ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 1, 2025

APPROXIMATION BOUNDS FOR CLUSTERING: AVERAGE LINKAGE, BISECTING K-MEANS

*K Yadagiri [1], A Bhasha [2], V Chandraprakash [3], S Veeresh Kumar [4]

[1,2,3,4] Assistant Professor, St. Martin's Engineering College, Secunderabad, Telangana – 500100

*Corresponding Author

Email: kyadagiriit@smec.ac.in

Abstract

Hierarchical clustering is a widely used method to analyze data. See Murtagh and Contreras (2012); Krishnamurthy et al. (2012); Heller and Ghahramani (2005) for an overview and pointers to relevant work. In a typical hierarchical clustering problem, one is given a set of *n* data points and a notion of similarity between the points. The output is a hierarchy of clusters on the input. Specifically, a dendrogram (tree) is constructed where the leaves correspond to the *n* input data points and the root corresponds to a cluster containing all data points. Each internal node of the tree corresponds to a cluster of the data points in its subtree. The clusters (internal nodes) become more refined as we move down the tree. The goal is to construct the tree so that these deeper clusters contain points that are relatively more similar.

1. Introduction

Hierarchical clustering is a widely used method to analyze data. See Murtagh and Contreras (2012); Krishnamurthy et al. (2012); Heller and Ghahramani (2005) for an overview and pointers to relevant work. In a typical hierarchical clustering problem, one is given a set of *n* data points and a notion of similarity between the points. The output is a hierarchy of clusters on the input. Specifically, a dendrogram (tree) is constructed where the leaves correspond to the *n* input data points and the root corresponds to a cluster containing all data points. Each internal node of the tree corresponds to a cluster of the data points in its subtree. The clusters (internal nodes) become more refined as we move down the tree. The goal is to construct the tree so that these deeper clusters contain points that are relatively more similar.

There are many reasons for the popularity of hierarchical clustering, including that the number of clusters is not predetermined and that the clusters produced induce taxonomies that give meaningful ways to interpret data.

Methods used to perform hierarchical clustering are divided into two classes: agglomerative and divisive. **Agglomerative** algorithms take a bottom-up approach and are more commonly used than divisive approaches (Hastie et al., 2009). In an agglomerative method, each of the *n* input data points starts as its own cluster. Then iteratively, pairs of similar clusters are merged according to some appropriate notion of similarity. Perhaps the most popular definition of similarity is **average linkage** where the similarity between two clusters is defined as the average similarity between all pairs of data points in the two clusters. In average linkage agglomerative clustering the two clusters with the highest average similarity are merged at each step. Other variants are also popular. Related examples include: **single linkage**, where the similarity between two clusters is the maximum similarity between any pair of single data points in different clusters, and **complete linkage**, where the distance is the minimum similarity between any pair of single data points in different clusters.

Divisive algorithms take a top-down approach where initially all data points are placed into a single cluster. They recursively perform splits, dividing a cluster into smaller clusters that will be further subdivided. The process continues until each cluster consists of a single data point. In each

step of the algorithm, the data points are partitioned such that points within each cluster are more similar than points across clusters. There are several approaches to perform divisive clustering. One example is bisecting k-means where k-means is used at each step with $k = 2$. For details on bisecting k-means, see Jain (2010).

Motivation: Hierarchical clustering has been used and studied for decades. There has beensome work on theoretically quantifying the quality of the solutions produced by algorithms, such as Ackerman et al. (2012); Ackerman and Ben-David (2016); Zadeh and Ben-David (2009); Ben-David and Ackerman (2008); Dasgupta (2016). Much of this work focuses on deriving the structure of solutions created by algorithms or analytically describing desirable properties of a clustering algorithm. Though the area has been well-studied, there is no widely accepted formal problem framework. Hierarchical clustering describes a class of algorithmic methods rather than a problem with an objective function. Studying a formal objective for the problem could lead to the ability to objectively compare different methods; there is a desire for the community to investigate potential objectives, which would further support the use of current methods and guide the development of improvements.

This paper is concerned with investigating objectives for hierarchical clustering. It givesa natural objective and leverages its structural connection to average linkage agglomerative clustering to prove this algorithm obtains a constant approximation to the best possible clustering. In contrast to this positive result, single linkage, complete linkage, and bisecting k-means are shown to have superconstant (i.e. scaling with the number of data points) approximation ratios. This paper also provides some divisive algorithms that have comparabletheoretical guarantees to average linkage.

Problem Formulation: Towards this paper's goal, we begin by trying to establish a formal problem framework for hierarchical clustering. Recently, Dasgupta (2016) introduced a new problem framework for hierarchical clustering. This work justified its proposed objective by establishing that for several sample problem instances, the resulting solution corresponds to what one might expect out of a desirable solution. This work spurred considerable interest and there have been several follow up papers (Charikar and Chatziafratis, 2017; Dasgupta,2016; Roy and Pokutta, 2016).

