[Project Code: PMHC-AS]

Preferences Match using Single Linkage Agglomerative (Bottom-Up) Clustering Technique

Roll no: 23AT61R04, Student name: Thummalabavi Sankshay reddy

---

Introduction: This report presents the results and analysis of a clustering experiment conducted on a dataset of land mines using Kmeans and Hierarchical Clustering algorithms. The objective was to identify the optimal number of clusters and assess the similarity between clusters obtained from the two algorithms.

Data Loading and Preprocessing: The dataset containing information about preferences was loaded from the file "preferences.csv". The dataset was preprocessed to prepare it for clustering.

Optimal Number of Clusters:

- The Silhouette Coefficient was computed for different values of k in the K-means algorithm.
- For k = 3, the Silhouette Coefficient is 0.1971,
  For k =4, the Silhouette Coefficient is 0.1555,
  For k = 5, the Silhouette Coefficient is 0.075,
  For k =6, the Silhouette Coefficient is 0.1142,

- **Therefore, the optimal number of clusters using the K-means algorithm is 3**.

Hierarchical Clustering:

- Hierarchical clustering was performed using single linkage Agglomerative (Bottom -Up) Clustering Technique with cosine similarity as the distance metric.
- The dataset was divided into 3 clusters using hierarchical clustering.

Similarity Analysis: Jaccard Similarity was computed between corresponding sets of clusters obtained from K-means and hierarchical clustering.

- For Cluster 1: Jaccard Similarity: 0.44020
- For Cluster 2: Jaccard Similarity: 0.2097
- For Cluster 3: Jaccard Similarity: 0.349

  The Jaccard Similarity scores indicate moderate similarity between clusters obtained from K-means and hierarchical clustering

Execution Time:

- The approximate time taken to execute kmeans 20.5629 seconds
- Hierarchical clustering took considerable time compared to K-means due to its iterative nature and the computation of pairwise distances.
  Exe time = 1551.269 seconds

Conclusion:

- The K-means algorithm with an optimal value of k=3 achieved a higher Silhouette Coefficient, suggesting better clustering quality compared to other values of k.
- Hierarchical clustering provided insights into the hierarchical structure of the data but required more time to execute.

```
Silhouette coefficient for k=3: 0.19719720387920467
Silhouette coefficient for k=4: 0.15554642703275448
Silhouette coefficient for k=5: 0.07588426821027972
Silhouette coefficient for k=6: 0.11424331517518536
Optimal value of k: 3 (highest Silhouette Coefficient)
```

```
Time taken for K-means: 20.562980890274048 seconds
```

```
Time taken for Agglomerative: 1551.26975107193 seconds
Jaccard Similarity : [0.4402061855670103, 0.2097107438016529, 0.34917355371900827]
```

This report summarizes the findings of the clustering analysis, providing insights into the optimal number of clusters, clustering methods used, similarity analysis, execution time, and recommendations based on the results

The clusters are in .txt files submitted with this report