A Comparative Study of Prototype Based Clustering: K-Medoids Clustering Vs K-Means Clustering

Name: Sandra Binu | Student ID: 22029960

Introduction:

Clustering algorithms are essential tools for data analysis and pattern recognition, especially in fields like marketing and retail. In this report, we present a comparative analysis of two prototype clustering algorithms, K-Medoids and K-Means, applied to a Sales Transaction Weekly dataset. The goal is to determine which algorithm yields better clustering results and insights for understanding product sales behaviour.

Data Description:

The Sales Transaction Weekly (UCI) dataset comprises weekly purchased quantities of 800 products over 52 weeks, with each row representing a unique product and each column representing sales for a specific week. Additionally, normalization features are included for each week. The dataset contains no missing values, facilitating a comprehensive analysis.

Methodology:

1. Data Preprocessing:
   - The dataset is split into features and normalization features
   - Outlier detection is performed using boxplots to ensure data quality.
   - The data is normalized using Min-Max scaling to ensure uniformity across features.
   - Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset to two dimensions, facilitating visualization and interpretation of results.
2. Clustering: K value (number of clusters) is found as 3 using Elbow method.
   A. **K-Medoids Clustering:** K-Medoids clustering (PAM, Partition around medoids) is a partitioning algorithm that, similar to K-Means, groups data points into 'K' clusters, but employs actual data points as cluster centres (medoids) to enhance robustness against outliers. Figure 1 shows the cluster plot of K-Medoids clustering on the chosen dataset.

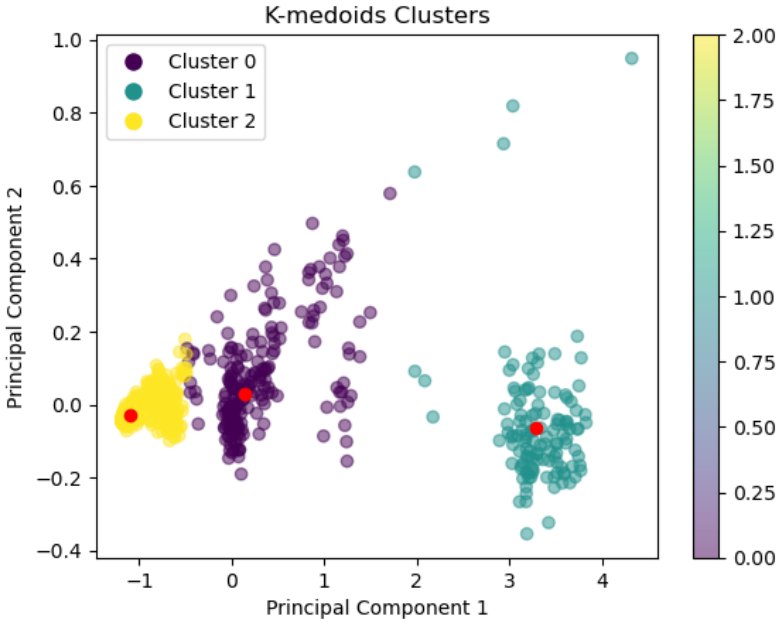

Figure 1

   B. **K-Means Clustering**: K-Means clustering is an unsupervised machine learning algorithm that divides data into 'K' separate, non-overlapping clusters through minimization of within-cluster variance. Figure 2 shows the cluster plot obtained from K-Means clustering on chosen dataset.

