Comparison of Lightweight Clustering Methods on SNP Data for Population Structure Exploration

Isaac Thomas

March 7, 2024

0 TODO

- download `plink` (right version) and get it to run ✓
- download pruned vcf data for chromosome 19 ✓
- get PCA results ✓
  - map each row of PCA data to its population code and subpopulation code ✓
- build and run k-means
  - plot silhouette score vs k ✓
  - compute silhouette score of optimal clustering ✓
  - plot the optimal clustering ✓
    - label/color points by their population/subpopulation ✓
    - map each point to plot to its shape (true label) and color (true label) ✓
    - label/color points by their population/subpopulation ✓
- build and run hierarchical clustering
  - compute optimal number of clusters for agglomerative clustering
  - plot silhouette score vs k for agglomerative clustering
  - plot optimal clustering
  - label/color points by their population/subpopulation
  - plot dendrogram

1 Introduction

1.1 Motivation

Single Nucleotide Polymorphisms (SNPs) have had massive intergenerational effects on human phenotypes. Such impact renders SNP data useful in describing phenotypes to an extent that allows informative clustering. As biologically/probabilistically informed methods for this task are more expensive and require more detailed data or heavy computation, we ask a crucial question: what is the simplest clustering method that one can "get away" with while informatively clustering SNP data with the goal of capturing population demographics? This paper explores answers to this inquiry in the context of two well-known methods: K-means clustering and Ward's method for agglomerative and divisive hierarchical clustering. We compare these two methods' abilities to cluster SNP data from the 1000 Genomes dataset and determine how much insight they can "get away" with providing without significant biological or probabilistic information/assumptions.

1.2 Overview

Section 2 covers the details of the 1000 Genomes dataset used, the preprocessing methods used, and the model training. We explore the structure of the 1000 Genomes dataset, and we go over methods used to filter out the SNPs that would be most useful for clustering tasks. Section 3 compares the abilities of these models to each other and against more biologically or probabilistically informed methods. We compare training procedures, training time, and hyperparameter tuning needs of these two methods. Section 3 compares the insights extracted by each model's clustering of the SNP data and discusses promising directions for future research in this area. We determine which method was more informative and discuss the performance-insight tradeoff relevant to both. We also discuss future research for lightweight methods that can informatively cluster SNP data.

2 Methods

2.1 Data & Preprocessing

2.1.1 Dataset

2.1.2 Preprocessing

2.2 Models

2.2.1 K-Means Clustering

2.2.2 Hierarchical Clustering (Ward's Method)

3 Discussion

3.1 Model Comparison

3.2 Promising Directions for Future Research