CSCI4150U: Data Mining

K-Means and Hierarchical Clustering

Syed Naqvi
Student ID: 100590852

November 10, 2024

Abstract

This report analyzes K-means and Hierarchical clustering on the Breast Cancer Wisconsin and Waveform datasets, using Sum of Squared Errors (SSE) as a metric and the elbow method for k-means model selection. Hierarchical clustering methods (Single, Complete, and Group Average Link) are visualized using dendrograms. Findings show K-means effectively identifies natural groupings in the data and Hierarchical clustering confirms these groupings by revealing a significantly reduced rate of inter-cluster distance reductions beyond optimal partitioning. The report highlights the effectiveness of K-means and Hierarchical clustering algorithms in finding natural data clusters.

1 Introduction

1.1 Methodology

Datasets used in this analysis are sourced from the *UC Irvine Machine Learning Repository* and include:

- **Breast Cancer Wisconsin (Diagnostic)**
- **Waveform Database Generator (Version 1)**

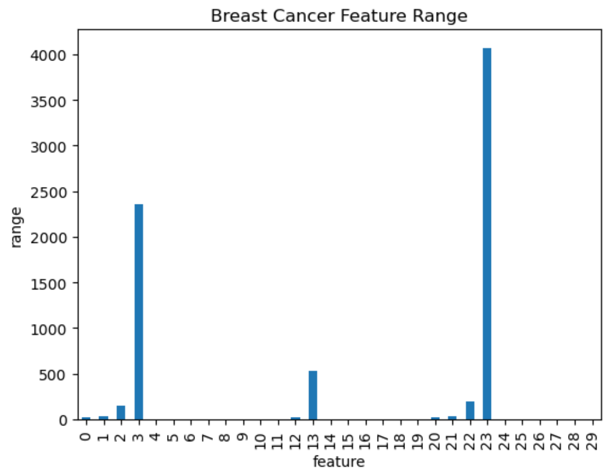
We evaluate the following clustering algorithms:

- **K-means Clustering**
- **Hierarchical Clustering** (Single Link, Complete Link, and Group Average)

For K-means clustering, model k values range from 1 to 6 for the *Breast Cancer Wisconsin (Diagnostic)* dataset and 2 to 6 for the *Waveform Database Generator (Version 1)* dataset. Clustering performance is assessed using the Sum of Squared Errors (SSE) with Euclidean distance as the metric.

1.2 Preprocessing

To perform accurate clustering, we analyze feature ranges to determine the dataset most needing of standardization.



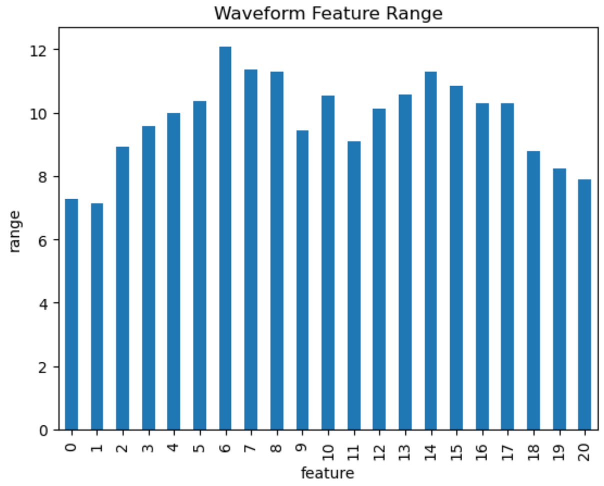
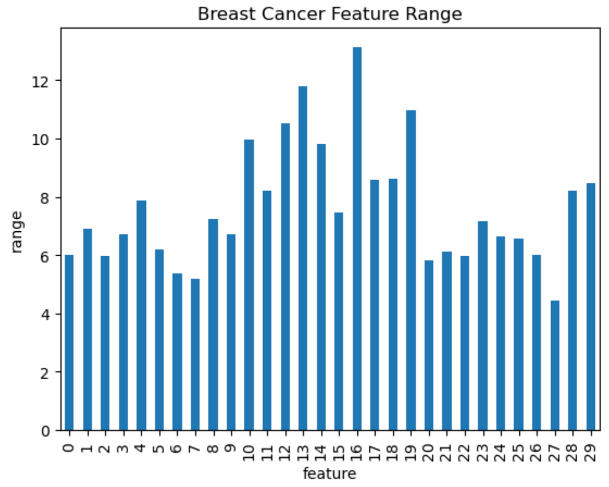


