Proceedings of CONF-SPML 2027 Symposium: Intelligent Network Security and Machine Learning for Communication Systems
DOI: 10.54254/2755-2721/2027.MELB38502

From "Birds of a Feather" to Data Insights: The Principles, Methods and Applications of Hierarchical Clustering Algorithms

Lexuan Yang

Overseas Study Service Center of Central China Normal University, Wuhan, China
yang.lexuan@outlook.com

Abstract. Cluster analysis aims to group similar objects into the same cluster. Hierarchical clustering, as an important method, does not produce a single grouping result but instead constructs a clustering tree that simultaneously presents the grouping structure of the data at different levels of granularity. This paper focuses on hierarchical clustering algorithms. Using a literature review approach, it outlines the basic principles of hierarchical clustering, its two construction approaches, the algorithmic workflow exemplified by AGNES, and linking methods such as single linkage, complete linkage, average linkage, and Ward's method. It also reviews their applications in fields such as bioinformatics, image segmentation, text clustering, and business analytics. The results indicate that hierarchical clustering does not require a predefined number of clusters and offers high interpretability, however, it is computationally complex, its merging decisions are irreversible, and it is sensitive to noise. Integrating hierarchical clustering with partitioning methods such as K-means, or with improved hierarchical algorithms such as BIRCH and CURE, represents an important direction.

Keywords: hierarchical clustering, agglomerative hierarchical clustering, dendrogram, cluster analysis

1. Introduction

Whether it's opening a map app on your phone to find a nearby bus stop, biologists trying to identify "families" of genes with similar functions, or e-commerce platforms aiming to group users with similar tastes—all these scenarios point to the same question: How can we automatically group similar items together? This is the task of cluster analysis.

As the scale and complexity of data have grown, clustering methods have continued to diversify. Partitioning methods are simple and efficient but require the number of clusters to be specified in advance and only produce groupings at a single level of granularity, density-based, grid-based, and graph-based methods, on the other hand, are designed for non-spherical clusters, large-scale data, and complex structures, respectively. Hierarchical clustering takes a different approach: rather than rushing to produce flat grouping results, it constructs a clustering tree that simultaneously reveals

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2027.MELB38502

the nested structure of the data at different levels of granularity. This method is widely used in fields such as bioinformatics, image segmentation, text clustering, and business analytics.

However, existing research on hierarchical clustering still has limitations: its merging and splitting decisions are irreversible, the computational cost is high, and the selection of different linking methods often relies on experience. This paper focuses on hierarchical clustering algorithms and employs a literature review approach to examine their two construction paths—agglomeration and partitioning—the impact of linking methods on clustering results, typical application boundaries, and existing improvement strategies, with the aim of providing guidance for method selection and hybrid clustering design in exploratory analysis.

2. The basic concepts of hierarchical clustering

2.1. Basic concepts of hierarchical clustering

The goal of cluster analysis is to partition a set of objects into several groups such that objects within the same group are similar to one another, while objects in different groups differ significantly. Hierarchical clustering is a type of clustering method characterized by the fact that it does not produce a single grouping result but instead constructs a multi-level nested structure: based on a specific distance metric, it progressively merges (or splits) the most similar objects or classes until all objects are grouped into a single class or remain independent. Unlike partitioning methods such as K-means, hierarchical clustering does not require the number of clusters to be specified in advance. It outputs a clustering tree rather than a flat grouping, simultaneously presenting the grouping structure of the data at different levels of granularity. This distinction gives hierarchical clustering a unique advantage in exploratory analysis, its specific applications will be discussed in Section 5 with the aid of examples.

2.2. Two approaches to hierarchical clustering

Based on the direction of the operation, hierarchical clustering is divided into two main strategies: (1) Agglomerative hierarchical clustering (bottom-up): Starting with each sample as its own cluster, the two nearest clusters are repeatedly merged until all samples are grouped into a single cluster, as in the AGNES algorithm. (2) Split-level clustering (top-down): All samples are initially treated as a single cluster and are progressively split into smaller subclusters until each sample is independent, DIANA is a representative example. Whether using agglomerative or split-level clustering, the process is irreversible: once two clusters are merged, they will not be split again, and improper mergers in the early stages will permanently affect the dendrogram structure.