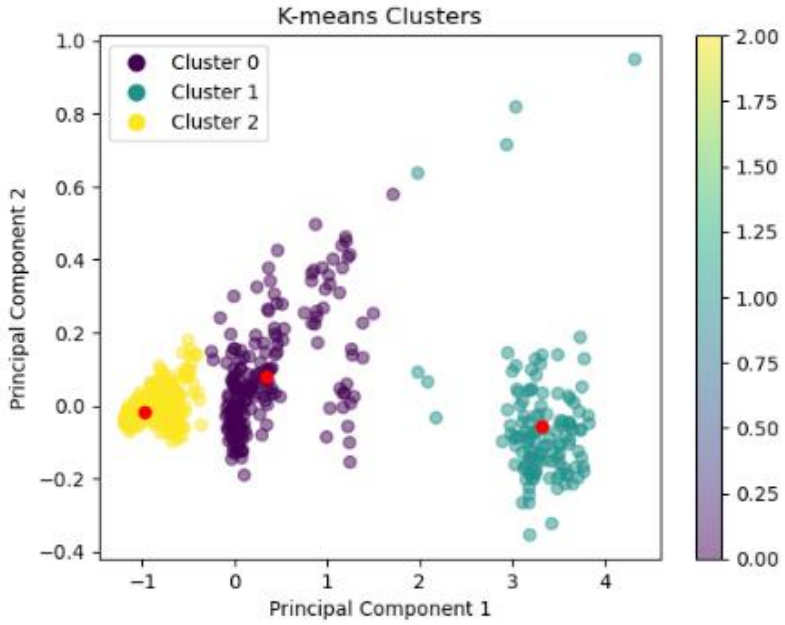

Figure 2

Comparison and Results Analysis:

- Since K-Means minimises the sum of squared distances to centroids, it is susceptible to outlier-induced shifting whereas K-Medoids use real data points as medoids, they are less affected by outliers and have less of an impact on centroid locations and cluster assignments.
- K-Means scales well for large datasets and frequently produces clusters with comparable sizes and shapes, making interpretation and visualisation simple. In contrast, K-Medoids can create a wider range of cluster sizes and shapes, but they may be harder to understand and less scalable because pairwise distances must be calculated.

Algorithms	K-Medoids	K-Means
Silhouette Score	0.7458	0.7529
Davies–Bouldin Index	0.3797	0.3756
Execution Time	0.0363 sec	0.0232 sec

Table 1

- Silhouette Score: In comparison to K-Medoids (0.7458), K-Means has a marginally higher silhouette score (0.7529), meaning that data points within clusters are, on average, closer to one another and farther from points in other clusters.
- Davies-Bouldin Index: Compared to K-Medoids (0.3797), K-Means has a slightly lower Davies-Bouldin Index (0.3756), indicating better cluster separation and coherence.
- Execution Time: K-Means demonstrated a slightly faster clustering time, completing in approximately 0.023 seconds, while K-Medoids took slightly longer at approximately 0.036 seconds.

Conclusion:

In conclusion, K-Means clustering seems to produce marginally better results than K-Medoids in terms of silhouette score, Davies–Bouldin Index, and computational efficiency based on the examination of the Sales Transaction Weekly dataset. Although K-medoids provide robustness against noise and outliers, large datasets may find it difficult to handle their computational cost and complexity. Though there aren't many differences between the two algorithms, choosing one over the other should take certain needs and dataset features into account. All things considered, both algorithms provide insightful information about the buying habits of their clients, supporting companies in making calculated decisions and focusing their marketing efforts.

References:

- Tan, P, Steinbach, M, Kumar, V, & Karpatne, A 2019, Introduction to Data Mining EBook: Global Edition, Pearson Education, Limited, Harlow. Available from: ProQuest Ebook Central. [24 April 2024].
- https://www.irjet.net/archives/V6/i3/IRJET-V6I3154.pdf
- https://www.sciencedirect.com/science/article/pii/S1877050916000971
- https://medium.com/@prasanNH/exploring-the-world-of-clustering-k-means-vs-k-medoids-f648ea738508

- **Data Set**: https://archive.ics.uci.edu/dataset/396/Sales+Transactions+Dataset+Weekly
- **GitHub Link**: https://github.com/sandrabinu3/Data-Mining/blob/main/K-Means_Vs_K-Medoids.ipynb