Figure 1: Pre-Standardized Feature Ranges (Breast Cancer Data)

Figure 2: Pre-Standardized Feature Ranges (Waveform Data)



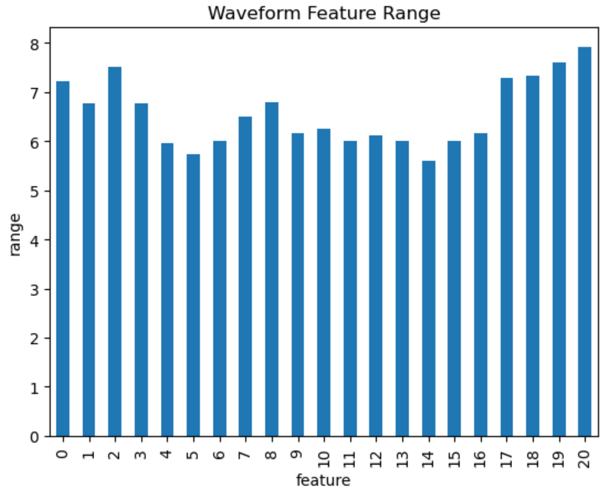


Figure 3: Post-Standardized Feature Ranges (Breast Cancer Data)

Figure 4: Post-Standardized Feature Ranges (Waveform Data)

The feature values have now been scaled and are significantly better suitable for clustering.

2 Part I: K-Means Clustering

2.1 Model Selection

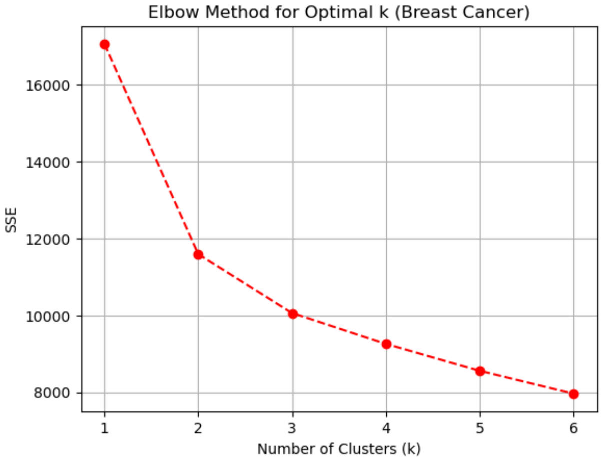


Figure 5: Elbow method for selecting optimal K-value (Breast Cancer Dataset)

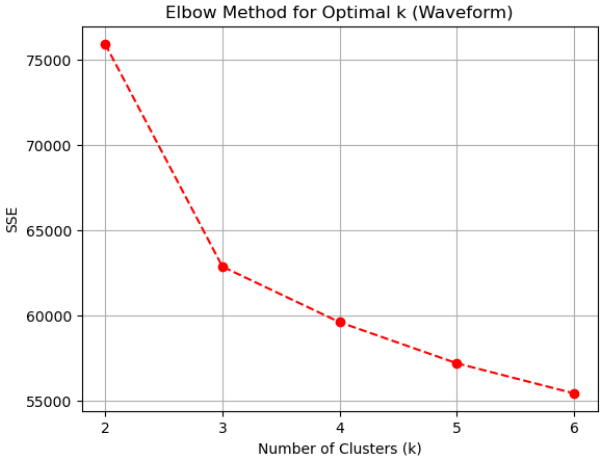
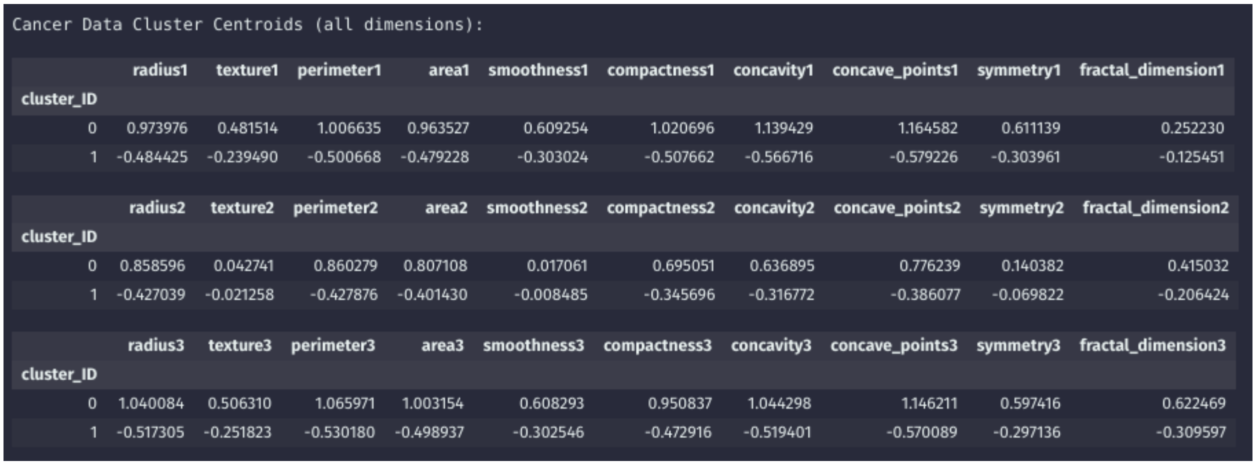


