Predicting the risk of diabetes

STATS/CSE 780 Course Project

Pao Zhu Vivian Hsu (400547994)
McMaster University

2023-11-30

Motivation

- Diabetes is a chronic disease that occurs when the body cannot effectively produce or use insulin to regulate blood sugar levels.
  - 422 million people have diabetes worldwide and 1.5 million deaths that occur every year are linked to diabetes (World Health Organization 2023).
- Machine learning techniques can be used to predict diabetes.
  - Different studies suggest different techniques to most accurately predict diabetes.
  - Islam et al.'s study compares 3 different techniques and states that their decision tree produced the most accurate results (2020). The study's data is publicly available for analysis.
- *GOAL*: Reproduce the decision tree in Islam et al.'s study to verify accuracy and compare with a neural network to assess whether this would be more accurate than a tree-based model.

Data (1 slide)

- Collected by Islam et al. from a hospital in Bangladesh (2020) and openly published on Kaggle (Larxel 2023) where it was downloaded.
- Data contains 17 variables
  - Response: Binary variable called class that indicates whether the patient has a positive (1) or negative risk (0) for diabetes
  - Predictors: 1 quantitative and 16 categorical variables describing the patient and if they experience common symptoms related to the disease, such as weakness, itching, and obesity
-
- Source?
- Is it a data frame?
  - What is in rows?
  - What is in columns?
- Results of exploratory analysis?
  - Data types, type of response if any?
  - Correlation analysis?
  - Outliers?
  - Missingness?

Data (Example)

