Clustering the Genetic Footprint

Maxwell Lee

Viterbi School of Engineering, University of Southern California,

Los Angeles, California 90089, USA

(Dated: January 19, 2022)

Abstract

In this report, we attempt to cluster the genetic footprint of thousands of patients in hopes to find similarities to an individual expressing immunity to a new SARS-CoV-2 variant. We used a K-Means clustering algorithm and determined the data exhibits 5 main clusters. The identified immune patient belongs comfortably to a cluster containing 2886 patients. Further research should be conducted with these patients in the hopes of advancing our knowledge of this disease and hopefully discovering protection from this new variant.

I. INTRODUCTION

For this task, we are given information about 14398 patients. Each patient has a genetic fingerprint consisting of a vector of 386 values. We will attempt to cluster these patients to reveal similarities in their genetic footprint. This is important, because a patient is believed to be immune to a new variant of the SARS-CoV-2 virus. We would like to know if this is related to their genetic makeup, but doing an experiment with only one subject is not good science. Instead, we look to find patients with similar genetic makeups so we can conduct research with the larger group.

II. DATA EXPLORATION

The data given to us is said to have already been cleaned, so the task of data exploration is less important than usual. However, there are still some insights to glean from the data. 115 columns contain no information, that is their values are all the same. For our purposes, we could drop these columns to reduce the computational time needed for our K-Means algorithms as no information is provided. However, if this were to be used in the future dropping these specific columns may be problematic, as a new patient or our test patient that does have new information for these columns would be incompatible with the model. For this reason, we decide to keep these columns, as they do not adversely affect the K-Means algorithm, as they all will give a distance of zero.

III. DATA PREPROCESSING

As the data is already cleaned, the main data preprocessing decision is whether or not to employ some standardization of the data. Each feature, or value in the vector, will have its own range. Without standardization, this will give greater weight to the features with a greater range, as their relative distances will be greater and K-Means treats all feature distances equally. In general, it usually does not hurt to standardize the data. The capacity for larger errors falls on the side of not standardizing. In this data set, all features already have similar ranges, with all values being positive, most features being in the range of 0 to 8, with no feature having a range larger than 0 to 25. Thus, the decision to standardize the data or not will likely not have a significant impact. We will attempt K-Means with both

to see if any noticeable differences arise.

IV. MODEL SELECTION

We will strictly be using the K-Means clustering algorithm as our model. We will attempt the clustering with two versions of our data set: the unaltered data and the standardized data. For each, we will attempt to find the optimal value for K by iterating different K values and computing the Silhouette Score. The Silhouette score is a measure of both intra and inter-cluster similarity, rewarding intra-cluster similarity and penalizing inter-cluster similarity. Ideally, there will be a clear maximum value for the Silhouette Score associated with the optimal K value for the task.

To perform these tasks, we use Sci-Kit Learn's K-Means algorithm along with its StandardScaler algorithm to create our standardized data set [1]. Each data set was tested with K values from 2 to 10. The resulting plots are shown in Figure 1 and Figure 2.

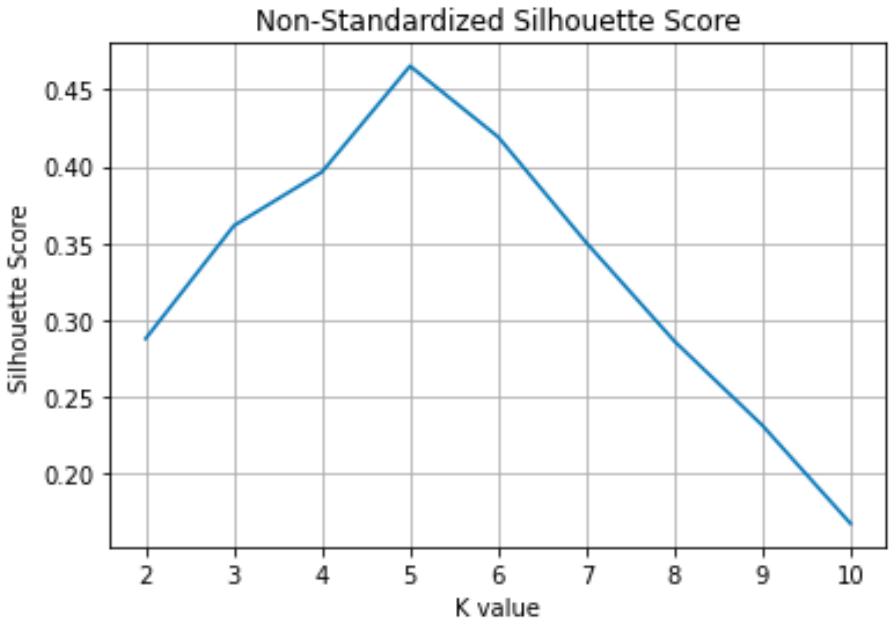

FIG. 1. Silhouette Scores for Non-Standardized Data Set

Clearly, both data sets have a maximum Silhouette Score at K=5, indicating there should be 5 clusters in our data. The actual values of the Silhouette Score should not be used to compare the two approaches as the data sets are different and thus the distances used to compute the Silhouette Score will be different.

