CS60050 – MACHINE LEARNING

PROJECT – 3 REPORT

Economic Outlook Categorization using Complete Linkage Divisive (Top-Down) Clustering Technique

Submission by:

Cheruvu Surya Sai Ram (19EC39008)

Under the guidance of

Prof. Aritra Hazra

and

TA Suryansh Kumar

Indian Institute of Technology Kharagpur

Preprocessing the Dataset :

- The dataset contains 368 instances with 14 features.
- The dataset contains **some NULL** values. They are removed accordingly.
- Then there is an attribute in the dataset which has **only one unique value** which is also removed since it would have no effect on clustering.
- There exists **two pairs of attributes** in the dataset which are equivalent i.e. there is a **one to one mapping** between them. One of the attributes in each pair is removed.
- We would remain with 2 categorical attributes and 9 continuous attributes. The **categorical attributes are converted into numerical** and then **standardization** is applied on the whole modified dataset.
- The outlier removal in the dataset is left to the discretion of the user. It is not done here but a **separate function is provided for the same** in case one wants to remove them before applying clustering algorithms.

Key Points in Implementation :

- It is asked to consider **cosine similarity** as a **distance measure**. The equivalent distance form is obtained as **1 – similarity** since distance should be minimum if two feature vectors are similar.
- The **K-Means clustering algorithm** is run for **20 iterations** using random initialization of distinct centroids. In every iteration, the points are clustered and the centroids are updated. However the seed is set in the code.
- The **clustering information is saved in the format** as mentioned in the problem statement and the **Silhouette Coefficient** values are calculated. The average Silhouette Score is obtained by taking the mean over all the points.
- **Divisive Clustering** is performed using **Complete Linkage**. Initially all points are chosen within the same cluster. Then the cluster with the farthest points is chosen and split accordingly taking those points as the centers. This is done until the required number of clusters is obtained.
- Clusters of K-Means and Divisive clustering are obtained for the optimal value of k. They are **mapped** and the corresponding **Jaccard Similarity scores** are obtained.

Results, Observations, Justifications and Interpretation:

i) K Means clustering algorithm

Optimal number of clusters obtained are 5 considering **Silhouette Score** as the evaluation metric. Silhouette Scores obtained for K-means clustering with different values of K are **tabulated below**.

K-values	Silhouette Score
3	0.533
4	0.639
5	0.701
6	0.664

In K-Means clustering, the overall **best Silhouette Score** was obtained as **0.701** for **5 clusters**. So we say that **we obtained the best clusters for k = 5** since the inter cluster distances are maximized and intra cluster distances are minimized. The K-Means algorithm can be run for more iterations until centroid convergence is achieved and an improvement in Silhouette Scores can be expected.

However, **the outputs vary across runs and depends upon the centroid initialization for K Means**. But our program always outputs the same values since we set a random seed. On changing the random seed, we would observe different outputs in different runs. The K-Means outputs widely depend on centroid initialization. **If two centroids are initialized in the same cluster**, they usually divide an existing cluster into halves resulting in sub-optimal clustering performance.

ii) Complete Linkage Divisive Clustering algorithm

Silhouette Score for Top Down Hierarchical (Divisive) Clustering algorithm using complete linkage strategy with **5 clusters** was obtained as **0.659.**

The silhouette score obtained using the optimal value of k **turned out to be almost the same** for both K-Means and Divisive Clustering. **This shows that** both the clustering algorithms are equally good in terms of clustering performance. Since K-Means clusters depend on initialization of centroids, it could yield a better match on choosing better centroids.

Silhouette coefficient for each data point is defined as

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

where a and b are the average distance from data-points belonging to the same cluster and the closest neighboring cluster respectively. Ideally, b >> a, so S is approximately equal to 1. We obtained best average silhouette coefficient of 0.701 with K-Means for k=5, **thus demonstrating good clustering performance**.

