cutTreeBalanced - Balanced cuts for dendrograms

Lewis J. Martin[1] (ljmartin@hey.com)
Vicente Reyes-Puerta (vr.github@outlook.com)

[1] University of Sydney, NSW, Australia

1 Abstract

Hierarchical clustering techniques are common in a wide range of fields. The output of these techniques are dendrograms, which are trees that define the linkages between input data leading to clusters. Generating clusters requires some method of cutting the dendrogram to separate groups of points from each other. Many existing cutting techniques, such as a straight cut along a constant height, lead to clusters with highly imbalanced sizes. In certain cases, such as with high-dimensional data, these clusters do not accurately represent the underlying structure of the data. This paper presents and demonstrates - using a real world example in the field of biochemistry - a novel approach named cutTreeBalanced for Python, a fast and simple method of cutting dendrograms that results in balanced cluster sizes.

2 Introduction

Hierarchical (alternatively agglomerative) clustering techniques are common in a wide range of academic fields (such as astronomy, biology, chemistry, economics, ecology and bioinformatics) and business domains (such as customer segmentation, recommender systems and online marketing), where they are used to group data based on their similarity in some (multi-dimensional) metric space. Ideally, these groupings organize the data by abstracting the underlying structure into the cluster labels, leading to new insights.[1] In a typical setting, a pairwise distance matrix is used to iteratively join individual objects, or groups of objects, together into incrementally larger clusters.[2] This results in a dendrogam which fully describes the linkages between the initial state (where each object is in a single-member cluster) to the end state (where all objects are grouped into one cluster). An example of a dendrogram is shown in Figure 1.

Joining nodes in a dendrogram requires a rubric to define the shortest similarity. For example, single-linkage clustering joins clusters based on the shortest distance between any member. Alternatively, complete linkage defines the shortest

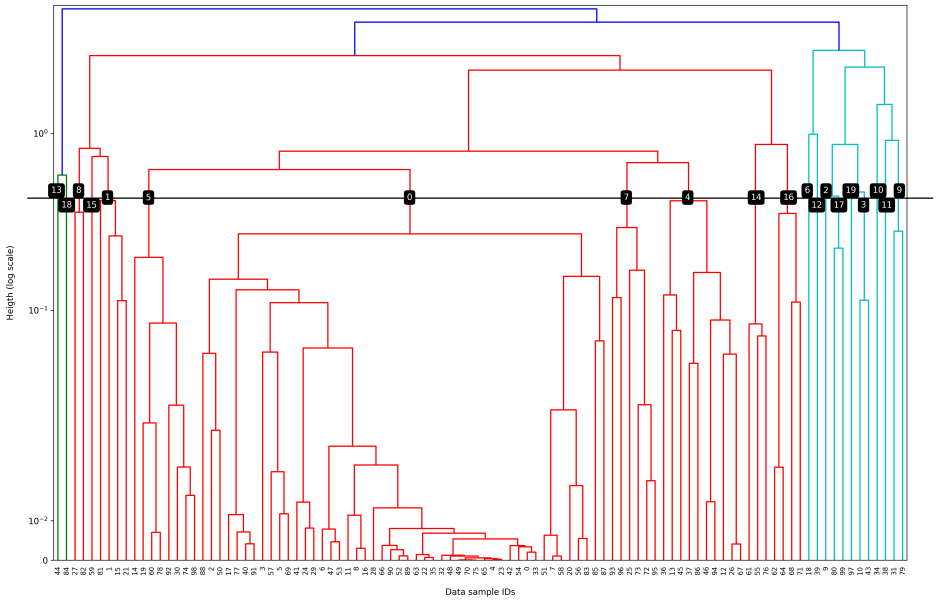

Figure 1: A dendrogram generated by single-linkage clustering, showing a straight cut at height XYZ. Straight cuts at any height generate clusters with imbalanced sizes (`can we add a histogram of cluster sizes at multiple heights? LJM`)

similarity as the smallest maximum distance between two clusters. A third method, Ward linkage, chooses clusters based on the minimum increase in within-cluster variance.[3] All of these methods result in dendrograms of the same form. To generate cluster labels, the dendrogram is 'cut' at a certain point, such that all cluster joins after that point are ignored (see 1). A common type of cut, which is also conceptually simple, is demonstrated in 1, whereby all linkages beyond a user-defined similarity are ignored. This may also be known as a constant height cut.