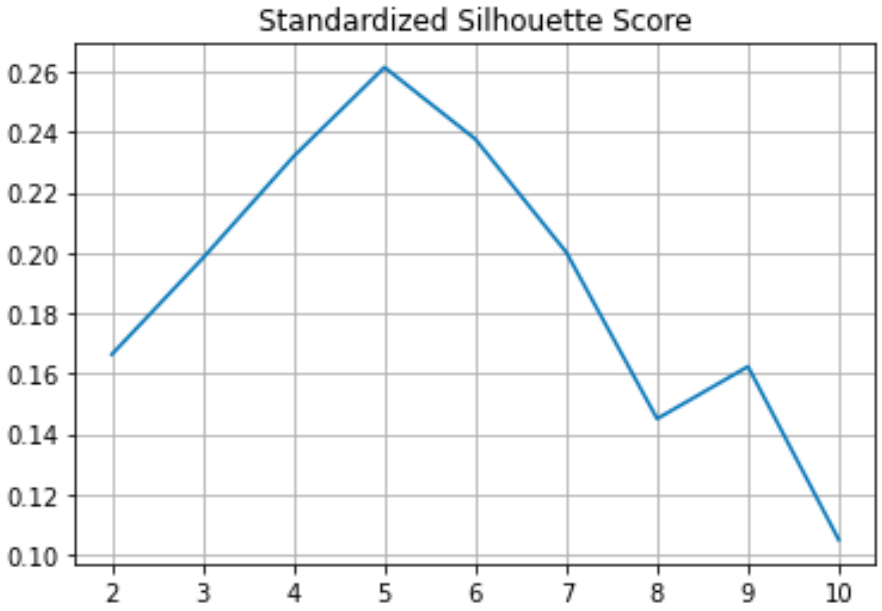

FIG. 2. Silhouette Scores for Standardized Data Set

To compare the two approaches, we fit the K-Means algorithm with K=5 for both data sets. We then use the resulting labels to calculate the Silhouette Score on each data set with each approach. For the non-standardized data set, both approaches produce a Silhouette Score of .465, and for the standardized data set, both approaches produce a Silhouette Score of .261. Any differences between the two approaches are simply a rounding error. As we predicted in Section II, the decision to standardize or not is not important to this particular problem. So, we will decide to standardize the data, as that is generally the better practice, and is inconsequential to the overall task.

V. MODEL EVALUATION

Evaluating any unsupervised task is inherently difficult. We do not have ground truth values to partition a test set and evaluate the effectiveness of our model. We have the Silhouette Score, but even that is mostly useful in the relative sense for determining model approaches or parameters. To determine if our data is clustered and if we captured those clusters, it is useful to visualize the data. While we obviously cannot visualize 386 dimensions, we can apply the PCA algorithm to reduce the dimensionality of our data down to 2 dimensions so that we can visualize the data. In doing so we of course lose some of the variance and signal in our data, but this trade off must be made to be able to visualize.

Applying the PCA algorithm from Sci-Kit Learn [1], we can visualize the clusters as shown below in Figure 3. The points are colored according to their cluster assigned by the 5 cluster K-Means algorithm on the standardized data set.

