

cutTreeBalanced: Balanced cuts for dendrograms

Cutting dendrograms to achieve hierarchical clustering with balanced cluster sizes.

1 Abstract

Hierarchical (a.k.a. agglomerative) linkage techniques are common in a wide range of fields such as economics, bioinformatics, (**some fields**). The output of these techniques are dendrograms, which are directed acyclic graphs (**note: not just DAGs - they have tree structure**) that define the linkage between input data. Generating clusters from dendrograms requires some method of cutting to separate groups of points from each other. Many existing cutting techniques, such as cutting at a constant height, lead to clusters with highly imbalanced sizes. This paper presents and demonstrates balancedCut for Python - a method of cutting dendrograms with balanced cluster sizes.

2 Introduction

- What is hierarchical clustering, dendrogram, tree cut.
- Straight cut (i.e. constant height cut).
- hdbscan: [1], dynamicTreeCut [2]
- High-dimensional data have reduced contrast [3], and often high ‘hubness’ [4], 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.

(a reference to the abstract: Abstract).

3 Algorithm Results

References

- [1] L. McInnes, J. Healy, and S. Astels, “Hdbscan: Hierarchical density based clustering,” *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.
- [2] P. Langfelder, B. Zhang, and S. Horvath, “Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for r,” *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.
- [3] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is ‘nearest neighbor’ meaningful?” in *International conference on database theory*, 1999, pp. 217–235.
- [4] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, “Hubs in space: Popular nearest neighbors in high-dimensional data,” *Journal of Machine Learning Research*, vol. 11, no. sept, pp. 2487–2531, 2010.