The score is bounded **between -1 for incorrect clustering and +1 for highly dense clustering**. **Scores around zero** indicate overlapping clusters. **Larger value of Silhouette Coefficient denotes that** clusters are denser and well-separated in adherence to the idea of clustering algorithms.

iii) Jaccard Similarity

As instructed in the question, **clusters obtained from K-Means are mapped with those obtained from Divisive Clustering** for K=5. The **corresponding Jaccard Similarity scores** are obtained. They are **tabulated** as follows.

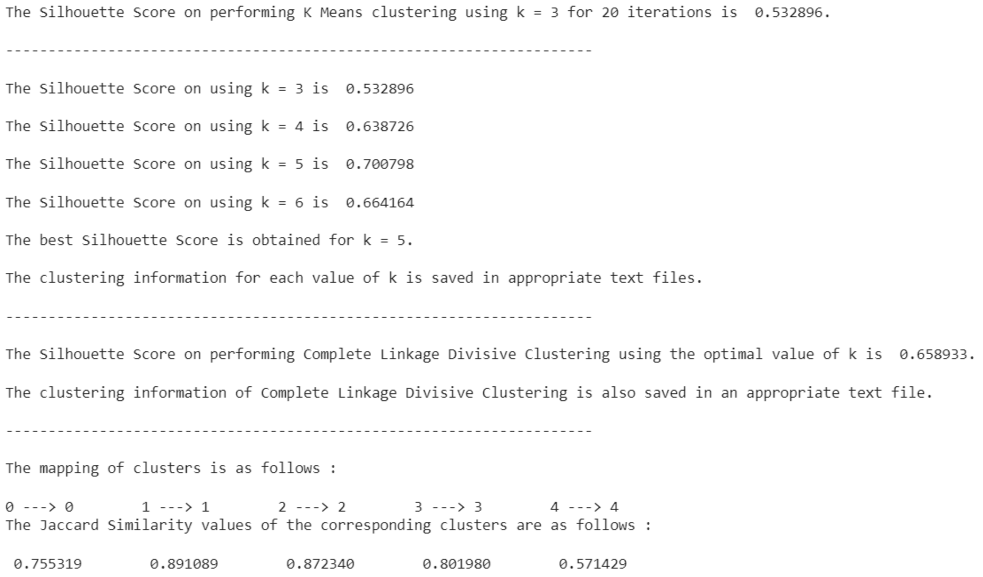
Cluster Mapping and Jaccard Similarity Score -

K-means cluster id	Divisive Cluster Id	Jaccard Similarity
0	0	0.755
1	1	0.891
2	2	0.872
3	3	0.802
4	4	0.571

The **Jaccard similarity scores for two clusters is defined as** the size of the intersection over the size of their union. Since the clusters obtained for divisive clustering almost matched with those obtained from K-Means for k=5, we could observe **good enough values of Jaccard similarity**. **This can also be justified by** looking at the cluster information in the corresponding text files.

Results can also be checked and compared after removing outliers from the dataset as per the user’s wish.

Below is a screenshot of **the output of the program file** after running.



```
The Silhouette Score on performing K Means clustering using k = 3 for 20 iterations is  0.532896.

----------------------------------------------------------------------

The Silhouette Score on using k = 3 is  0.532896

The Silhouette Score on using k = 4 is  0.638726

The Silhouette Score on using k = 5 is  0.700798

The Silhouette Score on using k = 6 is  0.664164

The best Silhouette Score is obtained for k = 5.

The clustering information for each value of k is saved in appropriate text files.

----------------------------------------------------------------------

The Silhouette Score on performing Complete Linkage Divisive Clustering using the optimal value of k is  0.658933.

The clustering information of Complete Linkage Divisive Clustering is also saved in an appropriate text file.

----------------------------------------------------------------------

The mapping of clusters is as follows :

0 ---> 0        1 ---> 1        2 ---> 2        3 ---> 3        4 ---> 4
The Jaccard Similarity values of the corresponding clusters are as follows :

 0.755319        0.891089        0.872340        0.801980        0.571429
```

Approximate Execution Time :

The approximate time taken by the program to run all the steps in a reasonable PC configuration is **30 seconds**.