Report on Interest Match using Complete Linkage Agglomerative (Bottom-Up) Clustering

[Project Code: IMHC-AC]

Paramananda Bhaskar
21CS30035

Introduction:

The objective of this project is to form clusters of similar people based on their interests using clustering algorithms. Specifically, we will be performing K-means clustering and Complete Linkage Agglomerative (Bottom-Up) Clustering. The dataset consists of information about individuals' interests in various categories such as Music, Movies, Politics, etc. Our goal is to identify the optimal number of clusters and evaluate the clustering algorithms using the Silhouette coefficient metric and Jaccard similarity.

Methodology:

The project was divided into the following steps:

Data Preprocessing: The dataset "interests.csv" containing **approximately 1000 instances and 12 features** was loaded and preprocessed. Missing values were handled, and the data was normalized.

K-means Clustering:

- K-means clustering was performed with **k=3** clusters **using cosine similarity as the distance measure**.
- Random initialization of cluster centroids was done as distinct data points.
- The algorithm was iterated for **20 iterations** to optimize the cluster assignments.

Evaluation of K-means Clustering:

- **The Silhouette coefficient metric** was computed to evaluate the quality of clustering.
- The Silhouette coefficient measures the density and separation of clusters, with values ranging from -1 to 1.

- A higher Silhouette coefficient indicates better clustering performance.

Finding Optimal Value of K:

- K-means clustering was repeated for **k=4, 5, and 6 clusters**.
- The Silhouette coefficient was calculated for each value of k.
- The optimal number of clusters was determined based on the highest Silhouette coefficient obtained.
- The optimal clustering information was saved in the file "kmeans.txt".

Hierarchical Clustering:

- **Complete Linkage Agglomerative (Bottom-Up) Clustering** was implemented using the same cosine similarity measure as K-means.
- The optimal number of clusters obtained from the previous step was used.
- Cluster assignments were saved in the file "agglomerative.txt".

Computing Jaccard Similarity:

- **Jaccard similarity** was computed between corresponding sets of clusters **obtained from K-means and Hierarchical clustering**.
- Each set of clusters from K-means was mapped to a distinct set of clusters from Hierarchical clustering based on Jaccard similarity.
- Jaccard similarity scores for all mappings were printed.

Results:

- Optimal Number of Clusters: **The optimal number of clusters was determined to be 3** based on the highest Silhouette coefficient.
- Silhouette Coefficient: The **Silhouette coefficient for the optimal number of clusters k=3** was **0.19713028527954768**.
- Jaccard Similarity: Jaccard similarity scores for mappings between K-means and Hierarchical clustering are as follows:

  Jaccard similarity for cluster 1: 0.42452830188679247

  Jaccard similarity for cluster 2: 0.1695095948827292

  Jaccard similarity for cluster 3: 0.3705263157894737

Discussion:

- The evaluation of clustering algorithms using the Silhouette coefficient helps in assessing the quality of cluster assignments.
- Jaccard similarity provides insights into the similarity between clusters obtained from different clustering techniques.
- The optimal number of clusters ensures that the data is appropriately segmented into meaningful groups based on similarity in interests.

Analysis:

Optimal Number of Clusters (k=3):

- The choice of 3 clusters appears to be appropriate based on the Silhouette coefficient, indicating that the clustering algorithm successfully identified meaningful groups within the dataset.
- With three clusters, the data is segmented into distinct groups, making it easier to analyze and understand.

Silhouette Coefficient:

- The Silhouette coefficient of approximately 0.20 indicates moderate clustering quality. While it's not exceptionally high, it still suggests that the clusters are reasonably well-defined.
- However, the presence of overlapping clusters (indicated by the coefficient being close to 0) suggests that there may be some ambiguity in the clustering results.

Jaccard Similarity:

- The Jaccard similarity scores provide insights into the similarity between clusters obtained from K-means and Hierarchical clustering.
- Cluster 1 has the highest Jaccard similarity, indicating that it is relatively consistent across both clustering methods.
- Cluster 2 has the lowest similarity, suggesting that there are significant differences in how this cluster is identified by the two algorithms.
- Cluster 3 shows moderate similarity, indicating some consistency but also some divergence in cluster assignments.

Conclusion:

- The project successfully implemented K-means clustering and Hierarchical clustering to form clusters of similar individuals based on their interests.

- The evaluation metrics helped in determining the optimal number of clusters and assessing the quality of clustering algorithms.
- Jaccard similarity analysis provided additional insights into the consistency of clustering results across different techniques.

Future Scopes:

- The results suggest that while the clustering algorithms were able to identify meaningful clusters, there is room for improvement, particularly in terms of reducing overlap and increasing consistency between different clustering methods.
- Further analysis and fine-tuning of parameters may lead to improved clustering results and better separation between clusters.
- Overall, the clustering analysis provides valuable insights into the structure of the dataset and can inform decision-making processes in various applications such as targeted marketing, recommendation systems, and social network analysis.

Computational Time:

- **The approximate time taken** by the program **to run all steps** on a reasonable PC configuration was **around 1 minute**, considering dataset size and algorithm complexity.