Constant-height cuts commonly lead to clusters with imbalanced sizes.

- hdbscan:,[4] dynamicTreeCut[5]
- High-dimensional data have reduced contrast,[6] and often high 'hubness',[7] as a result of curse of dimensionality. In these cases, similarity by usual distance metrics is not well defined. Single-linkage, Ward linkage, or other linkages may result in dendrograms that are impossible to cut while preserving underlying structure. In these cases, variable cutting distance is required.
- Example of hubness causing bad hierarchical clustering results:[8]

(a reference to the abstract: Abstract).

3 Algorithm

[algorithm explanation using toy example (random data)]

4 Results

[results using your molecular data]

5 Discussion

[comparison to other approaches and outlook]

References

1. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).
2. Bramer, M. *Principles of data mining.* **180**, (Springer, 2007).
3. Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 236–244 (1963).
4. McInnes, L., Healy, J. & Astels, S. Hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* **2**, 205 (2017).

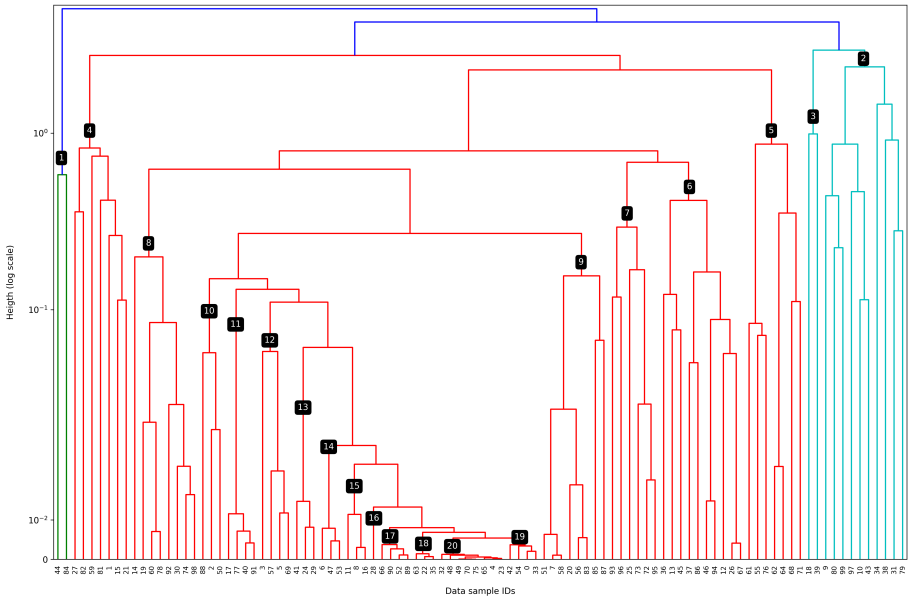

Figure 2: The same dendrogram as in 1, but demonstrating a balanced cut of size XYZ

5. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for r. *Bioinformatics* **24**, 719–720 (2008).

6. Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. When is 'nearest neighbor' meaningful? in *International conference on database theory* 217–235 (Springer, 1999).

7. Radovanovic, M., Nanopoulos, A. & Ivanovic, M. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* **11**, 2487–2531 (2010).

8. MacCuish, J., Nicolaou, C. & MacCuish, N. E. Ties in proximity and clustering compounds. *Journal of Chemical Information and Computer Sciences* **41**, 134–146 (2001).