3. Algorithm flow of agglomerative hierarchical clustering

The input for agglomerative hierarchical clustering is a set of data points and a selected distance metric, the output is a dendrogram. By performing horizontal cuts on the dendrogram, a specific number of clusters can be obtained. Taking AGNES as an example, its basic algorithmic steps are clear and intuitive, as shown in Figure 1.

DOI: 10.54254/2755-2721/2027.MELB38502

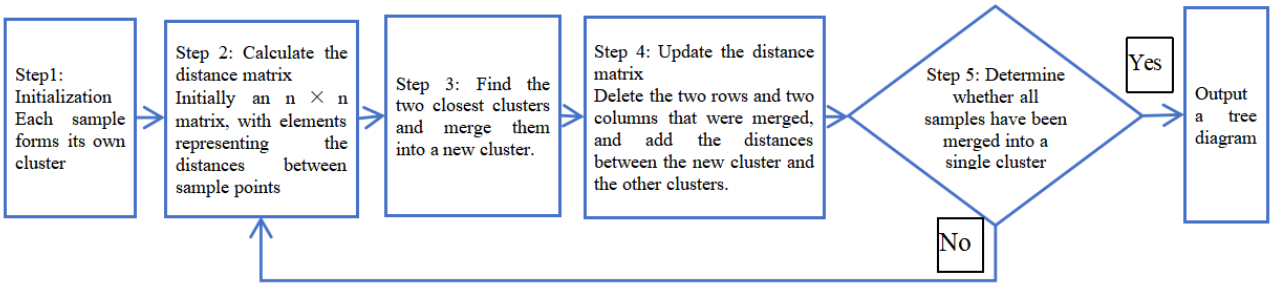

Figure 1. Flowchart of the agglomerative hierarchical clustering (AGNES) algorithm

Step 1: Initialization—each sample forms its own cluster.

Step 2: Based on the selected inter-cluster distance metric, calculate the distance matrix between the current clusters. The distance matrix is a symmetric matrix, the element in row i, column j represents the distance between cluster i and cluster j. Initially, the matrix is n×n, and each element is the distance between sample points.

Step 3: Identify the two closest clusters and merge them into a new cluster.

Step 4: Update the distance matrix—remove the two rows and two columns corresponding to the merged clusters, and add the distances between the new cluster and the remaining clusters. After merging, the two merged clusters no longer exist, so the distances between the new cluster and the remaining clusters must be recalculated based on the linking method.

Step 5: Determine whether all samples have been merged into a single cluster. If not, return to Step 3, if so, output the dendrogram.

This algorithm is irreversible. Furthermore, in terms of computational cost, a naive implementation requires scanning the distance matrix to find the minimum value at each step, resulting in a time complexity of approximately $O(n^3)$, after optimization using data structures such as priority queues, this can be reduced to $O(n^2 \log n)$. The high computational cost is the primary limitation on the application of hierarchical clustering to large-scale data.

4. Measures of inter-class distance: linkage methods

Each merging step in hierarchical clustering is based on "inter-cluster distance." However, there are two levels of distance here: the distance between points is measured by metrics such as Euclidean distance or Manhattan distance, while the distance between clusters is defined by the linking method. Applying different linking methods to the same dataset may yield vastly different clustering results. The main methods include single-linking, all-linking, average-linking, and the Ward method.

Single-link (shortest-distance method) uses the closest pair of points between two clusters as the inter-cluster distance. Its advantage is that it can identify non-spherical, elongated clusters, but it is prone to the "chain effect"—where a small number of boundary points link together groups that should be separate.

Complete linkage (longest-distance method) uses the pair of points with the greatest distance between two clusters as the inter-cluster distance, tending to produce compact clusters with clear boundaries. It does not allow "bridging" via a small number of boundary points, making it more sensitive to noise—a single outlier can significantly increase the distance between two clusters.

The average-linkage method computes the average distance between all pairs of points across two clusters, and strikes a balance between the flexibility of single-linkage clustering and the compactness of complete-linkage clustering. Compared with single-linkage and complete-linkage

DOI: 10.54254/2755-2721/2027.MELB38502

methods, it is less susceptible to the influence of individual outliers, but it also incurs a correspondingly higher computational cost.

The Ward method does not directly define the distance between clusters, instead, it adopts an analysis of variance (ANOVA) approach, selecting the two clusters whose merger results in the smallest increase in intra-cluster variance. It tends to produce clusters of balanced sizes and internal compactness, performs robustly on continuous numerical data, and is typically used in conjunction with Euclidean distance.

