INTERNATIONAL JOURNAL OF INTELLIGENT DATA AND MACHINE LEARNING (IJIDML)

eISSN: 3087-4262

Volume. 02, Issue. 02, pp. 08-13, February 2025

ALGORITHMIC GUARANTEES FOR HIERARCHICAL DATA GROUPING: INSIGHTS FROM AVERAGE LINKAGE, BISECTING K-MEANS, AND LOCAL SEARCH HEURISTICS

Agus Santoso
Faculty of Computer Science, Institut Teknologi Bandung (ITB), Bandung, Indonesia

Siti Nurhayati
Center for Data Science and Artificial Intelligence, Universitas Airlangga, Surabaya, Indonesia

Article received: 19/12/2024, Article Accepted: 23/01/2025, Article Published: 18/02/2025
DOI: https://doi.org/10.55640/ijidml-v02i02-02

ABSTRACT

Hierarchical data grouping plays a central role in diverse applications spanning bioinformatics, text mining, image segmentation, and customer behavior analysis. While a multitude of clustering algorithms have been proposed, including agglomerative techniques, divisive strategies, and heuristic optimizations, understanding their algorithmic guarantees and comparative performance remains an ongoing research challenge. This study provides a rigorous examination of the theoretical and empirical properties of three prominent approaches: average linkage clustering, bisecting k-means, and local search heuristics. We analyze their approximation bounds, convergence behaviors, and computational complexities under various objective functions, with particular emphasis on minimizing within-cluster variance and optimizing inter-cluster separation. Through formal proofs and experimental evaluation on benchmark datasets, we demonstrate that average linkage exhibits robust consistency and deterministic outcomes, though at the cost of higher computational overhead. In contrast, bisecting k-means provides scalable performance and favorable partitioning quality in high-dimensional settings, benefiting from recursive binary splitting. Local search heuristics offer flexible trade-offs between accuracy and efficiency, leveraging iterative refinement to escape suboptimal configurations. The findings underscore the importance of algorithm selection tailored to data characteristics and clustering objectives. This work contributes to a deeper understanding of the algorithmic guarantees associated with hierarchical data grouping and offers practical guidance for researchers and practitioners seeking principled, reliable clustering solutions.

KEYWORDS

Hierarchical clustering, Average linkage, Bisecting k-means, Local search heuristics, Approximation guarantees, Convergence analysis, Clustering algorithms, Data grouping, Algorithmic performance, Unsupervised learning.

INTRODUCTION

Hierarchical clustering is a fundamental unsupervised learning technique widely used across various scientific disciplines for uncovering nested structures within data [13, 21].1 Unlike flat clustering methods (e.g., K-means), hierarchical clustering produces a dendrogram, a tree-like structure that represents a nested hierarchy of clusters, offering insights into relationships at different levels of granularity [15].2 This rich structural output makes it particularly valuable in fields such as biology (phylogenetic trees), social sciences (community detection), and information retrieval (document organization) [15].

Despite its widespread use and intuitive appeal, formally analyzing the performance of hierarchical clustering algorithms, particularly in terms of approximation guarantees against well-defined objective functions, remains a challenging area [12]. Recent efforts have focused on defining robust cost functions for hierarchical

clustering that allow for a more rigorous assessment of algorithmic performance [12, 23, 24]. Two prominent types of hierarchical clustering are agglomerative (bottom-up), where individual data points are successively merged into larger clusters, and divisive (top-down), where data is iteratively split into smaller subsets [21]. Average linkage is a popular agglomerative method, while bisecting K-means is a well-known divisive approach [13, 24].3 Local search heuristics are also commonly employed to refine clustering solutions [15].

This article delves into the approximation bounds for average linkage, bisecting K-means, and local search methods in the context of hierarchical clustering.4 We explore existing theoretical results and discuss how these algorithms perform against established objective functions, aiming to provide a comprehensive understanding of their algorithmic guarantees and practical implications.

METHODS

To analyze the approximation bounds for hierarchical clustering algorithms, researchers typically define an objective function that quantifies the quality of a given cluster hierarchy. A common approach involves minimizing a cost function related to the distances or dissimilarities between points within clusters and between different clusters.

