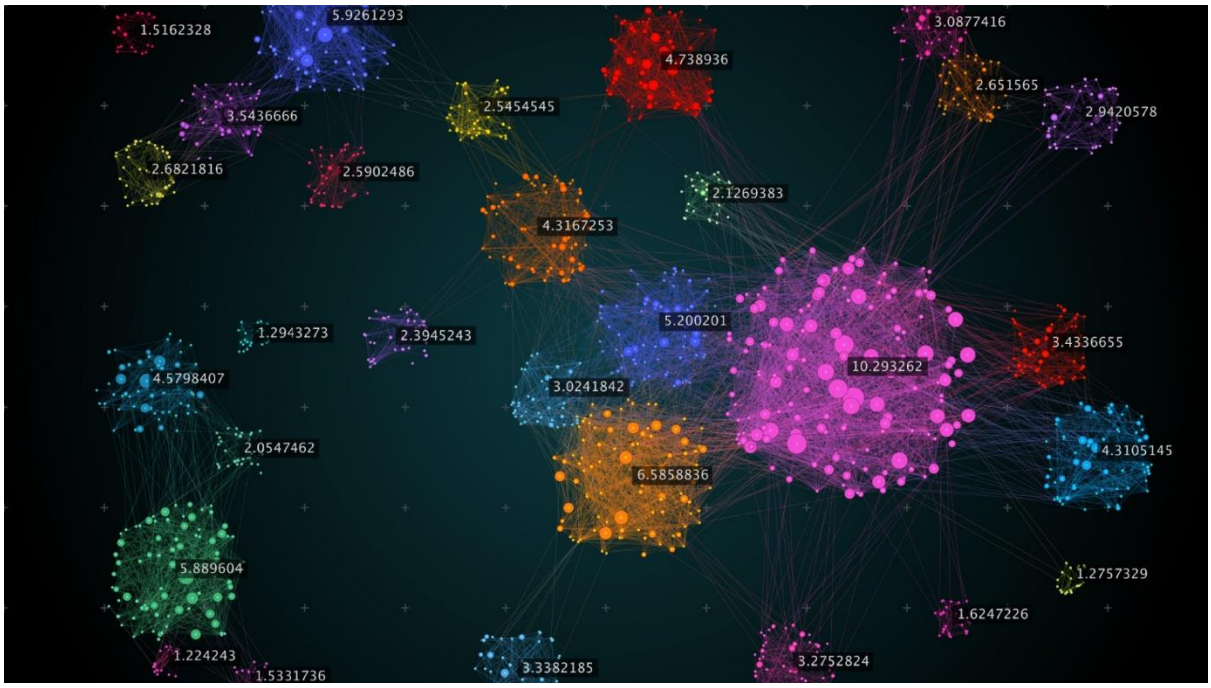
MACHINE LEARNING(CS60050): SPRING 2021

Virus Clustering using Complete Linkage Hierarchical Clustering Technique

Project Code: VC4

Submitted by: Akash Singh Sant(20CS60R40)



Introduction

We are given the data about features of different kinds of viruses found in the bodies of different people. These features include size of the virus, number of nucleic acid molecules and transmission rate. We are to cluster these viruses into groups.

GOAL: The Goal of this task is to cluster the dataset into an optimal number of clusters.

Sample rows are given below and the full dataset can be accessed here

	size	number_of_strands	number_of_protein_layers	sam ratio	latency
0	276.0	11	9	6.664	269
1	275.0	14	6	10.960	269
2	205.8	8	4	6.742	318
3	275.8	11	6	7.080	269
4	103.1	4	7	9.510	42

Sample rows after Preprocessing and Normalisation

	size	number_of_strands	number_of_protein_layers	sam ratio	latency
0	1.603344	-0.072941	0.572755	-0.310962	0.307881
1	1.591673	0.725679	-0.491846	1.871746	0.307881
2	0.784040	-0.871560	-1.201579	-0.271332	0.701426
3	1.601010	-0.072941	-0.491846	-0.099601	0.307881
4	-0.414571	-1.936387	-0.136979	1.135031	-1.515277

We'll Cluster the data In two ways:

1. K-means Clustering
2. Bottom-up hierarchical clustering

K Means Clustering

Method:

- Specify number of clusters *K*.
- Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids or a fixed number of iteration decided. In our algorithm we have chosen 20 iterations.
- Compute the Euclidean distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the mean of the all data points that belong to each cluster.