Related Work (Other Cost Functions): Recently a contemporaneous paper (Cohen- Addad et al., 2017) done independently has been published. This paper considers a class ofobjectives motivated by the work of Dasgupta (2016). For their objective, they also derive positive results for average linkage clustering and additionally give axiomatic properties that are desirable in an objective for hierarchical clustering. Ma and Dhavala (2018) consider combining Dasgupta's objective function with prior knowledge about the data set. Wang and Wang (2018) suggests an alternate objective with the goal of comparing the clusterability across different input graphs, rather than just different clusterings for a single input graph. Both Chierchia and Perret (2019) and Monath et al. (2019) propose continuous objective functions with the goal of applying gradient descent. Lattanzi et al. (2019) and Abboud et al. (2019) both strive to make more efficient (parallelized and raw runtime, respectively) versions of classical clustering algorithms, and judge the resulting quality via comparing iteration-by-iteration against what the classical algorithm would have done. Wang and Moseley (2020) propose an objective function for which bisecting k-means achieves a constant approximation.

2. Preliminaries

In this section, we give preliminaries including a formal definition of the problem consideredand basic building blocks for later algorithm analysis.

In the **Reward Hierarchical Clustering Problem** there are n input data points given as a set V. There is a weight $w_{i,j} \geq 0$ between each pairof points i and j denoting their similarity, represented as a complete graph G. The output of the problem is a rooted tree T where the leaves correspond to the data points and the internal nodes of the tree correspondto clusters of the points in the subtree. We will use the indices $1, 2, \ldots n$ to denote the leaves of the tree. For two leaves i and j, let $T[i \vee j]$ denote the subtree rooted at the least common ancestor of i and j and let the set leaves-outside($T[i$

$\vee j]$) denote the number ofleaves in T that are not in $T[i \vee j]$. The objective is to construct T to maximize the rewad optimal solution to maximizing reward$_G(T)$. Thus, there is an optimal solution for the reward$_G(T)$ objective that is binary.

2.1 Analyzing Agglomerative Algorithms

In this section, we discuss a method for bounding the performance of an agglomerative algorithm. When an agglomerative algorithm merges two clusters A, B, this determinesthe least common ancestor for any pair of nodes i, j where $i \in A$ and $j \in B$. Given this, we define the reward gain *due to* merging A and B as, merge-rew$_G(A, B) := (n - |A| - |B|) \sum_{a \in A, b \in B} w_{ab}$ Notice that the final reward reward$_G(T)$ is exactly the sum over iterations of the reward gains,since each edge is counted exactly once: when its endpoints are merged into a singlecluster. Hence, reward$_G(T) =$ merges A,

3. Conclusion

One motive for developing an analytic framework is that it may help clarify and explain our observations from practice. In this case, we have shown that average linkage is a $\frac{1}{3}$-approximation to a particular objective function (Theorem 1), and the analysis that does so helps to explain what average linkage is optimizing. We have also shown that average-linkage can be no better than a $\frac{1}{2}$-approximation One open problemis to devise new algorithms and determine the best approximation ratio possible for the problem. The current state of the art is a (0.336)-approximation based on semi-definite programming Can this be improved further? Another open problem is to find a characterization of graphs that excludes some of the worst-case ones used to provenegative results. Is there a formal way to restrict inputs that allows for better objective guarantees?

We mention that similar results to ours for average-linkage have been shown by Cohen- Addad et al. In this work, it is shown that average-linkage is a $\frac{1}{2}$-approximation for a related objective function when there are dissimilarity scores between the points.

Another open direction is the possibility of other objective functions. What are single linkage, complete linkage, and bisecting k-means optimizing for? One quirk shared by both Dasgupta's cost objective and our reward objective is that the optimal tree is always binary.This is not appropriate for all applications; for example, in the classical application of biological taxonomy, groups typically contain much more than two subgroups. Can we devise an objective function which incentivizes non-binary trees?

4. Acknowledgments

Benjamin Moseley was supported in part by a Google Research Award, a Yahoo Research Award and NSF Grants CCF-1617724, CCF-1733873, CCF-1725661, CCF-1824303, CCF-1845146 and CMMI-1938909. This work was partially done while the author was working at Washington University in St. Louis.

Joshua R. Wang was supported in part by NSF Grant CCF-1524062. This work waspartially done while the author was at Stanford University.

1. A follow-up work to ours shows that average-linkage cannot achieve a $\frac{1}{3} + g$ approximation for any $g > 0$ (Charikar et al., 2019a).

References

Amir Abboud, Vincent Cohen-Addad, and Hussein Houdrougé. Subquadratic high- dimensional hierarchical clustering. In *Advances in Neural Information Processing Systems*, pages 11576–11586, 2019.

Margareta Ackerman and Shai Ben-David. A characterization of linkage-based hierarchical clustering. *Journal of Machine Learning Research*, 17:232:1–232:17, 2016.

Margareta Ackerman, Shai Ben-David, Simina Brânzei, and David Loker. Weighted clustering. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.*, 2012.

Sanjeev Arora, Satish Rao, and Umesh V. Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56(2):5:1–5:37, 2009.

Pranjal Awasthi, Afonso S Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. 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, 2015.

Shai Ben-David and Margareta Ackerman. Measures of clustering quality: A work- ing 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, De- cember 8-11, 2008*, pages 121–128, 2008. URL http://papers.nips.cc/paper/ 3491-measures-of-clustering-quality-a-working-set-of-axioms-for-clustering.

Moses Charikar and Vaggos Chatziafratis. 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, 2017.

Moses Charikar, Vaggos Chatziafratis, and Rad Niazadeh. 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, 2019a.

Moses Charikar, Vaggos Chatziafratis, Rad Niazadeh, and Grigory Yaroslavtsev. Hierarchicalclustering for euclidean data. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 2721–2730,2019b.