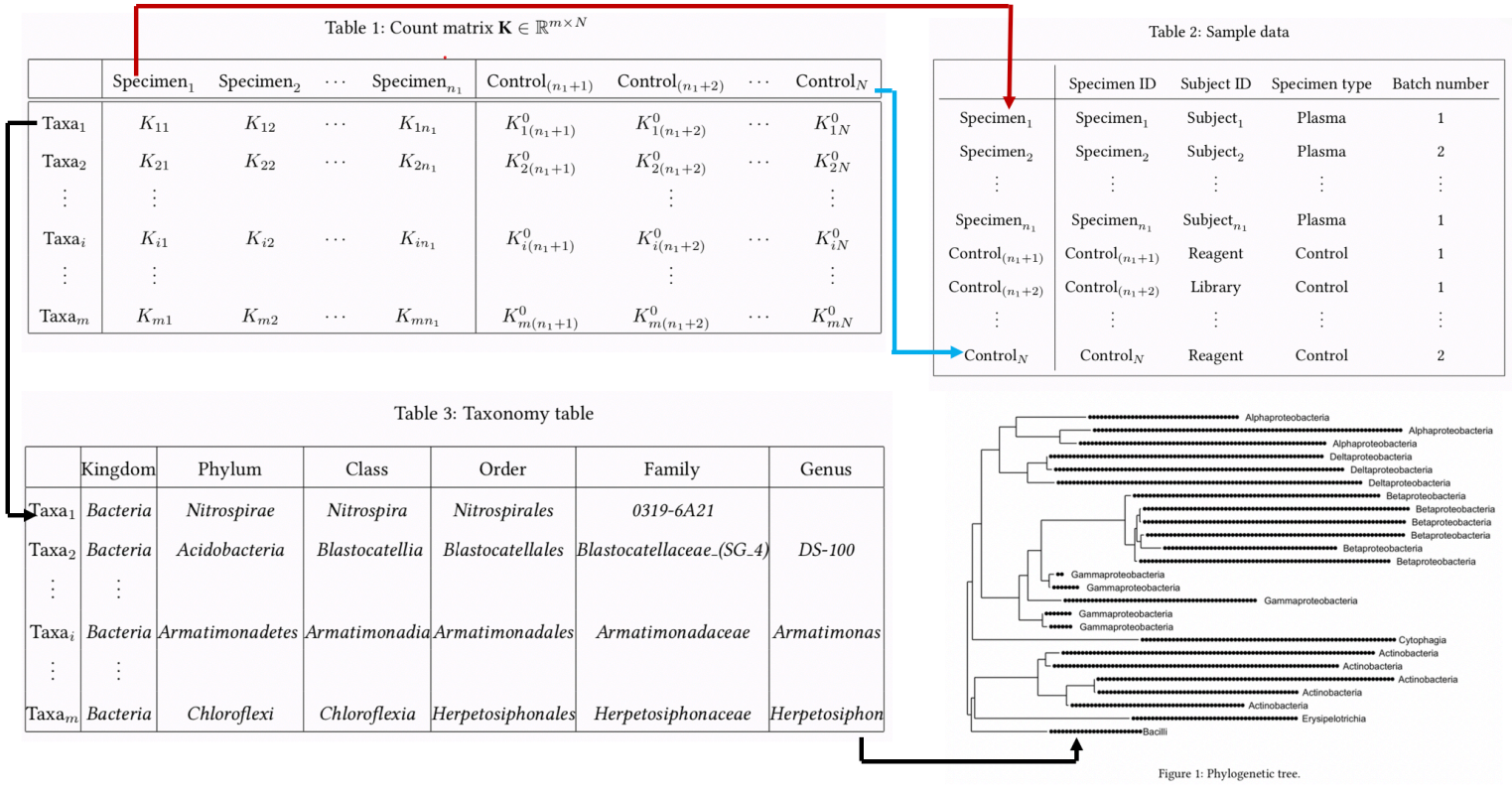

Table 1: Count matrix $\mathbf{K} \in \mathbb{R}^{m \times N}$

	Specimen_1	Specimen_2	$\cdots$	Specimen_{n_1}	$\text{Control}_{(n_1+1)}$	$\text{Control}_{(n_1+2)}$	$\cdots$	Control_N
Taxa_1	K_{11}	K_{12}	$\cdots$	K_{1n_1}	$K^0_{1(n_1+1)}$	$K^0_{1(n_1+2)}$	$\cdots$	K^0_{1N}
Taxa_2	K_{21}	K_{22}	$\cdots$	K_{2n_1}	$K^0_{2(n_1+1)}$	$K^0_{2(n_1+2)}$	$\cdots$	K^0_{2N}
$\vdots$	$\vdots$					$\vdots$		$\vdots$
Taxa_i	K_{i1}	K_{i2}	$\cdots$	K_{in_1}	$K^0_{i(n_1+1)}$	$K^0_{i(n_1+2)}$	$\cdots$	K^0_{iN}
$\vdots$	$\vdots$					$\vdots$		$\vdots$
Taxa_m	K_{m1}	K_{m2}	$\cdots$	K_{mn_1}	$K^0_{m(n_1+1)}$	$K^0_{m(n_1+2)}$	$\cdots$	K^0_{mN}

Table 2: Sample data

	Specimen ID	Subject ID	Specimen type	Batch number
Specimen_1	Specimen_1	Subject_1	Plasma	1
Specimen_2	Specimen_2	Subject_2	Plasma	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Specimen_{n_1}	Specimen_{n_1}	Subject_{n_1}	Plasma	1
$\text{Control}_{(n_1+1)}$	$\text{Control}_{(n_1+1)}$	Reagent	Control	1
$\text{Control}_{(n_1+2)}$	$\text{Control}_{(n_1+2)}$	Library	Control	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Control_N	Control_N	Reagent	Control	2

Table 3: Taxonomy table

	Kingdom	Phylum	Class	Order	Family	Genus
Taxa_1	*Bacteria*	*Nitrospirae*	*Nitrospira*	*Nitrospirales*	*0319-6A21*	
Taxa_2	*Bacteria*	*Acidobacteria*	*Blastocatellia*	*Blastocatellales*	*Blastocatellaceae_(SG_4)*	*DS-100*
$\vdots$	$\vdots$					
Taxa_i	*Bacteria*	*Armatimonadetes*	*Armatimonadia*	*Armatimonadales*	*Armatimonadaceae*	*Armatimonas*
$\vdots$	$\vdots$					
Taxa_m	*Bacteria*	*Chloroflexi*	*Chloroflexia*	*Herpetosiphonales*	*Herpetosiphonaceae*	*Herpetosiphon*

Figure 1: Phylogenetic tree.

Figure 1: Source: xxx

K_{ij} abundance of i-taxon in j-th sample.

Methods (2 slide)

- What are the two methods you compared?
- Why those two methods?
- Algorithms of the methods?
- Any statistical transformation used?
- Any other pre-processing (feature engineering) used?
- Any feature selection (filter, or wrapper, or embedded) used?
- etc.?

Methods (Example)[1]

- KNN and DT for classification.
- Decision trees - partition the predictor space into simple regions.
  - Predict y_0 of a new data point x_0 using the response of training observations in the region to which x_0 belongs.
  - How to find the partitions?
- KNN -

[1] An introduction to statistical learning (**james2013introduction?**)

Results (1 slide)

- What are the results of applying the methods?
  - Visualize the results?
  - Compare the methods using graphs?
  - Interpret the model/results?
  - etc.?

Discussion (1 slide)

- Discuss problems related to the methods and data -
  - Curse of dimensionality?
  - Multiple data types?
  - Interpretablity?
  - Reproducibility?
  - Stability?
  - etc. ?

Thank You!

References

Islam, M. M. Faniqul, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. 2020. "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques." In *Computer Vision and Machine Intelligence in Medical Image Analysis*, edited by Mousumi Gupta, Debanjan Konar, Siddhartha Bhattacharyya, and Sambhunath Biswas, 113–25. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-8798-2_12.

Larxel. 2023. "Early Classification of Diabetes." https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

World Health Organization. 2023. "Diabetes." https://www.who.int/health-topics/diabetes#tab=tab_1.