DENDROGRAM SLICING THROUGH A PERMUTATION TEST APPROACH RECONSIDERED

L. Palazzo, A. Iodice D'Enza, F. Palumbo, and D. Vistocco
Department of Political Sciences, University of Naples Federico II
(e-mail: [lucio.palazzo, iodicede, domenico.vistocco, fpalumbo]@unina.it)

ABSTRACT: DESPOTA is a clustering method that cuts the tree branches at various heights to find the best division among those that can be achieved from the hierarchical clustering tree through the use of a permutation test at each node. In order to reduce the computational cost and increase the applicability of DESPOTA to huge data sets, the present study suggests two improvements to the DESPOTA original implementation that combine aggregation with either splitting or partitioning approaches. A dataset of the Italian universities' five-year periodical accreditation by the Italian national agency (ANVUR) is used to test the suggested approach.

KEYWORDS: hierarchical clustering, permutation test, top-down splitting

1 Introduction

Hierarchical algorithms represent excellent solutions for data clustering when one aims to get nested partitions in the data that can be easily visualized through tree-like representations, also referred to as dendrograms (from the Greek word δὲνδρον= tree). Cutting the tree at a given level defines data partitioning into disjoint clusters. Nonetheless, the optimal cutting level (corresponding to the optimal number of clusters) remains a ticklish problem, and the choice is generally left to the user's heuristic criteria. DESPOTA (Bruzzese & Vistocco, 2015, DEndrogram Slicing through a PermutatiOn Test Approach) seeks the best partition among the possible ones achievable from a hierarchical clustering tree, cutting the tree branches at different heterogeneity levels. DESPOTA performs a permutation test at each node under the null hypothesis that the two descending branches sustain only one cluster. It ensures that the optimal number of clusters is based on the decision made using independent permutation tests, considering the minimum cost required for joining two branches and the cost incurred in the merging process. DESPOTA does not require any distributional assumption and works in a purely data-driven approach. The use of permutations to test for clusteredness in abundance/species data has been proposed by Greenacre (2011). DESPOTA needs a considerable

computational burden, even for moderately large data sets. At each node of the dendrogram, an agglomerative procedure is applied on each branch and for each permutation.

This paper proposes two modifications of the DESPOTA original implementation, aiming to limit the computational effort and favor the applicability of DESPOTA to large data sets. In particular, while the DESPOTA original procedure is purely agglomerative, we propose two variations combining the agglomerative with divisive and partitioning approaches. The divisive approach-based proposal is based only on distances and is suitable for categorical and mixed data. The partitioning-based approach provides further computational efficiency, yet it requires continuous data.

The paper presents some main results concerning a dataset containing some variables that refer to the efficiency and effectiveness of education at Italian universities. These variables are a subset of those that are considered for the five-year periodical accreditation by the Italian national agency (ANVUR).

The paper is organized as follows: Section 2 recalls the DESPOTA test statistic while Section 3 describes the proposed enhancements; Section 4 provides an example and concludes the paper.

2 DESPOTA: general idea and test statistics

Any indexed hierarchy defines a sequence of nested partitions, and at each partitioning, it corresponds to a level of heterogeneity $h(\cdot)$ dictating whether observations/groups are clustered. The choice of $h(\cdot)$ and the corresponding cluster solution is up to the user's expertise and knowledge of the domain.

In order to provide a data-driven choice, Bruzzese & Vistocco (2015) provided a test statistic that evaluates whether two subgroups should be kept separated or merged together. Under the null hypothesis, it is assumed there is no gain in splitting the subgroups at hand. Let us consider a generic dendrogram and let $h(L_k)$ and $h(R_k)$ be, respectively, the left and right branch heterogeneity levels at the node k; then, the test statistic is obtained through the ratio of the minimum cost to the actual cost. Hence, for a generic node k the quantity $h(L_k \cup R_k)$ indicates the heterogeneity level merging the nodes L_k and R_k, and the test statistics is defined as:

$$rc_k = \frac{|h(L_k) - h(R_k)|}{h(L_k \cup R_k) - \min\{h(L_k), h(R_k)\}}, \tag{1}$$

is the ratio between the minimum and actual merging costs, which ranges in $[0,1]$. If rc_k is close to 1 means that L_k and R_k should be kept together.

