Homework 3 Report

Jae Young Kim

Applied Data Science, University of Southern California,

Los Angeles, California 90089, USA

(Dated: March 11, 2021)

Abstract

A new SARS-CoV-2 variant appeared and it is more dangerous than other strains. However, an individual Z who seems to be immune to this new variant was found. Therefore, the patients who has same type of genetic composition as Z were identified through this research. Genetic fingerprint data of the patients were used and K-means method was implemented to identify the clusters of the patients. With these clusters, patients who are in the same cluster with individual Z were identified. Applying K-means, 5 different clusters were found and 2800 patients were in the same cluster with individual Z. The final result was submitted to Kaggle and scored 0.9940 F1 score.

I. INTRODUCTION

To make clusters of the patients, I explored "ps3_genetic_fingerprints.npy" data set and trained a clustering model. As there were 386 features and it is quite high dimensional, most data exploration and preprocessing steps have been focused on reducing the dimension of the data without losing information. Before dealing with individual columns, minmax scaler was applied to all columns. Then, the columns with only one value for all of the patients were dropped and columns with high correlation were dropped. After that, to reduce dimension, t-distributed stochastic neighbor embedding(tsne) was applied. The number of dimension of the data was dropped to 3.

After reducing dimension of the data, to find the appropriate number of clusters(k), silhouette score and within cluster sum of squares(WCSS) for each k were calculated. As a result, k=5 was chosen to be the best number of clusters.

With the number of clusters, 5, k-means algorithm was applied to the data and the patients in the same cluster with individual Z were identified. This result was submitted to Kaggle and scored 0.9928 F1 score.

II. DATA EXPLORATION AND PREPROCESSING

To preprocess data, first and foremost, minmax scaler was applied to each columns. Then, the columns with unique value were dropped. 115 columns out of 398 columns were dropped in this process. For the columns with two values, to scale it, those values were converted into 0 and 1. There were 19 columns which has two values.

To explore data, histograms and hit map of correlation matrix were drawn. Every column had many zero values. And some histograms seems to be composed of two different distributions. For example, histogram in figure 2 is a histogram of 76th feature and it seems to be composed of one long-tail distribution and one normal distribution.

After scanning feature distributions, hit map of correlation matrix was drawn(figure 3). Many white spots and black spots were found and the dots mean that there are many features which are highly correlated to each other. Therefore, the columns which has above 0.8 absolute correlation were dropped. 58 columns were dropped through this process. After dropping highly correlated features, hitmap of correlation matrix is drawn once more(Figure