The choice of linking method directly affects the clustering results, there is no single "best" option, and the decision must be based on the data distribution and the task objectives.

5. Applications of hierarchical clustering

Hierarchical clustering is applied in a variety of fields. In bioinformatics, hierarchical clustering is commonly used to identify functionally related gene families when analyzing gene expression data, in the field of species classification, the structure of clustering dendrograms is highly consistent with that of phylogenetic trees. The inherent hierarchical structure of biological data is the fundamental reason for the widespread application of hierarchical clustering in this discipline. Azwar et al. [1] successfully distinguished drug-resistant from sensitive cell lines by analyzing the gene expression profiles of colorectal cancer cell lines using hierarchical clustering. Lin Qianmin [2] proposed the CCA-hierarchical clustering algorithm, which demonstrated superior internal stability and biological functionality compared to traditional hierarchical clustering when applied to rice gene expression data.

In the field of image and pattern recognition, Khanykov et al. [3] combined Ward hierarchical clustering with split-and-merge and K-means methods for object detection in color and grayscale images, overcoming computational complexity issues under limited memory constraints through the Algebraic Multi-Layer Network data structure. Guo Xinggang et al. [4] integrated hierarchical clustering with graph-based segmentation algorithms, effectively resolving the under-segmentation problem and improving processing speed through multithreaded parallelism.

Hierarchical clustering is also widely applied in document and text clustering. In citation network analysis, Havemann et al. [5] utilized the branching structure of dendrograms to identify topics at different levels of aggregation. In news topic grouping, hierarchical clustering likewise does not require a predefined number of topics, and dendrograms can visualize the nested relationship between "coarse-grained topics" and "fine-grained subtopics."

Hierarchical clustering is also applied in scenarios such as social networks and recommendation systems. Huang Haoying et al. [6] obtained node influence vectors through signal propagation and combined them with modularity to perform hierarchical clustering for community detection, when validated on benchmark networks such as the Zachary Karate Club, their accuracy outperformed the Newman algorithm. Due to computational complexity constraints, such methods are currently mostly used for small- to medium-scale network data.

In business and market analysis, customer segmentation is the most intuitive application of hierarchical clustering. By grouping customers based on purchasing behavior and demographic characteristics, dendrograms help analysts determine "how many categories are most reasonable." In exploratory scenarios, where the number of clusters is uncertain, this is precisely where hierarchical clustering excels. Sumit Kumar et al. [7] used brand search behavior as a feature to compare the performance of K-means and hierarchical clustering in e-commerce customer segmentation. K-means achieved a better coefficient of variation (0.293) than hierarchical clustering (0.254), identifying three customer segments: "high search, low purchase," "loyal customers," and

DOI: 10.54254/2755-2721/2027.MELB38502

"moderately engaged" groups. However, the dendrogram generated by hierarchical clustering still provided a basis for determining a reasonable number of clusters. Ha Thi Thu Nguyen et al. [8] used hierarchical clustering to segment customer opinions based on online reviews of the Sheraton Hanoi Hotel, compared to a manual method, the accuracy rate was approximately 83.6%, validating the consistency of opinions within the same customer group.

In summary, the application of hierarchical clustering can be categorized into three typical scenarios: scenarios requiring multi-level grouping (e.g., gene expression, citation networks), exploratory scenarios where the number of clusters is uncertain (e.g., customer segmentation, news topic grouping), and scenarios where it can be integrated with other methods to overcome computational bottlenecks (e.g., image segmentation). However, due to computational complexity, hierarchical clustering still needs to be combined with other methods when applied to large-scale data.

6. Advantages, disadvantages and areas for improvement of hierarchical clustering

The main advantage of hierarchical clustering is that it does not require the number of clusters to be specified in advance, making it suitable for exploratory analysis. It produces a dendrogram, allowing users to observe grouping structures at multiple levels of granularity simultaneously, and offers good interpretability. It also offers flexibility in the selection of distance metrics and linking methods.

The limitations of hierarchical clustering should not be overlooked: it has high computational complexity and is not suitable for extremely large datasets. Merge (or split) decisions are irreversible, and early errors cannot be corrected. It is sensitive to noise and outliers, with the degree of sensitivity depending on the linking method.

