CS60050 - Machine Learning

Assignment 3

Country Grouping using Complete Linkage Divisive (Top-Down) Clustering Technique

Bratin Mondal

21CS10016

Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur

1 Problem Statement

- Implement K-means clustering algorithm from scratch for $k = 3, 4, 5, 6$.
- Find the optimal value of k using the silhouette coefficient metric.
- Implement Complete Linkage Divisive (Top-Down) Clustering Technique from scratch for the optimal value of k found in previous step.
- Establish a one-to-one correspondence between the clusters obtained from K-means and Complete Linkage Divisive (Top-Down) Clustering Technique and calculate the Jaccard Similarity.

2 Model Description

2.1 K-means Clustering Algorithm:

Algorithm:

```
Initialize $K$ cluster centroids randomly
repeat
    Assign each data point to the nearest centroid
    Update each centroid as the mean of the data points assigned to it
until Convergence
```

2.2 Silhouette Coefficient:

Formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where,

- $s(i)$ is the silhouette coefficient of object i
- $a(i)$ is the average distance from i to other objects in the same cluster
- $b(i)$ is the average distance from i to objects in the nearest cluster that i is not a part of

2.3 Complete Linkage Divisive (Top-Down) Clustering Technique:

Pseudocode:

```
Start with all data points in one cluster
Until the desired number of clusters is reached:
    Find the cluster with the maximum diameter to split
    Split the cluster into two based on the farthest points
    Update the cluster assignments
```

2.4 Jaccard Similarity:

Formula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

where,

- $J(A, B)$ is the Jaccard similarity coefficient between sets A and B

- $|A \cap B|$ is the size of the intersection of sets A and B
- $|A \cup B|$ is the size of the union of sets A and B

2.5 Cosine Similarity:

Formula:

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

where,

- cosine_similarity(A, B) is the cosine similarity between vectors A and B
- $A \cdot B$ is the dot product of vectors A and B
- $\|A\|$ and $\|B\|$ are the Euclidean norms of vectors A and B respectively

3 Results

3.1 K-means Clustering

The optimal value of k was found to be 3 using the silhouette coefficient metric. We observe that the silhouette coefficient decreases as the value of k increases. Hence, the optimal value of k is 3. Analysing the change of silhouette coefficient with respect to k is shown below

Value of K	Silhouette Coefficient
3	0.6988985969993166
4	0.6137622384616047
5	0.5641146777848881
6	0.4879572689267768

Table 1: Silhouette Coefficients for Different Values of K

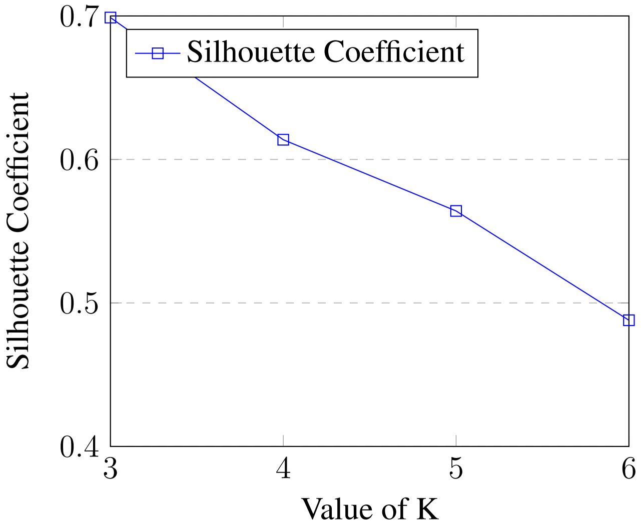

Figure 1: Silhouette Coefficients for Different Values of K

The decrease in silhouette coefficient with increasing k suggests that the clusters become less distinct as more clusters are introduced. This indicates that the data may not naturally divide into more than three clusters, leading to a decrease in the quality of clustering as k increases.

3.2 Complete Linkage Divisive (Top-Down) Clustering Technique

The Silhouette Coefficient for the optimal value of k was found to be 0.6895572393550293.

This high Silhouette Coefficient indicates that the clusters formed by the Complete Linkage Divisive clustering technique are well-separated and dense, with each data point closely resembling its own cluster centroid compared to others. It suggests that the chosen value of k effectively captures the underlying structure of the data. However, further analysis and validation may be necessary to ensure the robustness and generalizability of the clustering results.

3.3 Jaccard Similarity

The Jaccard Similarity between the clusters obtained from K-means and Complete Linkage Divisive (Top-Down) Clustering Technique is shown in the Table below.

