Cluster analysis of multi-morbidity in UK Biobank participants

Data Analysis Plan

Introduction

This data analysis plan describes the aims, methodology **and outcomes** that will be used for the cluster analysis of multi-morbidity in UK Biobank participants. These analyses will be carried out by the Barracudas (MSc Health Data Analytics and Machine Learning students).

Aims

The overall objective of this study is to identify subsets of individuals suffering from multi-morbidity who share common environmental and/or biological pathways. To do this we are going to analyse UK Biobank participants.

* We will identify sub-groups of individuals suffering from multi-morbidity
* We will compare cluster analysis methods
* We will attempt to identify possible shared and distinct underlying physio-pathological processes between clusters

Dataset

UK Biobank data.

UK Biobank is a prospective study investigating the contributions of genetic predisposition, and lifestyle and environmental exposures to the development of disease in 500 000 people aged 40-69 in the UK.

The dataset contains numerical, binary and categorical variables.

Using the six disease variables present in the dataset, there are 61 000 UK Biobank participants with multimorbidity.

Exploratory Data Analysis

Dataset overview (Josh-Table 1)

* Numbers of diseases across important epidemiological factors e.g. age, sex, smoking, alcohol

Distributions of diseases across important epidemiological factors

Outcomes

* Clusters of Biobank participants with multi-morbidity (across different methods)
* Data similarity measurements for clusters/distance metrics for clusters

Data Analysis Clustering

**Algorithms :**

1. Kmeans
2. Joël: Partitioning around Medoids (Josh could also look into that since it’s probably

close to what he does.

1. Joël: DBSCAN
2. Gaussian Mixture Models
3. Josh to choose

**Pre- processing:**

Clustering methods require specific inputs to compute clusters.

Some only work with continuous features (kmeans) while others require a similarity/distance matrix. None of the algorithms directly work with the initial mixed data.

We can thus have two approaches to our mixed dataset :

1. Make meaningful continuous features out of our initial dataset through adapted dimensionality reduction
2. Make meaningful similarity matrices out of the mixed data

How to get 1)

Explore dimensionality reduction techniques for mixed datasets:

* FAMD => ~ Mixed PCA with the FactoMineR package
* Auto-Encoder => One hot encode everything and throw a Neural Network at it

How to get 2)

* Use 1) then some distance measures for continuous features
* Use Gower distance/Random Forests on the whole dataset to compute similarities from the mixed data.

**Pipelines Joël :**

Kmeans

FAMD = > Kmeans

Auto-Encoder => Kmeans

RF => DBSCAN

RF => PAM

Gower distance matrix => PAM

Gower distance matrix => DBSCAN

Auto-Encoder => Some distance calculation => DBSCAN

Auto-Encoder => Some distance calculation => PAM

FAMD => Some distance calculation => PAM

FAMD => Some distance calculation => DBSCAN

Stability analyses?

Methods for cluster evaluation

* Silhouette coefficient- how well defined are the clusters for each model
* Calinski-Harabasz criterion
* Interpretability of the clusters (this will be the big thing)

Interpretation of clusters

* Dimensionality reduction and interpretation of components?
* Univariate tests for all variables?
* Interpretation plots?