International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Unsupervised learning Algorithm in Clustering: A Comparison of Hierarchical and K-means

[1]Dr.V.Maniraj
Associate professor,Research Supervisor,
Head of the Department, Department of Computer Science, A.V.V.M Sri pushpam College (Autonomous), Poondi, Thanjavur (Dt),
Affiliated to Bharathidasan University, Thiruchirapalli, Tamilnadu,

[2]P. Akshaya M.Sc., Computer Science

Department of Computer Science, A.V.V.M Sri pushpam College (Autonomous), Poondi, Thanjavur (Dt), Affiliated to Bharathidasan University, Thiruchirapalli, Tamilnadu,

ABSTRACT:

Unsupervised learning algorithms play a crucial role in discovering hidden patterns and structures within the data This paper delves into two prominent clustering approaches K-means and Hierarchical clustering. Evaluating their performance, strengths and weakness and their methodology and their process. The results highlight the strength of Hierarchical clustering in identifying complex clusters and k-means in handling well separated clusters. This study provides the insights for choosing the suitable algorithm for specific clustering tasks.

Keywords: Machine learning, Unsupervised learning, Clustering, K-means, Hierarchical.

INTRODUCTION:

Machine learning is a subset of Artificial Intelligence (AI) that involves training algorithms to learn from data and make predictions. Where Unsupervised learning is a type of machine learning algorithm can learn patterns, relationships, or groupings in data without prior knowledge of the expected output. Unsupervised learning is a way to automatically discover hidden patterns or structure in data without human guidance. Unsupervised learning is like giving a puzzle to a machine without showing it the completed picture. The machine then analyzes the pieces(data), figuring out how they might fit together or group based on their similarities, all on its own. It doesn't have a guide to fallow, but instead, uncovers hidden patterns, relationships, structures in the data.

WORKFLOW DIAGRAM FOR UNSUPERVISED LEARNING :

International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

UNSUPERVISED LEARNIG ALGORITHMS:

HIERARCHICAL CLUSTERING:

clustering is the most established subcategory of UL where hierarchical clustering is one of its types. Hierarchical clustering also known as hierarchical cluster analysis, is an algorithm is used to group similar objects into groups called clusters. Hierarchical clustering can build a hierarchy of clusters by merging or splitting existing clusters. Represented visually as a dendrogram.

Key concepts:

Distance metric: Determines the similarity or dissimilarity between two data points or clusters common metrics include:

- Euclidean distance
- Manhattan distance
- Cosine similarity
- Hamming distance

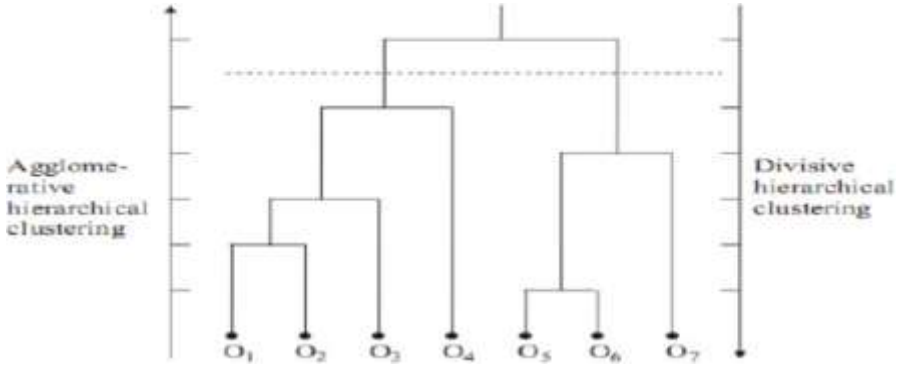
Linkage criterion: Defines how the distance between two clusters is calculated based on the distance between their individual members. Common linkage criteria include:

- Single Linkage
- Complete Linkage
- Average Linkage
- Centroid Linkage

Types of Hierarchical clustering:

1. **Agglomerative Clustering (Bottom-up):** This is the most common type of hierarchical clustering that can starts with each data point as an individual cluster. In each iteration, the two closed clusters are merged into a single cluster. This process continues until all data points are merged into a single large cluster.

2. **Divisive clustering (Top-down):** This is also one of the types of hierarchical clustering that can begins with all data points in a single large cluster. In each iteration, a cluster is divided into two smaller clusters. This process continues until each data point from its own individual cluster.



Strengths of Hierarchical Clustering:

- No need for pre-defined number of clusters
- The resulting dendrogram provides a clear visual hierarchical structure
- Can handle various data types and distance measures

Weakness of Hierarchical clustering:

- Can be computationally expensive
- Can be memory intensive for large datasets
- Not well-suited for dynamic data streams

K-Means clustering:

K-means clustering is the most popular unsupervised learning algorithm used to partition a dataset into a pre-defined number of clusters. It aims to group similar data points together and discover underlying patterns or structures within the data. K-means algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

How k-means works

1.Initialization:

Choose the number of clusters (k)

Randomly select k data points as initial centroids

2. Assignment:

Assign each data point to the nearest centroid

The distance between a data point and a centroid is typically calculated using Euclidean distance, but other distance metrics can also be used.

3. Update Centroids:

Recalculate the centroid of each cluster the new centroid is the mean of all data points assigned to that cluster.

4. **Repeat**:

Repeat steps 2 and 3 until the centroids no longer move significantly or a maximum number of iterations is reached.