Figure 6: Elbow method for selecting optimal K-value (Waveform Dataset)

From the above figures, the ideal number of clusters is **k=2** for the Breast Cancer dataset and **k=3** for the Waveform dataset. This suggests there are likely 3 distinct waveforms in the Waveform dataset and two distinct tumor categorizations (malignant or benign) in the Breast Cancer dataset which is of course consistent with the actual number of unique labels in both datasets.

2.2 Results

Using the ideal cluster amounts for each dataset, the following tables display centroids with lowest SSE scores achieved through convergence with a maximum of 300 iterations and 10 rounds of unique seed placement.



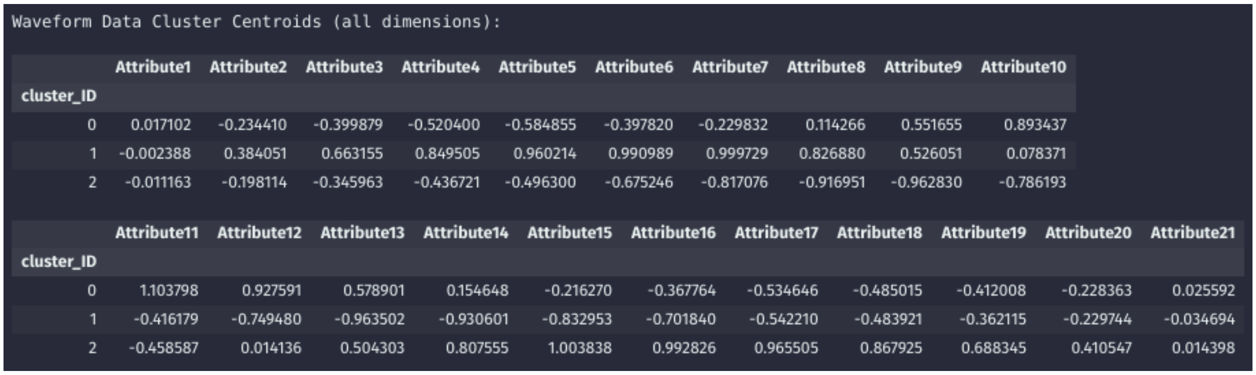
```
Cancer Data Cluster Centroids (all dimensions):
```

	radius1	texture1	perimeter1	area1	smoothness1	compactness1	concavity1	concave_points1	symmetry1	fractal_dimension1
cluster_ID										
0	0.973976	0.481514	1.006635	0.963527	0.609254	1.020696	1.139429	1.164582	0.611139	0.252230
1	-0.484425	-0.239490	-0.500668	-0.479228	-0.303024	-0.507662	-0.566716	-0.579226	-0.303961	-0.125451

	radius2	texture2	perimeter2	area2	smoothness2	compactness2	concavity2	concave_points2	symmetry2	fractal_dimension2
cluster_ID										
0	0.858596	0.042741	0.860279	0.807108	0.017061	0.695051	0.636895	0.776239	0.140382	0.415032
1	-0.427039	-0.021258	-0.427876	-0.401430	-0.008485	-0.345696	-0.316772	-0.386077	-0.069822	-0.206424

	radius3	texture3	perimeter3	area3	smoothness3	compactness3	concavity3	concave_points3	symmetry3	fractal_dimension3
cluster_ID										
0	1.040084	0.506310	1.065971	1.003154	0.608293	0.950837	1.044298	1.146211	0.597416	0.622469
1	-0.517305	-0.251823	-0.530180	-0.498937	-0.302546	-0.472916	-0.519401	-0.570089	-0.297136	-0.309597