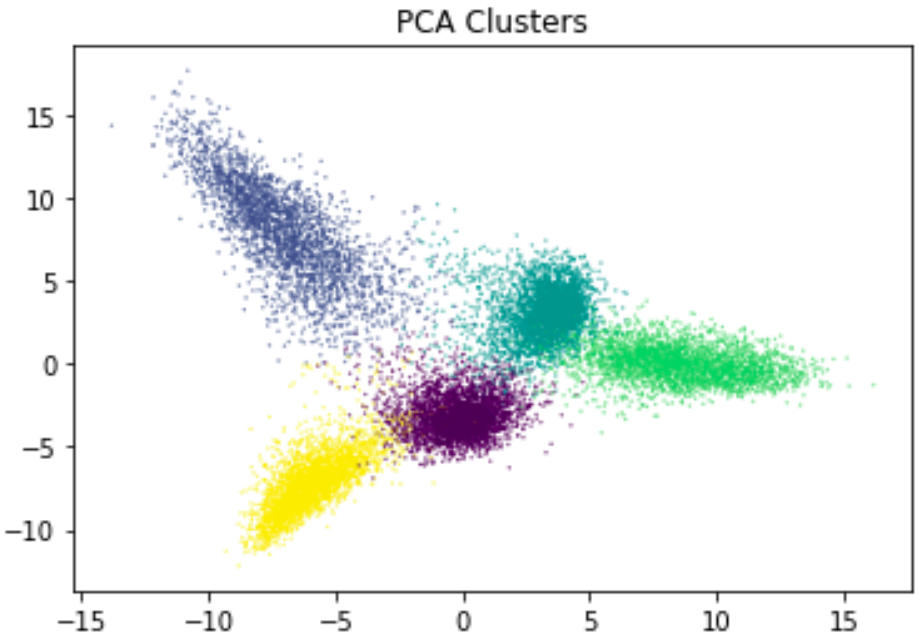

FIG. 3. 5 clusters displayed with their colors after PCA reduction

As we can see, the clustering appears to be rather successful. The choice of 5 clusters appears to be appropriate and the coloring of points aligns with the location of the clusters. There are points in between the blue, teal, and purple clusters that may not truly belong to any of the clusters, but overall the model appears to be a success.

VI. INTERPRETATION

To apply this model, we need to identify which cluster our test patient belongs to, so we can perform research with that cluster's patients. We can take the genetic makeup of the test patient and evaluate which centroid the test patient would be closest to. Doing so results in a classification of the purple cluster. We can verify this by visualizing the test patient on the PCA plot, shown below in Figure 4.

We can see that the test patient is very near the centroid of the purple cluster, verifying the result of the K-Means prediction. 2886 individuals belong to this cluster and are thus worth studying. As mentioned in Section V some of the fringe individuals in between the

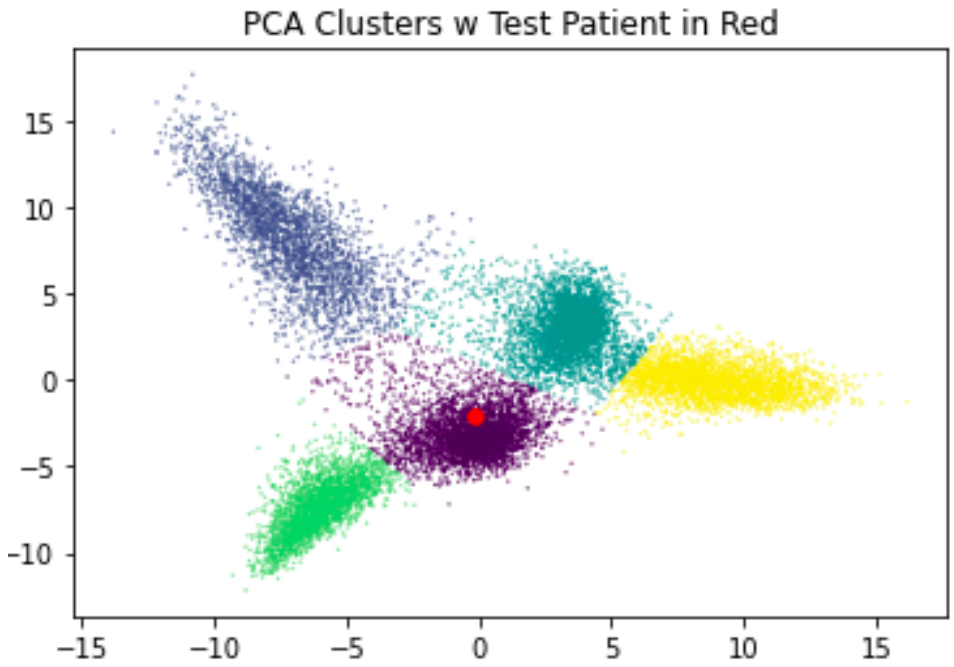

FIG. 4. Test patient shown in red in the PCA clusters

blue, teal, and purple clusters may be included in this purple cluster without sharing an appropriate level of similarity to our test patient. Further work could be done to identify some of these fringe patients if inclusion of these individuals would be costly and not worth studying.

While this will rarely be the case in a real world application, it is worth mentioned the model achieved an F1 score of .975 in identifying patients similar to our test patients in the Kaggle competition.

VII. CONCLUSIONS

In this task, we were able to identify appropriate clusters for the genetic makeup of the patients with relative ease. We were able to identify five as the appropriate number of clusters, as well as create appropriate centroids for the K-Means algorithm. We are able to validate both of these findings visually using the PCA algorithm. And most importantly, we identified 2886 patients that have a potentially similar genetic makeup to our test patient

that should be studied to further our knowledge of this new SARS-CoV-2 variant.

---

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research **12**, 2825 (2011).