**Prototype-based Clustering: Fuzzy C-Means Vs K-Means on Sales Data**

Name: Srikavya Vasala Student ID:23115572

Introduction:

Clustering is an unsupervised learning technique, that groups similar data points. In this report, we will dive into prototype-based clustering on the Sales Transactions weekly dataset. Two prototype based approaches: K-Means and Fuzzy C-Means are used to evaluate the dataset and compare the methods on dataset.

The comparison focuses on clustering performance, interpretability, and potential business insights for the dataset.

Data Preprocessing

The Sales Transaction weekly dataset is taken from UCI. The dataset contains weekly purchased quantities of 800 products over 52 weeks. Initial preprocessing to ensure robust clustering outcomes was: handling duplicate values, standardizing features with z-score normalization, and detecting and handling outliers. To preserve data integrity outliers were removed and replaced.

Clustering

Prototype-based clustering (also known as centroid-based clustering) is an important clustering subcategory which characterizes each cluster through a prototype and employs a relocation scheme to iteratively redistribute the data instances into the clusters, guided by an objective function.

The two Prominent prototype-based clustering algorithms are brought into focus: K-means and Fuzzy C-means.

K-means clustering is a widely used algorithm that aims to partition the data into k clusters based on similarity based on instances. The elbow method is used to determine optimal clusters by plotting the sum of squared errors(SSE) against various k values. K=2 was selected from the elbow method.

To understand the clustering structure, PCA and t-SNE were applied to reduce dimensionality of the data and project it onto 2D space. Figure 1 shows the resulting projections with cluster centroid overlaid to provide insights into clustering boundaries.

Fuzzy C-Means Clustering is an algorithm run by fuzzy logic, FCM works with fuzzy instances of data points to multiple clusters. Each data point has a degree of membership in each cluster, rather than being assigned to just one. Each data point’s membership in a cluster is calculated based on its distance to centroid, Raised by power m(fuzziness factor). In the figure 1 with t-SNE with FCM shows the membership degrees and FPC from the metrics.

Evaluation Metrics:

|  |  |  |
| --- | --- | --- |
| Metric | KMeans FCM |  |
| Silhouette Score | 0.597 0.597 |  |
| Davies-Bouldin Index | 0.737 0.738 |  |
| Fuzzy Partition Coefficient | - 0.842 |  |

* From the evaluation metrics, Both K-Means and FCM performed equally well in terms of overall cluster quality.
* However silhouette scores suggest better clustering with an average score.
* Fuzzy C-Means offer additional flexibility, and the high FPC value (0.842**)** shows it handled overlapping or borderline cases well.

Figure 1

t-SNE representation of K-means and FCM on weekly sales dataset


The models can suggest business insights and inventory management, Cluster 0 contains consistently sold products while cluster 1 suggests product seasonal sales accordingly.

**K-Means Clustering** provides clear, definitive clusters, making it ideal for segmenting products into distinct categories based on sales patterns. Unlike K-Means, FCM does not force hard boundaries between clusters, and products can belong to multiple clusters with different degrees of membership. This is useful in scenarios where products have overlapping characteristics. For example, some products may be both seasonal and consistently demanded, and FCM can capture this overlapping nature.

Conclusion:

Both clustering algorithms reveal meaningful clusters and structures in weekly product sales.

While K-means is computationally efficient and interpretable, fuzzy c-means offers greater efficiency.

In the sales data, based on metrics FCM outperforms K-Means, with the added advantage of partial membership scores.

References:

1. Dataset: https://archive.ics.uci.edu/dataset/396/sales+transactions+dataset+weekly
2. Introduction to Data Mining by Tan et al (Second edition) is Chapter  
   8 Section 2 (pp. 641-657).
3. https://ieeexplore.ieee.org/document/10154917