To address these limitations, researchers have proposed various improvement strategies. To address the "irreversibility" issue, hierarchical clustering can be combined with K-means [9]: First, hierarchical clustering is used to determine a reasonable value of K, followed by K-means to refine the results. To address the issue of "high computational cost," BIRCH [10] implements incremental clustering by constructing a clustering feature tree, reducing the complexity to near-linear levels. To address the issue of "incompatibility with non-spherical data," CURE [11] represents clusters using multiple representative points rather than a single center of mass, enabling the detection of non-spherical clusters.

7. Conclusion

The value of hierarchical clustering lies less in any single partition than in the multi-level view it offers on data structure. This paper reviewed the two construction paths and the influence of linkage methods on clustering outcomes.

Behind these methods, the key issue is not which linkage method is "best," but what assumptions about data structure each distance definition carries. Single-linkage, complete-linkage, average-linkage, and Ward's method differ not only in formula but in what they consider a meaningful cluster. This explains why the choice of linkage method should be treated as a modeling decision rather than a technical detail.

The efficiency, stability, and noise robustness of hierarchical clustering on large-scale data remain unresolved. Improvement strategies generally follow two paths: one is to combine hierarchical clustering with partitioning methods such as K-means, using it to suggest a reasonable K before refining the result, the other is to change how clusters are represented, as in BIRCH and CURE, to reduce cost or adapt to non-spherical structures. These two directions differ in emphasis—one works

DOI: 10.54254/2755-2721/2027.MELB38502

within the existing clustering framework, while the other redefines how clusters are represented—but both aim to make hierarchical clustering practical at larger scales. The central challenge of these hybrid approaches is how to improve efficiency while preserving the multi-granularity perspective and interpretability that make hierarchical clustering valuable in the first place. How to achieve this balance remains an open question. This review suggests that the value of hierarchical clustering lies not in a single algorithm, but in the flexibility of viewing data at multiple resolutions—and that is precisely what future hybrid methods should aim to preserve.

References

[1] Azwar, S., Ng, C. T., Zahari Sham, S. Y., et al. (2024). Possible involvement of long non-coding RNAs GNAS-AS1 and MIR205HG in the modulation of 5-fluorouracil chemosensitivity in colon cancer cells through increased extracellular release of exosomes. *Non-Coding RNA, 10*(2), 25.

[2] Lin, Q. M. (2023). A gene clustering algorithm based on the CCA-hierarchical clustering. *Journal of Harbin University of Science and Technology, 28*(5), 85–90.

[3] Khanykov, I., Nenashev, V., & Kharinov, M. (2023). Algebraic multi-layer network: Key concepts. *Journal of Imaging, 9*(7), 146.

[4] Guo, X. G., Wang, J., & Cheng, C. (2022). Image segmentation algorithm combining hierarchical clustering algorithm and graph-based segmentation algorithm. *Journal of National University of Defense Technology, 44*(3), 194–200.

[5] Havemann, F., Gläser, J., Heinz, M., et al. (2012). Identifying overlapping and hierarchical thematic structures in networks of scholarly papers: A comparison of three approaches. *PLoS ONE, 7*(3), e33255.

[6] Huang, H. Y., & Ma, Y. H. (2010). Community discovery algorithm based on signal transmission and hierarchical clustering. *Computer Engineering and Applications, 46*(9), 51–54.

[7] Kumar, S., Rani, R., Pippal, S. K., et al. (2025). Customer segmentation in e-commerce: K-means vs hierarchical clustering. *TELKOMNIKA, 23*(1), 119–128.

[8] Nguyen, H. T. T., Dao Thi Lan, H., Nguyen Thi Thu, H., et al. (2025). Customer opinion segmentation by using the hierarchical clustering method with the online reviews data. *Proc. 2025 10th Int. Conf. Information and Network Technologies*, 196–200.

[9] Zhu, E. Z., Sun, Y., Zhang, Y. X., et al. (2021). An optimal cluster number determination algorithm using a novel clustering method. *Journal of Software, 32*(10), 3085–3103.

[10] Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record, 25*(2), 103–114.

[11] Safdari-Vaigh, A., & Salehpour, P. (2021). Detecting non-spherical clusters using modified CURE algorithm. *Proc. 2021 5th Int. Conf. Computer and Knowledge Engineering (ICCKE)*, 369–373.