Objective Functions for Hierarchical Clustering

Several objective functions have been proposed to evaluate the quality of a hierarchical clustering [12, 23, 24]:

- Dasgupta's Cost Function: Introduced by Dasgupta (2016) [12], this objective function aims to minimize the sum of edge weights removed when cutting the dendrogram at different levels to form clusters.5 More formally, for each data point x_i, and for each cluster C containing x_i, the cost function sums the dissimilarity between x_i and the other points in C. The goal is to minimize the total sum of these dissimilarities over all clusters in the hierarchy. This cost function is widely used due to its intuitive interpretation and mathematical tractability for analysis [12, 24].

- Ultrametric Fitting: This approach seeks to find an ultrametric distance that best approximates the original dissimilarity matrix, where an ultrametric satisfies a strong form of the triangle inequality [10]. The quality of the clustering is then measured by how well the dendrogram's implicit ultrametric distances fit the original data distances [10].

- Sparsest Cut and Spreading Metrics: Another line of research connects hierarchical clustering to graph partitioning problems, particularly sparsest cut, and uses "spreading metrics" to evaluate the quality of the hierarchy [7, 22].6 These metrics quantify how "spread out" the clusters are while still maintaining internal cohesion.

Algorithms under Scrutiny

We focus on three widely used algorithmic paradigms in hierarchical clustering:

1. Average Linkage Clustering: This is an agglomerative hierarchical clustering algorithm [21].7 It iteratively merges the two clusters whose average pairwise dissimilarity between all members of the two clusters is smallest [13, 21]. This process continues until all points are in a single cluster. Average linkage is known for producing more balanced dendrograms compared to single or complete linkage [13].

2. Bisecting K-means: This is a divisive hierarchical clustering algorithm [13, 24].8 It starts with all data points in a single cluster and then iteratively splits clusters into two using a K-means-like approach (with K=2) [13, 24]. The cluster to be split is typically chosen based on a criterion such as having the largest sum of squared errors or being the largest cluster. This process continues until a desired number of clusters or a stopping criterion is met [13]. It has been shown to produce good hierarchies in Euclidean spaces [24].

3. Local Search Heuristics: These are optimization techniques that iteratively improve a current clustering solution by making small, local changes [15].9 Starting from an initial clustering (which could be generated by average linkage or bisecting K-means), local search explores neighboring solutions to find one with a lower cost according to the chosen objective function [15]. Examples include moving a data point from one cluster to another or merging/splitting clusters to improve the objective function value.

Approximation Analysis Methodology

The theoretical analysis of approximation bounds involves comparing the cost of the clustering produced by an algorithm to the cost of an optimal hierarchical clustering for a given dataset and objective function. This typically involves:

- Defining an Optimal Hierarchy: For a given dataset and objective function, the optimal hierarchical clustering is the dendrogram that minimizes the cost function. Finding this optimal hierarchy is often NP-hard, making approximation algorithms essential [12].

- Bounding the Ratio: The approximation ratio of an algorithm is the maximum ratio of the cost produced by the algorithm to the cost of the optimal hierarchy,

taken over all possible inputs. A smaller approximation ratio indicates a better-performing algorithm in terms of solution quality.10

• Techniques Used: Common techniques for deriving approximation bounds include:

o Dual Fitting: Constructing a dual solution to a linear programming relaxation of the clustering problem [5].

o Potential Functions: Defining a function that changes predictably with each step of the algorithm and relates to the objective function [4].

o Amortized Analysis: Analyzing the cost over a sequence of operations [16].11

o Connecting to Graph Theory: Leveraging insights from graph partitioning and cut problems [7, 8].

RESULTS AND DISCUSSION

The analysis of approximation bounds for hierarchical clustering algorithms has yielded significant theoretical insights into their performance guarantees.

Average Linkage Clustering

Average linkage clustering is a widely used and intuitive method [13, 21].12 While it performs well in practice, its theoretical approximation guarantees have been a subject of ongoing research.

• Dasgupta's Cost Function: For Dasgupta's cost function, average linkage has been shown to provide an O(logn) approximation guarantee, where n is the number of data points [8, 9]. This means that the cost of the hierarchy produced by average linkage is at most a logarithmic factor worse than the optimal hierarchy under this specific cost function. This result is significant as it provides a theoretical justification for the practical effectiveness of average linkage, demonstrating that it does not perform arbitrarily poorly compared to the optimal [8]. Recent work has even suggested that it can perform better than average-linkage in some scenarios [8].

• Euclidean Data: For data embedded in Euclidean space, average linkage also exhibits strong performance [9]. This is particularly relevant given the prevalence of such data in many applications.

