Agglomerative Clustering

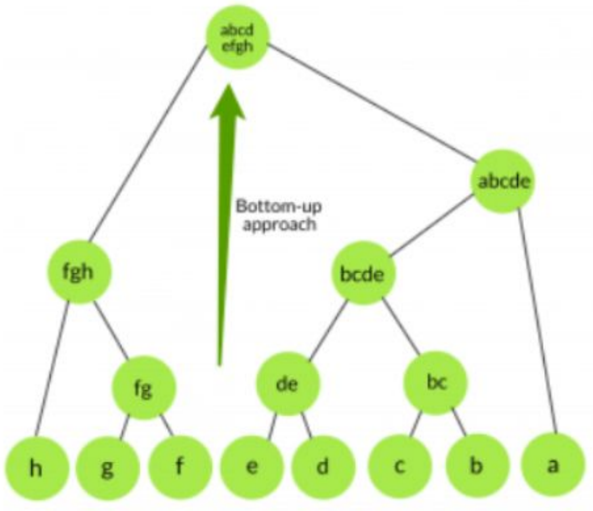
Group 02

210146N - Dissanayake D.M.P.S.
210471F - Perera S.A.I.M.
210518H - Ranasinghe K.S.
210554M - Sajeev Kugarajah
210639E - Theesan L.M.
210647C - Thevinka M.A.D.
210404F - Nanayakkara A.H.M.
210372D - Manawathilake K.C.K.

Introduction

- Agglomerative Clustering is a hierarchical algorithm that uses a bottom-up approach.
- Each data point is initially considered as a cluster.
- The algorithm proceeds by iteratively merges the most similar clusters based on a chosen distance metric, forming larger clusters. This process continues until a stopping criterion is met, resulting in a dendrogram that illustrates the hierarchical relationships between clusters.

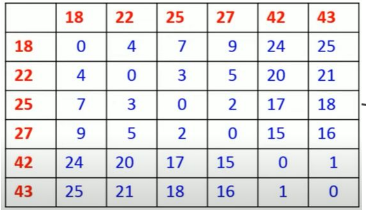
Agglomerative Clustering



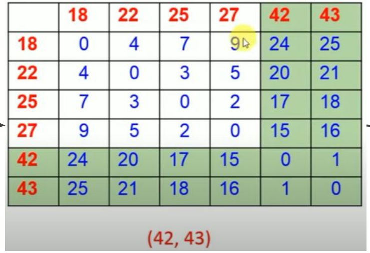
Hierarchical agglomerative clustering

Algorithmic Overview

- A bottom-up hierarchical clustering algorithm
  - First we'll build the proximity matrix and merge the clusters according to their minimum distance.
  - Then using that we'll build our dendrogram.

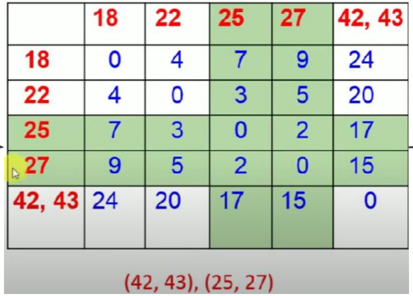


	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0



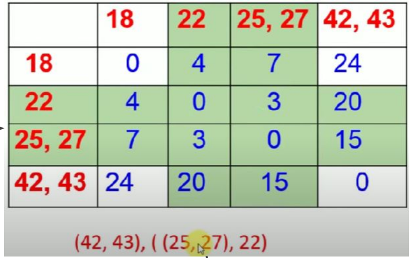
	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

(42, 43)



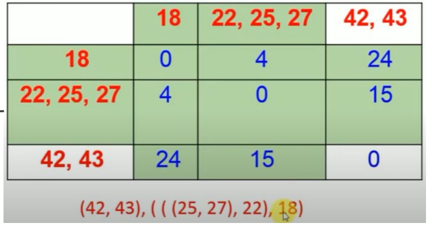
	18	22	25	27	42, 43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42, 43	24	20	17	15	0

(42, 43), (25, 27)



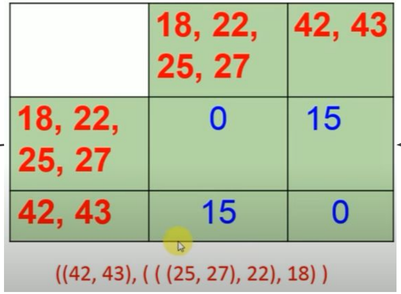
	18	22	25, 27	42, 43
18	0	4	7	24
22	4	0	3	20
25, 27	7	3	0	15
42, 43	24	20	15	0

(42, 43), ((25, 27), 22)



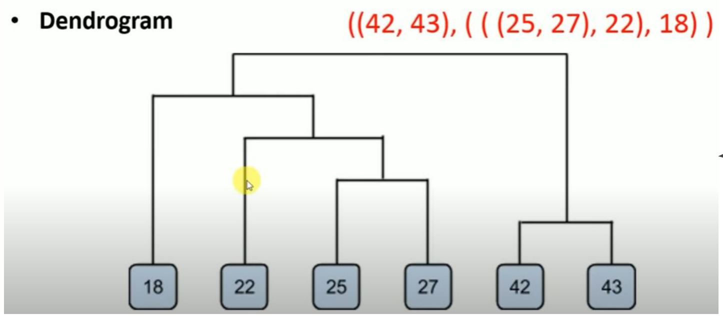
	18	22, 25, 27	42, 43
18	0	4	24
22, 25, 27	4	0	15
42, 43	24	15	0

(42, 43), (((25, 27), 22), 18)



	18, 22, 25, 27	42, 43
18, 22, 25, 27	0	15
42, 43	15	0

((42, 43), (((25, 27), 22), 18))