I have implemented this k means clustering algorithm for k=3,4,5,6

To determine which k value will give us the optimal clustering I have used Silhouette Coefficient metric

$$S= (b-a)/max(a,b)$$

The Silhouette Coefficient is defined for each sample and is composed of two scores:

a: The mean distance between a sample and all other points in the same cluster.

b: The mean distance between a sample and all other points in the next nearest cluster.

The k value which gives the maximum value of S is the optimal value of k.

Bottom-up hierarchical clustering using Complete linkage

The optimal k value found earlier is used here.

The algorithm begin considering every data point as a cluster and then these cluster are merged in a bottom up fashion and is run till k clusters are obtained.

In complete-linkage clustering, the link between two clusters contains all element pairs, and the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other. The shortest of these links that remains at any step causes the fusion of the two clusters whose elements are involved.

Mathematically, the complete linkage function, the distance D(X,Y) between clusters X and Y is described the following expression :

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y)$$

Where d(x,y) is the distance between elements x$\in X$ and y$\in Y$

X and Y are two sets of elements (clusters).

A run of the Algorithm:

I ran the k-means Algorithm on the given dataset for k=3,4,5,6 and the following

Silhouette coefficient value.

```
Silhoutte Coefficient for k = 3 ==> 0.33010860425140515
Silhoutte Coefficient for k = 4 ==> 0.33933201523946094
Silhoutte Coefficient for k = 5 ==> 0.33390115318286595
Silhoutte Coefficient for k = 6 ==> 0.32076762579365053
```

Seeing the value of Silhouette we can see that optimal k=4

```
Optimal K = 4
```

The Cluster information for k=4 is stored in the file kmeans.txt.

Now using this optimal value of k i have run the Bottom-Up Hierarchical Clustering Algorithm using the same notions of similarity as k-means clustering (Euclidean distance).

Also the k clusters are found using the complete linkage strategy in hierarchical clustering algorithm.

The k cluster obtained using the hierarchical algorithm are stored in agglomerative.txt

After obtaining the k cluster from k-means algorithm and k cluster from hierarchical clustering algorithm I have computed the Jaccard similarity between corresponding sets of both the cases.

The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets.

Now since we have 4 cluster in both the sets we'll have 16 values of Jaccard Similarity.

$$J(A,B) = (|A \cap B|)/(|A \cup B|)$$

Where A is some cluster of group A and B is some cluster of group B.

Following pair of Jaccard Similarity values were obtained.

```
Jaccard similarity, kmeans Cluster 1 <==> Heirarchical Cluster 1 = 0.14516129032258066
Jaccard similarity, kmeans Cluster 1 <==> Heirarchical Cluster 2 = 0.035714285714285714
Jaccard similarity, kmeans Cluster 1 <==> Heirarchical Cluster 3 = 0.0
Jaccard similarity, kmeans Cluster 1 <==> Heirarchical Cluster 4 = 0.28888888888888886

Jaccard similarity, kmeans Cluster 2 <==> Heirarchical Cluster 1 = 0.11811023622047244
Jaccard similarity, kmeans Cluster 2 <==> Heirarchical Cluster 2 = 0.2608695652173913
Jaccard similarity, kmeans Cluster 2 <==> Heirarchical Cluster 3 = 0.0
Jaccard similarity, kmeans Cluster 2 <==> Heirarchical Cluster 4 = 0.0

Jaccard similarity, kmeans Cluster 3 <==> Heirarchical Cluster 1 = 0.18213058419243985
Jaccard similarity, kmeans Cluster 3 <==> Heirarchical Cluster 2 = 0.032154340836012864
Jaccard similarity, kmeans Cluster 3 <==> Heirarchical Cluster 3 = 0.22916666666666666
Jaccard similarity, kmeans Cluster 3 <==> Heirarchical Cluster 4 = 0.25806451612903225

Jaccard similarity, kmeans Cluster 4 <==> Heirarchical Cluster 1 = 0.26384364820846906
Jaccard similarity, kmeans Cluster 4 <==> Heirarchical Cluster 2 = 0.39846743295019155
Jaccard similarity, kmeans Cluster 4 <==> Heirarchical Cluster 3 = 0.0
Jaccard similarity, kmeans Cluster 4 <==> Heirarchical Cluster 4 = 0.010909090909090911
```

Conclusion:

On multiple runs of the algorithm I found the Silhouette coefficient value to be in the range (0.25,0.42) thereby concluding that clusters are moderately dense. On considering the Jaccard Similarity of some cluster from case A with all the clusters of case B, I found one cluster is showing high similarity while the other three are quite dissimilar.