Figure 7: Centroids (Breast Cancer Dataset)



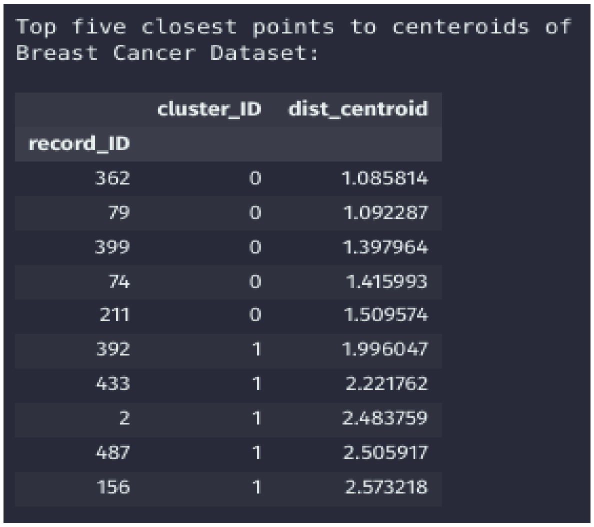
Waveform Data Cluster Centroids (all dimensions):

	Attribute1	Attribute2	Attribute3	Attribute4	Attribute5	Attribute6	Attribute7	Attribute8	Attribute9	Attribute10
cluster_ID										
0	0.017102	-0.234410	-0.399879	-0.520400	-0.584855	-0.397820	-0.229832	0.114266	0.551655	0.893437
1	-0.002388	0.384051	0.663155	0.849505	0.960214	0.990989	0.999729	0.826880	0.526051	0.078371
2	-0.011163	-0.198114	-0.345963	-0.436721	-0.496300	-0.675246	-0.817076	-0.916951	-0.962830	-0.786193

	Attribute11	Attribute12	Attribute13	Attribute14	Attribute15	Attribute16	Attribute17	Attribute18	Attribute19	Attribute20	Attribute21
cluster_ID											
0	1.103798	0.927591	0.578901	0.154648	-0.216270	-0.367764	-0.534646	-0.485015	-0.412008	-0.228363	0.025592
1	-0.416179	-0.749480	-0.963502	-0.930601	-0.832953	-0.701840	-0.542210	-0.483921	-0.362115	-0.229744	-0.034694
2	-0.458587	0.014136	0.504303	0.807555	1.003838	0.992826	0.965505	0.867925	0.688345	0.410547	0.014398

Figure 8: Centroids (Waveform Dataset)

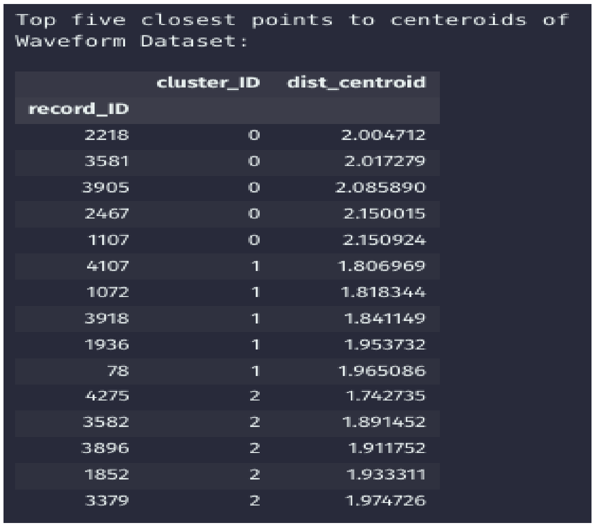
Using Euclidean distance as our metric, we can view the top five closest points in each cluster to their centroid across both datasets.



