early stopping)
- Hyperparameter tuning for models trained with LDA-modified data
- Train with few subsets and check its performance on the other sets for binary classification
- Few more feature extraction methods.

References

Truong, Duy Tin and Franzosa, Eric A, Et al. (2015)

MetaPhlAn2 for enhanced metagenomic taxonomic profiling

Nature Methods

Waldron, Levi and Edoardo, Truong, Et al. (2016)

Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights

PLOS Computational Biology

Thank you