NAME: COVENANT OLUDAPOMOLA OJO

COURSE: DATA MINING AND DISCOVERY

STUDENT ID: 23013682

PROTOTYPE-BASED CLUSTERING AND K-MEANS CLUSTERING

INTRODUCTION

Prototype-based clustering is a fundamental technique in unsupervised learning, organizing data into groups using prototypes such as centroids (average positions) or medoids (representative data points). Each data point is assigned to the cluster with the nearest prototype, measured by metrics like Euclidean distance (Zhou et al., 2024). Its simplicity and efficiency make it applicable in fields like customer segmentation, document categorization, and image compression.

This report focuses on prototype-based clustering with an emphasis on K-means clustering. Using a real-world dataset, the performance of K-means is compared to Spectral Clustering to evaluate their capabilities and limitations.

METHODOLOGY

Overview of Prototype-Based Clustering The

clustering process follows these steps:

- Initialization: Prototypes are chosen either randomly or using systematic heuristics.
- Assignment: Data points are allocated to the nearest prototype.
- Update: Prototypes are recalculated based on the data points assigned to them.
- Termination: The process stops when assignments stabilize or meet a predefined condition.

K-MEANS CLUSTERING ALGORITHM

K-means is an iterative clustering method aimed at reducing intra-cluster variability. The process begins with k randomly selected initial centroids. Data points are then assigned to the nearest centroid, and the centroids are recalculated based on the current cluster members. These steps repeat until no significant changes occur in centroid positions.

The algorithm minimizes the Sum of Squared Errors (SSE), defined as:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \| (x - C_i) \|^2 =$$

DATASET AND PREPROCESSING

The analysis utilized the UCI Stock Keeping Units dataset. Preprocessing steps included data normalization and managing missing values. Key features used for clustering were pallet weight and height, enabling intuitive visualizations and performance assessment.

RESULTS AND DISCUSSION

Selecting the Number of Clusters

The optimal number of clusters (k) was identified as three using the Elbow Method, which balances model complexity and intra-cluster variance.

PERFORMANCE COMPARISON

The clustering results were analyzed using the Silhouette Score:

- K-Means: Silhouette Score: 0.264926415303179, suggesting distinct and well-separated clusters.

Works efficiently with spherical cluster shapes but shows vulnerability to outliers. Using columns for visualization: Pal gross weight, Pal height

GAUSSIAN MIXTURE MODEL (GMM)

- GMM: Silhouette score: 0.14349941357640825

GMM calculates the probability that a point belongs to each cluster and assigns it accordingly.

OBSERVATIONS AND CHALLENGES

Initialization Sensitivity: Random initialization in K-means occasionally led to suboptimal clusters. Strategies like K-means++ can address this issue by improving the starting centroids.

Impact of Outliers: The reliance on mean centroids caused distortions in the presence of outliers, affecting cluster quality.

CLUSTER SIZES

- K-Means: 1330, 947,2.
- Spectral Clustering: 1274, 825,2.

PERFORMANCE SUMMARY

K-Means outperformed GMM slightly, achieving a higher Silhouette Score and clearer cluster separations. However, GMM demonstrated greater flexibility in handling clusters with different shapes and sizes.

CONCLUSION

Prototype-based clustering methods, particularly K-Means and GMM, offer a balance of simplicity and efficiency for large datasets. K-Means works best for well-separated spherical clusters, while GMM handles overlapping or non-spherical clusters better. Spectral Clustering, although not compared directly, showed potential for non-linear separability. Future research should explore hybrid approaches combining these methods to improve clustering performance. Additionally, careful parameter tuning can address initialization sensitivity and outlier impacts. Overall, both K-Means and GMM provide valuable insights into data patterns depending on the use case.

. **REFERENCES**

1. Zhou, S., Zhang, P. & Chen, H. (2024) *Latent Prototype-Based Clustering: A Novel Exploratory Electroencephalography Analysis Approach.* Sensors, 24(15). Available at: https://www.mdpi.com/1424-8220/24/15/4920 .
2. Tan, P.-N., Steinbach, M. & Kumar, V. (2018) *Introduction to Data Mining.* 2nd edn. Pearson.
3. UCI Machine Learning Repository (n.d.) *Stock Keeping Units Dataset.* Available at: https://archive.ics.uci.edu/dataset/531/stock+keeping+units .
4. UCI Machine Learning Repository (2019) *Stock keeping units* [Dataset]. Available at: https://doi.org/10.24432/C5CG7S .