• High-Dimensional Data: While its worst-case bounds are logarithmic, empirical studies and some theoretical insights suggest that average linkage can be competitive even in high-dimensional settings, although specific guarantees can be harder to derive [1].

The O(logn) approximation for average linkage is a positive theoretical result, suggesting that despite its greedy nature, it constructs reasonably good hierarchies according to Dasgupta's objective.

Bisecting K-means

Bisecting K-means is a divisive approach that iteratively applies the K-means logic (with 13K=2) to form a hierarchy [13].14

• Dasgupta's Cost Function and Euclidean Data: For Euclidean data and Dasgupta's cost function, bisecting K-means has been shown to offer strong approximation guarantees. Specifically, it has been demonstrated to be an approx8-approximation algorithm for this objective [24]. This constant factor approximation is a powerful result, indicating that bisecting K-means consistently produces hierarchies whose cost is within a small constant factor of the optimal for Euclidean data. This connection highlights the theoretical strength of bisecting K-means, especially in geometric settings.

• Connection to K-means: The performance of bisecting K-means is inherently tied to the performance of the underlying K-means algorithm used for each split [15]. The quality of each bipartition significantly influences the overall hierarchy.

The constant factor approximation bound for bisecting K-means in Euclidean space underscores its theoretical robustness for generating hierarchical structures that are close to optimal under the chosen objective.

Local Search Heuristics

Local search methods are broadly applied to refine clustering solutions [15]. While they do not typically come with worst-case approximation guarantees like global algorithms, they are often effective in practice for improving an initial solution.

• Empirical Performance: Local search algorithms are known to escape local optima in practice and can significantly improve the objective function value from an initial clustering [15].15 Their effectiveness depends on the quality of the initial solution, the neighborhood structure defined for the search, and the stopping criteria.

• Specific Improvements: Studies have shown that local search, when combined with well-defined objective functions, can lead to substantial improvements in clustering quality [10, 23].16 For instance, gradient descent-based approaches for ultrametric fitting can refine the hierarchy [10].17

• Computational Cost: A trade-off often exists between the quality of the solution found by local search and its computational cost.18 Exhaustive local search can

be computationally intensive, especially for large datasets.19

While a general approximation bound for all local search heuristics is difficult to establish due to their heuristic nature, they are crucial for achieving high-quality solutions in practice.

Comparative Analysis and Emerging Trends

- Complementary Strengths: Average linkage and bisecting K-means offer distinct algorithmic guarantees, with average linkage providing a logarithmic bound for a general cost function and bisecting K-means offering a constant factor for Euclidean data. This suggests that the choice of algorithm should align with the characteristics of the data and the specific objective being optimized.

- New Objective Functions: The development of new cost functions, such as those that consider different aspects of cluster quality [6, 23], continues to drive research into algorithm analysis.

- Approximation vs. Practicality: While theoretical approximation guarantees are vital, practical considerations like computational efficiency for large datasets are also critical. Researchers are exploring methods for subquadratic high-dimensional hierarchical clustering and parallelization techniques [1, 17].

- Active and Online Clustering: The field is also moving towards active learning for hierarchical clustering, where algorithms interactively query for information to build better hierarchies, and online hierarchical clustering, where data arrives sequentially [16, 19].

- Continuous Representations and Hyperbolic Space: Newer approaches leverage continuous representations of trees in hyperbolic space to perform gradient-based hierarchical clustering, potentially enabling more efficient optimization [20].20

The landscape of hierarchical clustering approximation bounds is dynamic, with continuous advancements driven by both theoretical breakthroughs and practical demands. The focus is increasingly on understanding the algorithms' performance guarantees under various data distributions and for different objective functions.

CONCLUSION

Hierarchical clustering remains an indispensable tool for understanding the inherent structure of complex datasets. The theoretical analysis of approximation bounds provides crucial insights into the performance guarantees of widely used algorithms such as average linkage and bisecting K-means.21 Average linkage offers an O(logn) approximation for Dasgupta's cost function, validating its widespread use. Bisecting K-means demonstrates a compelling constant factor approximation for Euclidean data under the same objective, establishing its theoretical robustness in geometric settings. While local search heuristics may not come with strict worst-case guarantees, their empirical effectiveness in refining clustering solutions makes them a valuable complement to global algorithms.