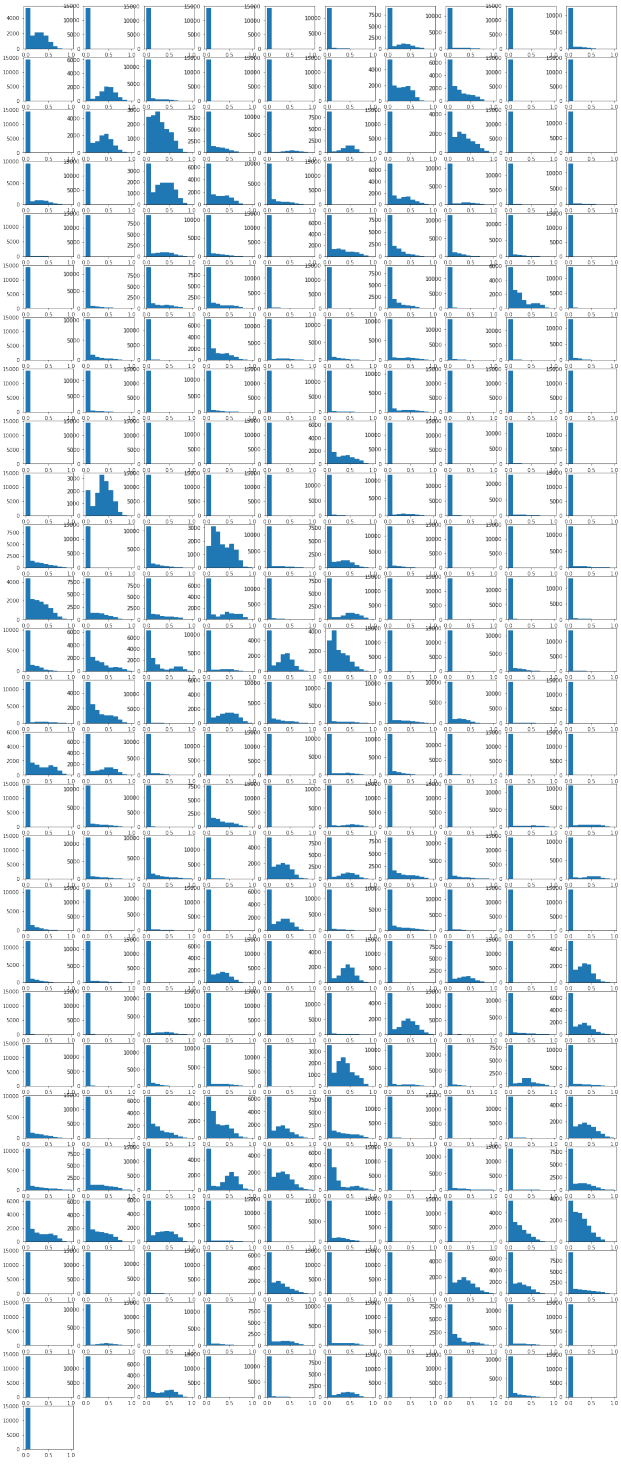

FIG. 1. Histograms of features

4). It is possible to check that white dots and Black dots were removed.

Finally, T-sne is applied to the data to reduce the dimension to 3.

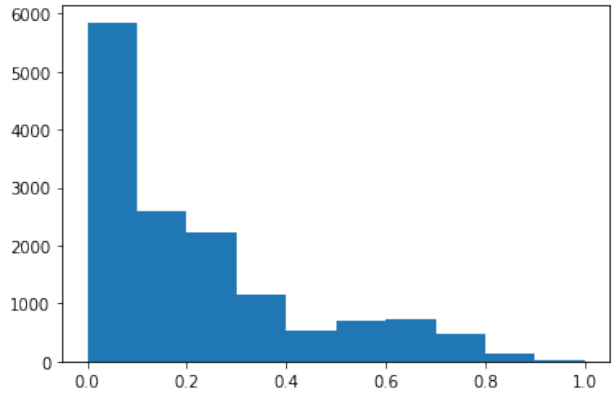

FIG. 2. Histograms of 76th feature

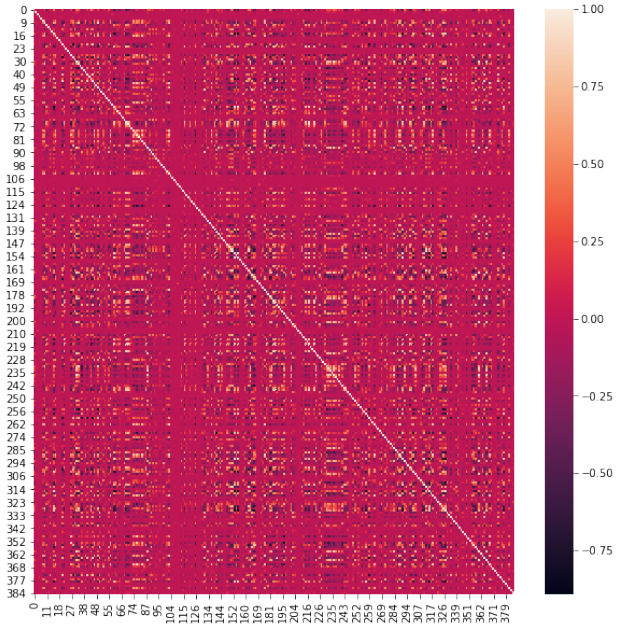

FIG. 3. Hitmap of correlation matrix before feature drop

III. MODEL SELECTION AND EVALUATION

As K-means is the most intuitive clustering model, K-means was used. To select the number of clusters, silhouette score and within cluster sum of squares(WCSS) were calculated for each number of clusters(k) and plotted in Figure 5 and Figure 6.

As the plot shows peak at k = 5(silhouette score = 0.46) in Figure 4 and elbow at k = 5 in Figure 5, 5 is chosen to be the number of clusters.

The final result is visualized by using PCA in Figure 7 and 8. The red cluster and cobalt cluster seems to be mixed in 2D space but it looks well separated when it is visualized on

