Interest Match using Complete Linkage Divisive (Top-Down) Clustering Technique

Ishaan Sinha (21CS30064)

1.) Introduction

Clustering is a fundamental task in machine learning and data analysis, aiming to group similar data points together based on certain features or attributes. In this project, we are tasked with clustering individuals based on their interests, with the objective of forming social clusters. This task serves as a preliminary step towards understanding social dynamics and behavior patterns among individuals.

2.) Dataset

The dataset provided contains information about the interests of individuals, with features such as Music, Movies, Politics, Economy-Management, Reading, Cars, Religion, Dancing, Writing, Active-Sport, Theatre, and Pets. We are required to perform clustering using K-means algorithm with cosine similarity as the distance measure, and evaluate the results using the Silhouette coefficient metric.

3.) Methods and Experiments

a. **K-means clustering:** In this step, we employ the K-means algorithm to partition the dataset into k=3 clusters based on cosine similarity as the distance measure. We randomly initialize k cluster means as distinct data points and iterate for 20 iterations to converge on stable clusters. The clustering information is then saved in a file for further analysis.

b. **Evaluation of the clustering algorithm:** The Silhouette coefficient metric is utilized to evaluate the quality of the clusters obtained from the K-means algorithm. This metric measures the mean distance between a sample and all other points in the same cluster (a), and the mean distance between a sample and all other points in the next nearest cluster (b). By calculating the Silhouette coefficient for each sample and taking the mean, we obtain a score that indicates the density and separation of clusters.

c. **Find optimal value of k:** To determine the optimal number of clusters, we repeat steps 1 and 2 for k=4, 5, and 6, and compute the Silhouette coefficient for each value of k. The value of k that yields the highest Silhouette coefficient is considered as the optimal number of clusters.

d. **Hierarchical clustering:** In this step, we implement a Complete Linkage Divisive (Top-Down) Clustering algorithm using the same notion of similarity as in step 1. We find k clusters, where k is the optimal number of clusters obtained from step 3, using the complete linkage strategy. We then compare the clustering results obtained from K-means and hierarchical clustering by computing the Jaccard similarity between corresponding sets of clusters. This involves mapping each set of clusters from K-means to a distinct set of clusters from hierarchical clustering, considering the Jaccard similarity, and printing the Jaccard similarity scores for all the k mappings.

4.) Results

Silhouette Coefficient:
For k = 3, Silhouette Coefficient: -0.1466375193058558
For k = 4, Silhouette Coefficient: -0.09645624017887024
For k = 5, Silhouette Coefficient: -0.061490565007007995
For k = 6, Silhouette Coefficient: -0.024205641207828095

Jaccard Similarity:
Jaccard Similarity for 1st cluster is 0.2346002621231979
Jaccard Similarity for 2nd cluster is 0.175
Jaccard Similarity for 3rd cluster is 0.1507537688442211
Jaccard Similarity for 4th cluster is 0.16133333333333333
Jaccard Similarity for 5th cluster is 0.2248995983935743
Jaccard Similarity for 6th cluster is 0.16339869281045752

Optimal number of clusters (k): **6** with Silhouette Coefficient: **-0.024205641207828095**

410.71196603775024 seconds runtime

5.) Conclusion

In this project, I embarked on a journey to explore clustering techniques for grouping individuals based on their interests. Beginning with the K-means clustering algorithm, I successfully partitioned the dataset into clusters and evaluated the clustering quality using the Silhouette coefficient metric. This allowed me to understand the density and separation of clusters, aiding in the determination of the optimal number of clusters.

Through experimentation with different values of k, I identified the optimal number of clusters that best captures the underlying structure of the data. Leveraging this optimal

k-value, I proceeded to implement a hierarchical clustering algorithm, specifically the Complete Linkage Divisive approach, to obtain an alternative clustering solution.

Finally, I compared the clustering results obtained from both K-means and hierarchical clustering using the Jaccard similarity metric. This comparison provided insights into the similarity and dissimilarity between the clustering solutions, shedding light on the robustness and effectiveness of the clustering techniques employed.

Overall, this project highlights the importance of clustering algorithms in uncovering patterns and structures within data, paving the way for deeper insights into social dynamics and behavior patterns among individuals. Further exploration and refinement of clustering techniques could lead to enhanced understanding and application in various domains, ranging from social sciences to marketing and beyond.