Top five closest points to centeroids of Breast Cancer Dataset:

	cluster_ID	dist_centroid
record_ID		
362	0	1.085814
79	0	1.092287
399	0	1.397964
74	0	1.415993
211	0	1.509574
392	1	1.996047
433	1	2.221762
2	1	2.483759
487	1	2.505917
156	1	2.573218

Figure 9: Top five closest points to their cluster centroids (Breast Cancer Dataset)



Top five closest points to centeroids of Waveform Dataset:

	cluster_ID	dist_centroid
record_ID		
2218	0	2.004712
3581	0	2.017279
3905	0	2.085890
2467	0	2.150015
1107	0	2.150924
4107	1	1.806969
1072	1	1.818344
3918	1	1.841149
1936	1	1.953732
78	1	1.965086
4275	2	1.742735
3582	2	1.891452
3896	2	1.911752
1852	2	1.933311
3379	2	1.974726

Figure 10: Top five closest points to their cluster centroids (Waveform Dataset)

3 Part II: Clustering Task

Next, we use dendrograms to visualize the Single Link, Complete Link and Group Average forms of Hierarchical clustering. We truncate the graphs above 30 separate clusters and visualize when the various distance metrics seems to converge.

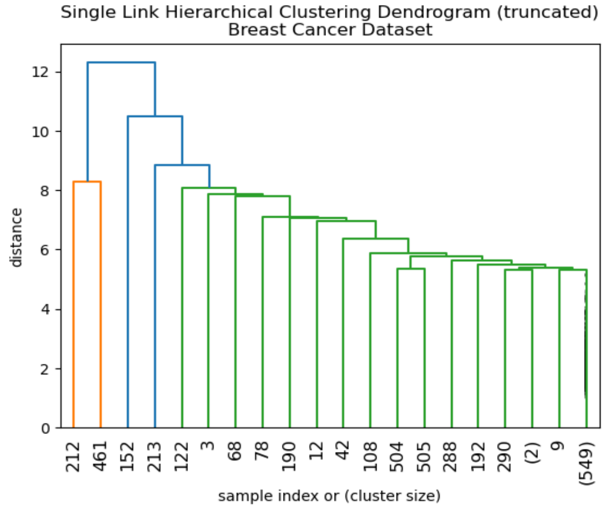


Figure 11: Single Link Hierarchical Clustering (Breast Cancer Dataset)

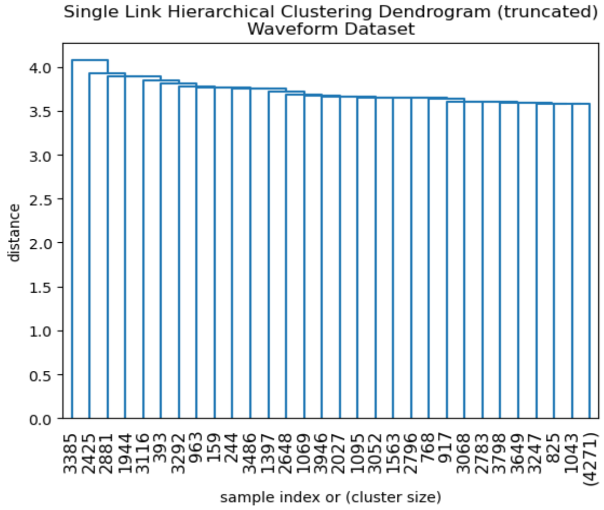


Figure 12: Single Link Hierarchical Clustering (Waveform Dataset)

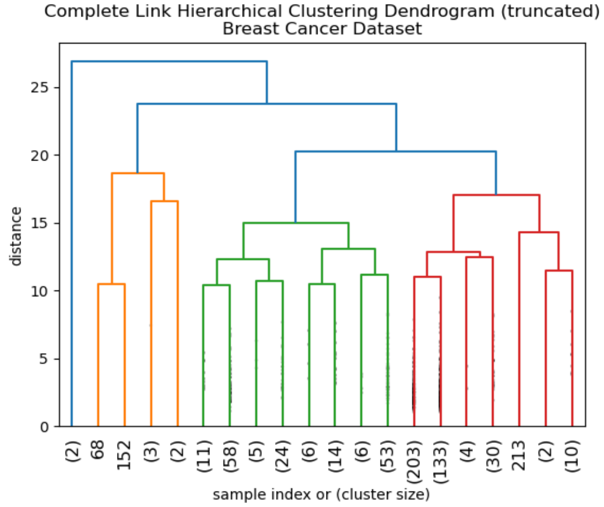


Figure 13: Complete Link Hierarchical Clustering (Breast Cancer Dataset)