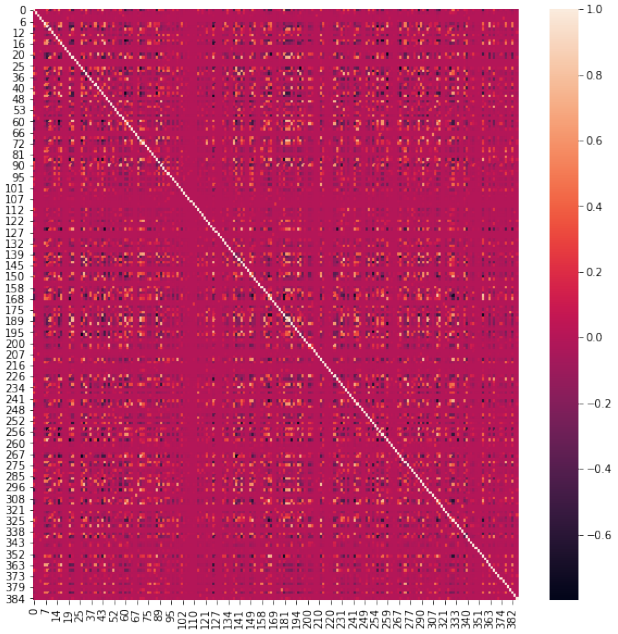

FIG. 4. Hitmap of correlation matrix after feature drop

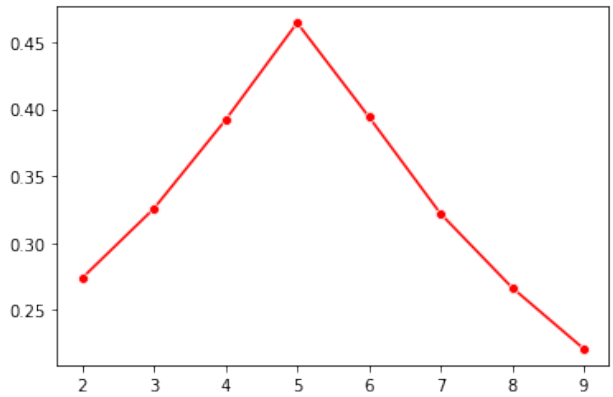

FIG. 5. Plot of silhouette score

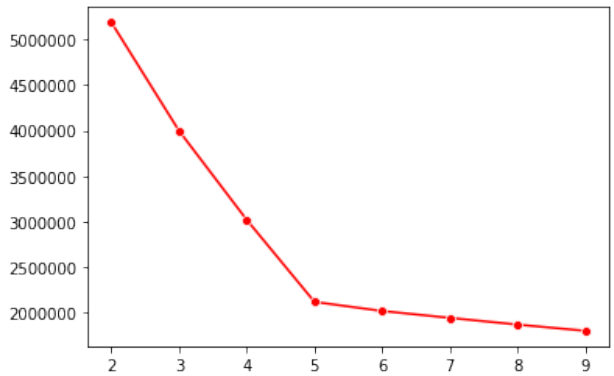

FIG. 6. Plot of WCSS

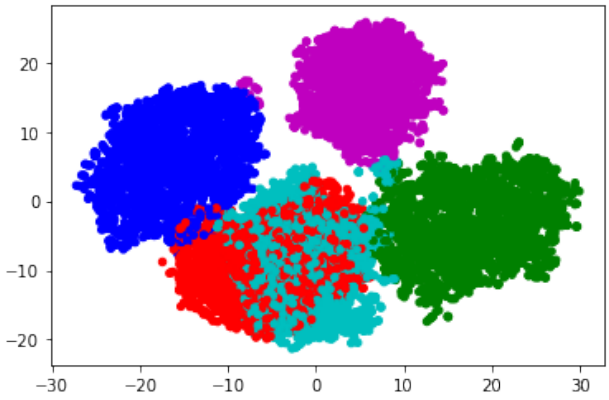

FIG. 7. K-means result visualized on 2D space

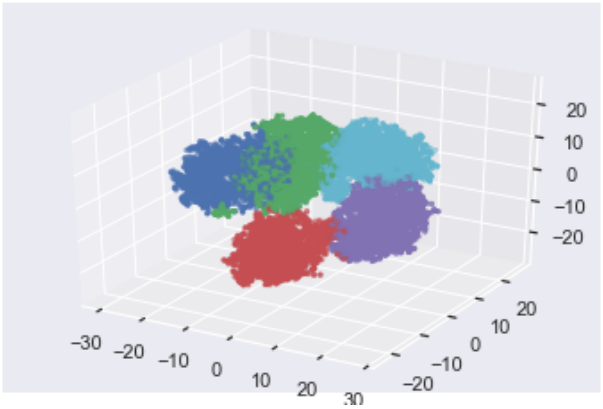

FIG. 8. K-means result visualized on 3D space

3D space.

Purple cluster is the cluster where patient Z is included. In addition, there were 2800 patients who were in the same group with patient Z. For evaluation, the result was submitted to Kaggle and graded with F1 score. The final score was 0.9940.

IV. CONCLUSIONS

To sum up, patients were clustered into five different groups according to genetic fingerprint data. K-means algorithm was used to solve this problem. Five different clusters were found and Individual Z, a patient seems to be immune to new SARS-CoV-2 variant, was clustered in one of the groups with 2800 other patients. The result is graded on Kaggle and scored 0.9940 F1 score.

DATA AVAILABILITY

Data is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps3-jeayoung114/data

CODE AVAILABILITY

Code is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps3-jeayoung114/code/hw3