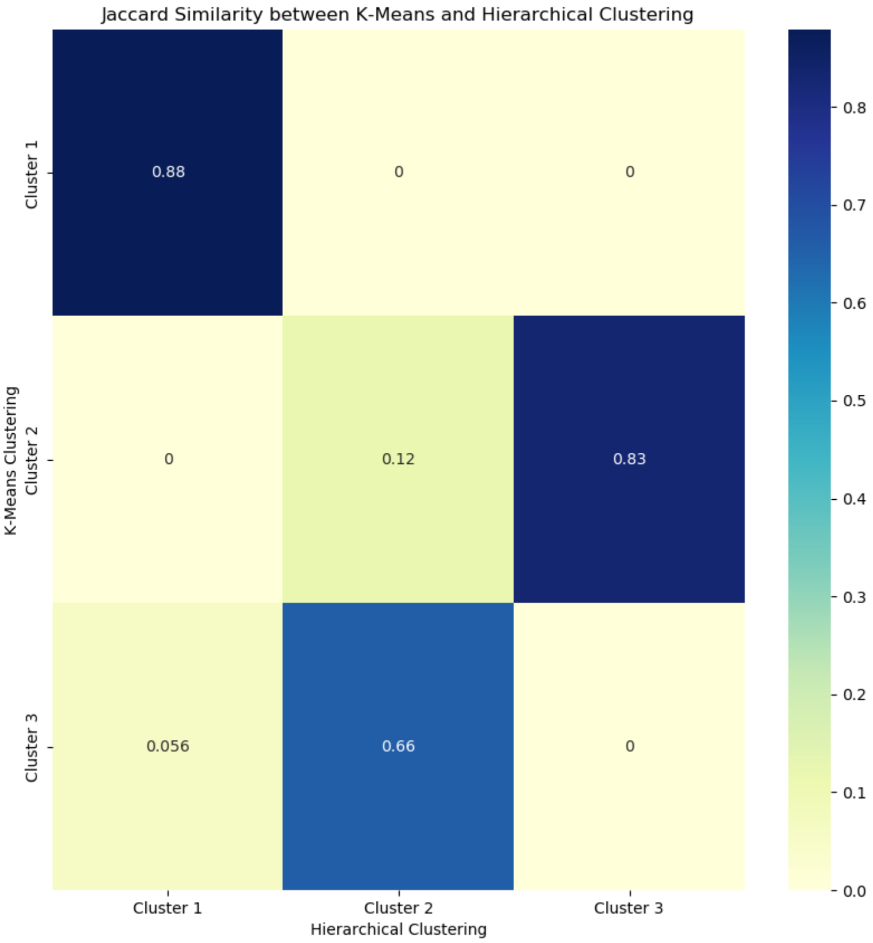

Figure 2: Jaccard Similarity

K-Means Cluster Index	Divisive Cluster Index	Jaccard Similarity
1	1	0.8787878787878788
2	3	0.8315789473684211
3	2	0.6610169491525424

Table 2: Comparison of K-Means and Divisive Clustering Indices with Jaccard Similarity

The Jaccard Similarity measures the similarity between two sets by comparing their intersection to their union. In the context of clustering, it quantifies the similarity of cluster assignments between different clustering techniques.

From Table 2, we observe that the Jaccard Similarity values range between 0.66 and 0.88, indicating moderate to high similarity between the cluster assignments produced by K-means and Complete Linkage Divisive clustering. Higher Jaccard Similarity values suggest a greater agreement between the clustering results of the two techniques.

However, it's essential to note that while Jaccard Similarity provides valuable insights into the agreement between clustering results, it does not account for the structure or quality of the clusters themselves. Therefore, further analysis, such as examining the cluster centroids or silhouette scores, may be necessary to fully assess the clustering performance.

4 Analysis

Table 3 shows the time taken for each step of the process. The "Load Data" step, which involves loading the dataset, took a minimal amount of time, while steps such as "Find Best K" and "Hierarchical Clustering" required more computation.

The total time taken for all steps was **2.0712053775787354** seconds.

This breakdown of time taken for each step provides valuable insights into the computational overhead of the process and can help identify potential bottlenecks or areas for optimization.

Step	**Time Taken (seconds)**
Load Data	0.0031189918518066406
Clustering with K = 3	0.26228976249694824
Clustering with K = 4	0.24265837669372559
Clustering with K = 5	0.25972580909729004
Clustering with K = 6	0.27604103088378906
Find Best K	1.0408053398132324
Hierarchical Clustering	1.0254216194152832
Calculate Jaccard Similarity	0.00032305717468261 72
Total Time Taken	**2.0712053775787354**

Table 3: Time Taken for Each Step

5 Conclusion

In this study, we implemented the K-means clustering algorithm from scratch for $k = 3, 4, 5, 6$ and found the optimal value of k to be 3 using the silhouette coefficient metric. We also implemented the Complete Linkage Divisive (Top-Down) Clustering Technique for the optimal value of k and obtained a high Silhouette Coefficient, indicating well-separated and dense clusters.

Additionally, we established a one-to-one correspondence between the clusters obtained from K-means and Complete Linkage Divisive clustering and calculated the Jaccard Similarity. The moderate to high Jaccard Similarity values suggested a reasonable agreement between the cluster assignments produced by the two techniques.

Furthermore, the breakdown of time taken for each step provided valuable insights into the computational overhead of the process. Steps such as finding the best k and performing hierarchical clustering contributed significantly to the total computation time.

Overall, this analysis sheds light on the effectiveness and efficiency of different clustering techniques in partitioning data and provides a foundation for further exploration and optimization in future studies.