Key Concepts

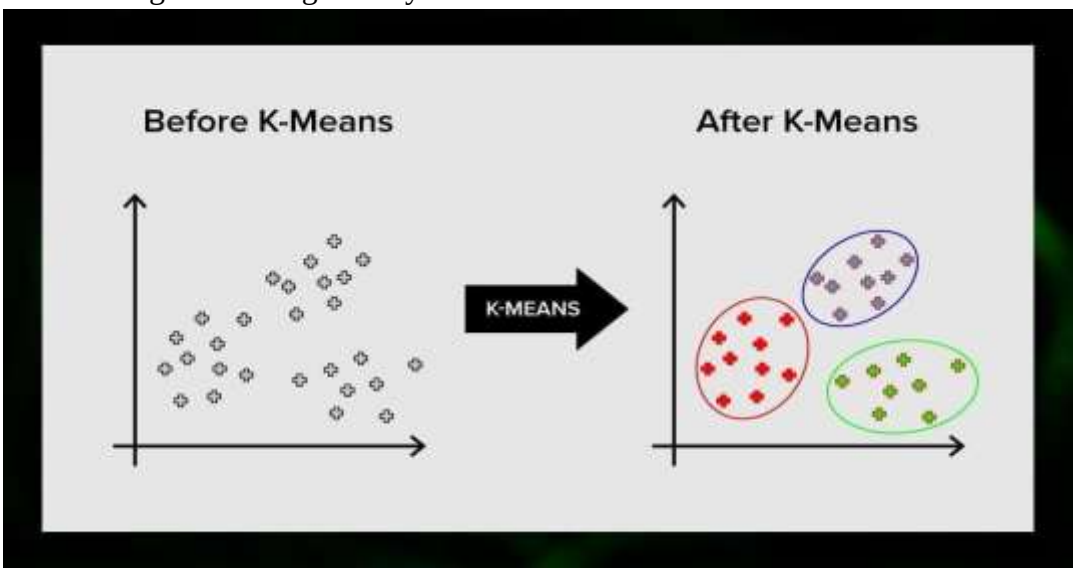
Centroid: The center point of a cluster, representing the average of all data points within a cluster.

Within-Cluster Sum of Squares (WCSS): A measure of how tightly clustered the data points are around their respective centroids. K-means aims to minimize the WCSS.

Choosing the Number of Clusters (k)

Elbow method: Plot the WCSS against different values of k. The "elbow" point, where the curve starts to bend, often indicates a good choice for k.

Silhouette Score: Measures how similar a data point is to its own cluster compared to other clusters. Higher scores generally indicate better-defined clusters.



Dataset for clustering:

S.NO	PLAYER NAME	AGE	GENDER	INTEREST
1	RAHUL	21	M-0	Football
2	JOHN	22	M-0	Cricket
3	KAMALI	18	F-1	Cricket
4	ANITHA	16	F-1	Football
5	GURU	25	M-0	Cricket
6	SAI	26	M-0	Football
7	SATHISH	27	M-0	Cricket
8	VISHNU	19	M-0	Football
9	ASWIN	15	M-0	Cricket
10	JANANI	29	F-0	Cricket

K=2 (randomly take two vectors)….V1, v2 Consider that the vectors. V1=(22,0) and v2=(18,1) for the data points for the players as a1,a2,a3,….,a10. By using the formula of Euclidean distance **d = √[(x2 - x1)2 + (y2 - y1)2]**

Finally, the output of the clustering V1,

Datapoints	Assigned Center	Interest
a1(21,0)	V1	Cricket
a2(22,0)	V1	Cricket
a5(25,0)	V1	Cricket
a6(26,1)	V1	Cricket
a7(27,0)	V1	Cricket
a10(29,1)	V1	Cricket

The output of the clustering V2,

Datapoint	Assigned Center	Interest
a3(18,1)	V2	Football
a4(16,1)	V2	Football
a8(19,0)	V2	Football
a9(15,0)	V2	Football

Finally, by clustering the players, we got the players interested on the Cricket more than Football.

Strengths of K-means clustering:

- k-means is a relatively simple to understand and implement
- k-means is effective for clustering spherical clusters, where the clusters are roughly equal in size and density
- k-means can handle high-dimensional data, making suitable for clustering datasets with many features.

Weakness of K-means clustering:

- k-means struggles with non-linear clusters, such as clusters with varying densities or irregular shapes
- k-means does not handle missing values, which can lead to poor clustering results.
- k-means is sensitive to initial placement of centroids, which can affect the final clustering results.

Conclusion:

This study has provided a comprehensive comparison of Unsupervised learning algorithms such as k-means and Hierarchical clustering, both this popular Unsupervised algorithm have their strengths and weaknesses, among them k-means is comparatively faster and efficient than Hierarchical for handling large datasets with high accuracy. By the results of the player dataset the k-means clustering has the ability to handling fast and

efficient for complex datasets for unsupervised learning in Artificial Intelligence (AI).

References:

1. ”A survey on Unsupervised Learning for Big data” by S.M.A. sharif et al. (2022)
2. “Unsupervised machine Learning for clustering and dimensionality reduction” by J. Zhao et al. (2022)
3. “Comparative study of Unsupervised Learning Algorithms for Anomaly detection” by A.S.S. Rao et al. (2022)
4. “Unsupervised clustering with Deep Neural Networks” by Aljalbout et al. (2021)
5. “Unsupervised Domain Adaptation with Adversarial Training” by Li et al(2020)
6. “A comparative study of Unsupervised Learning Algorithms for clustering and Dimensionality reduction” by S.K. Goyal, S. Sharma, and A.K. pujari et al (2018)
7. “A comparative Analysis of Unsupervised Learning Algorithms” by A.K. Jain et al (2018)