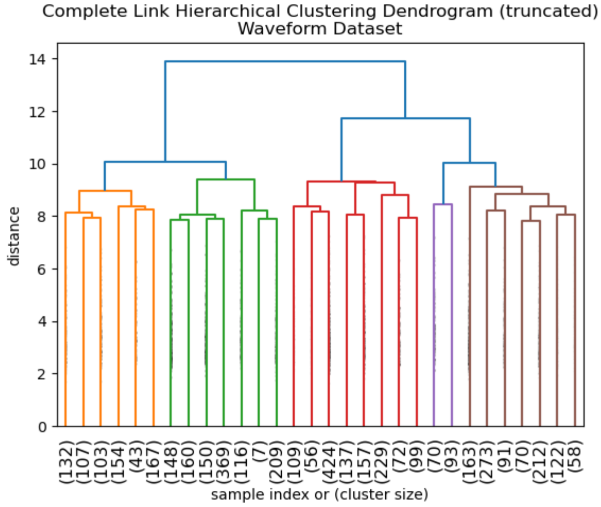


Figure 14: Complete Link Hierarchical Clustering (Waveform Dataset)

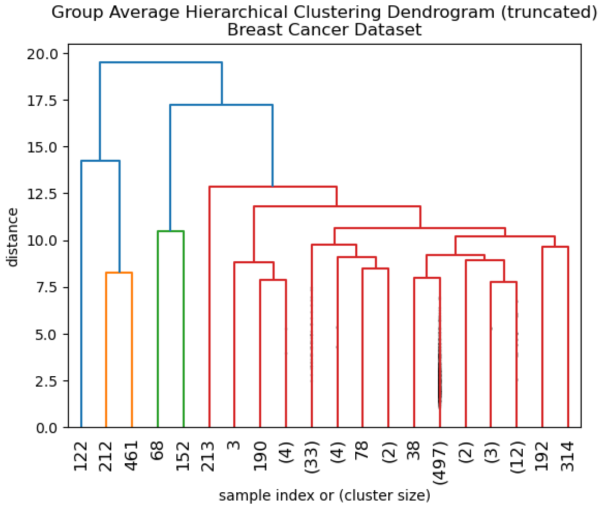


Figure 15: Group Average Hierarchical Clustering (Breast Cancer Dataset)

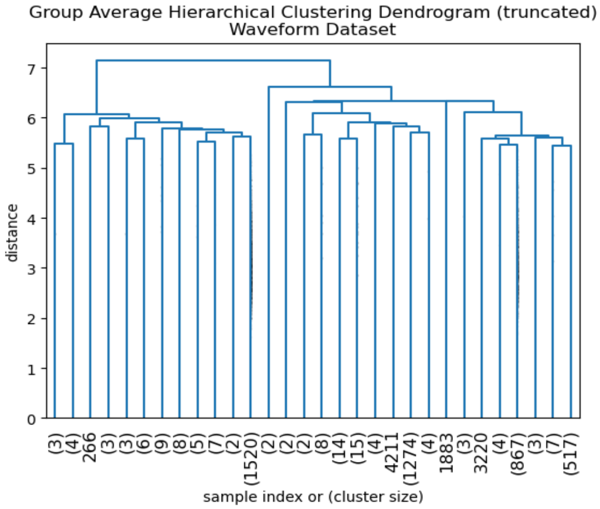


Figure 16: Group Average Hierarchical Clustering (Waveform Dataset)

4 Conclusion

This report demonstrates that both K-means and hierarchical clustering methods effectively identify natural groupings given relatively simpler and small datasets. Using Sum of Squared Errors (SSE) and the ”elbow method” for optimal k-selection, the K-means algorithm can effectively identify natural groupings in the data while the Single Link, Complete Link and Group Average Hierarchical clustering variants further corroborate these groupings through dendrogram visualizations. These visualizations reveal significantly reduced *rate of reduction* in inter-cluster distance beyond a certain level of partitioning, indicating a natural stopping point for clustering. Overall, the analysis highlights the complementary strengths of K-means and hierarchical clustering in data mining tasks.