- **Dendrogram**

((42, 43), (((25, 27), 22), 18))

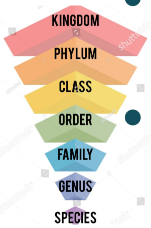
Source : https://www.youtube.com/watch?v=YH0r47m0kFM

Interpreting Dendrograms

- **Cluster Merging Visualization**: Dendrograms provide a visual representation of how clusters merge throughout the hierarchical clustering process.
- **Vertical Height Interpretation**: The vertical height on the dendrogram indicates the distance between clusters. Lower height signifies higher similarity between clusters.
- **Similarity Measure**: Objects or clusters merging at lower heights demonstrate greater similarity compared to those merging at higher heights.
- **Relative Heights**: Focus primarily on the relative heights rather than the absolute values, as the y-axis doesn't hold specific quantitative meaning.
- **Clustering Insights**: Analyzing dendrograms aids in understanding the underlying structure of the data and determining the optimal number of clusters based on the patterns of cluster merging.

Examples and Applications

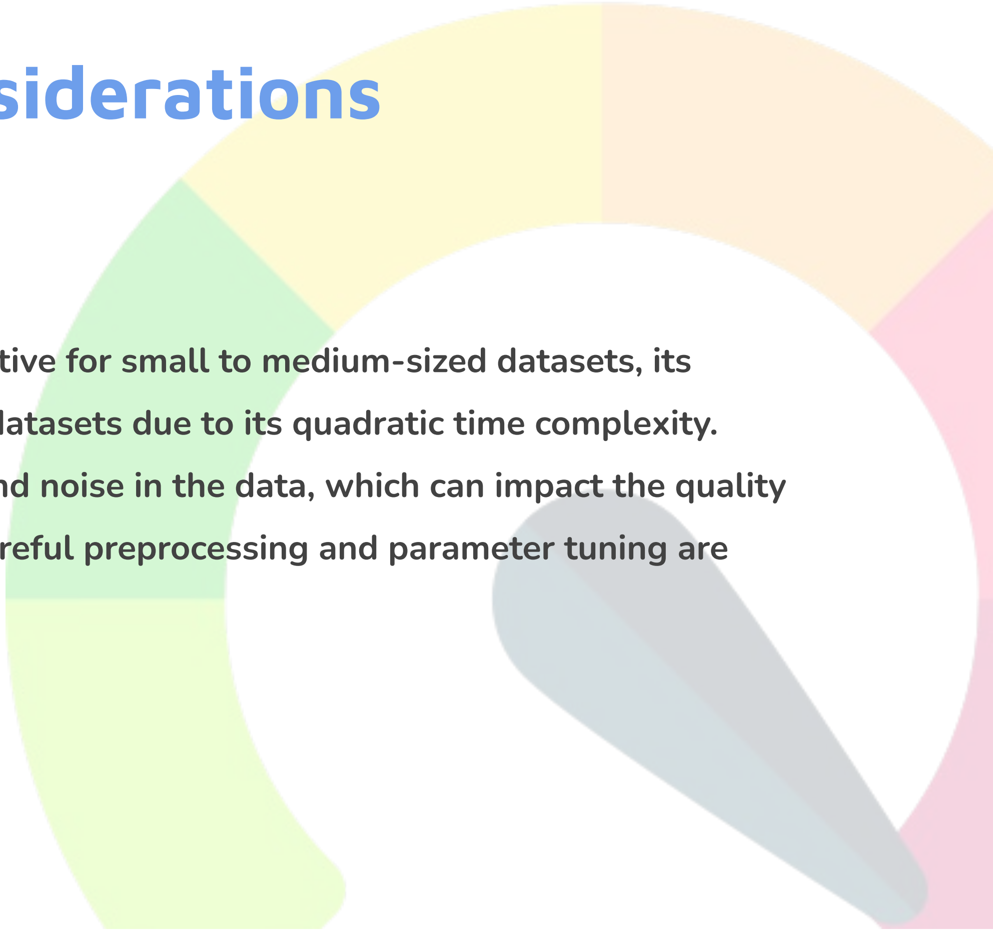
- **Customer Segmentation** - Different client segments, such as regular buyers, infrequent shoppers, budget-conscious buyers, etc., can be identified by the clustering analysis.
- **Image Segmentation** - For instance, segmenting distinct organs or tissues from an MRI or CT scan can help with diagnosis and therapy planning in medical imaging.



- **Biological Taxonomy** - Researchers can learn more about species distributions, evolutionary links, and ecosystem dynamics by locating groups of closely related species.
- **Social Network Analysis** - Social network clustering can make hidden patterns and structures in the network, like influencers, interest groups, and buddy circles, visible.
- **Fraud Detection** - Transactions that depart from normal behaviour, such as abnormally big transactions, transactions from unknown places, or transactions that happen at strange hours, can be identified using clustering analysis.

Performance Considerations

While agglomerative clustering is effective for small to medium-sized datasets, its performance may degrade with larger datasets due to its quadratic time complexity. Additionally, it is sensitive to outliers and noise in the data, which can impact the quality of the clusters generated. Therefore, careful preprocessing and parameter tuning are essential for optimal results.



Comparative Analysis

Compared to divisive clustering methods, such as k-means, agglomerative clustering is more intuitive and requires fewer assumptions about the underlying data distribution. However, it may struggle with non-globular clusters and is computationally more expensive.

Best Practices and Conclusion

In conclusion, agglomerative clustering offers a powerful framework for exploring the structure of complex datasets and uncovering meaningful insights. To maximize its effectiveness, practitioners should carefully consider the choice of linkage criterion, handle outliers appropriately, and validate the resulting clusters using domain knowledge or external evaluation metrics. With these best practices, agglomerative clustering can be a valuable tool for data analysis and decision-making.

Thank You