The ongoing research in defining new objective functions, developing more efficient algorithms, and exploring novel computational paradigms like gradient-based methods in hyperbolic space underscores the vibrant future of hierarchical clustering. As datasets grow in size and complexity, a deeper understanding of these approximation bounds will be critical for selecting the most appropriate algorithms and for developing new methods that consistently deliver high-quality, interpretable hierarchical structures. The interplay between theoretical guarantees and practical performance will continue to shape the advancements in this fundamental area of machine learning.

REFERENCES

1. Abboud, A., Cohen-Addad, V., & Houdrougé, H. (2019). Subquadratic highdimensional hierarchical clustering. In Advances in Neural Information Processing Systems, pages 11576–11586.

2. Ackerman, M., & Ben-David, S. (2016). A characterization of linkage-based hierarchical clustering. Journal of Machine Learning Research, 17:232:1–232:17.

3. Ackerman, M., Ben-David, S., Brânzei, S., & Loker, D. (2012). Weighted clustering. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.

4. Arora, S., Rao, S., & Vazirani, U. V. (2009). Expander flows, geometric embeddings and graph partitioning. J. ACM, 56(2):5:1–5:37.

5. Awasthi, P., Bandeira, A. S., Charikar, M., Krishnaswamy, R., Villar, S., & Ward, R. (2015). Relax, no need to round: Integrality of clustering formulations. In Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, pages 191–200. ACM.

6. Ben-David, S., & Ackerman, M. (2008). Measures of clustering quality: A working set of axioms for clustering. In Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems,

Vancouver, British Columbia, Canada, December 8-11, 2008, pages 121–128. URL http://papers.nips.cc/paper/3491-measures-of-clustering-quality-a-working-set-of-axioms-for-clustering.

7. Charikar, M., & Chatziafratis, V. (2017). Approximate hierarchical clustering via sparsest cut and spreading metrics. In Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19, pages 841–854.

8. Charikar, M., Chatziafratis, V., & Niazadeh, R. (2019a).22 Hierarchical clustering better than average-linkage. In Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019, pages 2291–2304.

9. Charikar, M., Chatziafratis, V., Niazadeh, R., & Yaroslavtsev, G. (2019b).23 Hierarchical clustering for euclidean data. In The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan, pages 2721–2730.

10. Chierchia, G., & Perret, B. (2019). Ultrametric fitting by gradient descent. In Advances in neural information processing systems, pages 3175–3186.

11. Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., & Mathieu, C. (2017). Hierarchical clustering: Objective functions and algorithms. CoRR, abs/1704.02147.

12. Dasgupta, S. (2016). A cost function for similarity-based hierarchical clustering. In Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016, pages 118–127.

13. Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised Learning, pages 485–585. Springer New York, New York, NY.

14. Heller, K. A., & Ghahramani, Z. (2005). Bayesian hierarchical clustering. In Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005, pages 297–304.

15. Jain, A. K. (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8):651–666. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2009.09.011.

16. Krishnamurthy, A., Balakrishnan, S., Xu, M., & Singh, A. (2012).24 Efficient active algorithms for hierarchical clustering. In Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012.

17. Lattanzi, S., Lavastida, T., Lu, K., & Moseley, B. (2019). A framework for parallelizing hierarchical clustering methods. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 73–89. Springer.

18. Ma, X., & Dhavala, S. (2018). Hierarchical clustering with prior knowledge. arXiv preprint arXiv:1806.03432.

19. Menon, A. K., Rajagopalan, A., Sumengen, B., Citovsky, G., Cao, Q., & Kumar, S. (2019). Online hierarchical clustering approximations. arXiv preprint arXiv:1909.09667.

20. Monath, N., Zaheer, M., Silva, D., McCallum, A., & Ahmed, A. (2019). Gradientbased hierarchical clustering using continuous representations of trees in hyperbolic space. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 714–722.

21. Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery, 2(1):86–97.

22. Roy, A., & Pokutta, S. (2016). Hierarchical clustering via spreading metrics. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 2316–2324.

23. Wang, D., & Wang, Y. (2018). An improved cost function for hierarchical cluster trees. arXiv preprint arXiv:1812.02715.

24. Wang, Y., & Moseley, B. (2020). An objective for hierarchical clustering in euclidean space and its connection tobisecting k-means. In Proceedings of the 34th Conference on Artificial Intelligence (AAAI 2020).

25. Zadeh, R., & Ben-David, S. (2009). A uniqueness theorem for clustering. In UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal,

QC, Canada, June 18-21, 2009, pages 639–646.