The null hypothesis distribution is obtained via permutation: at each node k of the original hierarchy, M (usually M=999) permutations of the L_k-vs-R_k membership are considered and the corresponding rc_k values computed.

3 Using permutations to compute the null hypothesis distribution

The computation of quantities in 1 of the shuffled sets under the null hypothesis is a critical point in DESPOTA. In fact, an agglomerative procedure is applied on L_k and R_k to obtain $h(L_k)$ and $h(R_k)$. Finally, the M obtained values for rc_k (see Formula 1) will define the null distribution of the test statistics. For the general node k, the computation of rc_k only involves the second- and third-last aggregation levels. Since the agglomerative approach is bottom-up, the whole hierarchy is needed to compute the second- and third-last aggregation levels. When the complete linkage is considered, given a set $\mathcal{A}$ of observations, the following relation holds: $h(\mathcal{A}) = max(d(i,i')),\ i,i' \in \mathcal{A}$. In this case, to compute the second- and third-last aggregation levels of the hierarchy, a top-down approach can be used, doing just the first split.

A classic implementation of divisive clustering (see, e.g., DIANA, Kaufman & Rousseeuw, 2009) has a complexity of $O(n^4)$ as opposed to the $O(n^3)$ of agglomerative procedures. Several proposals in the literature enhance the computational performance of divisive approaches, making them substantially more efficient than agglomerative procedures (seeRoux (2018) for a comparative review). To compute the rc_k null distribution, a single step of a divisive approach is used at each permuted node. A further enhancement to split up the permuted nodes is using a partitioning procedure like k-means with careful seeding (Arthur & Vassilvitskii, 2006) to avoid random starts and ensure quality bi-partitions.

4 Example and Conclusions

The considered data for the application consist of three standardized indicators: *iC13* (credits earned at the first year), *iC17* (students graduating up to one year late), and *iC28* (first-year students/faculty members ratio) measured over 68 Italian universities.

Both the agglomerative and the DESPOTA procedures are applied by using the Euclidean metric and the complete linkage aggregation. In Fig1 the results of the two clustering approaches are summarized.

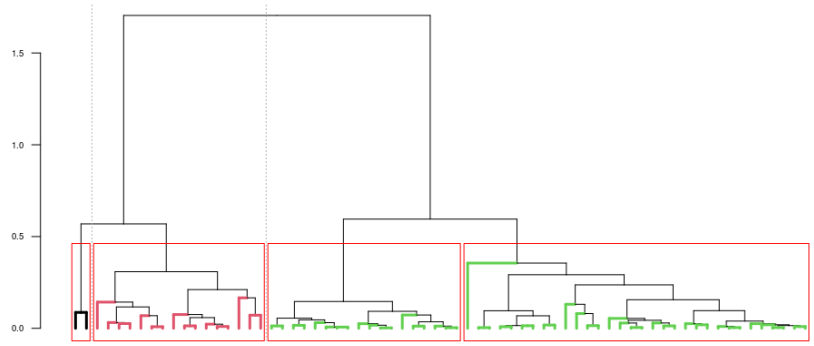

Figure 1. *Comparison between dendrogram cutting rules. The boxes depict the four clusters detected by the classical horizontal rule, while the colored leaves show the clusters selected by DESPOTA.*

DESPOTA and classical hierarchical clustering solutions disagree in the choice of the lower levels of the hierarchy: the horizontal cut splits the large group on the right-hand side of Figure 1, albeit there is no substantial difference between the two groups. DESPOTA sets Bicocca and Bocconi Universities in the same group as they present high values in all the indicators. While the best clustering solution is better interpretable, having a non-subjective procedure to pick a clustering solution is valid, even as a baseline.

References

ARTHUR, DAVID, & VASSILVITSKII, SERGEI. 2006. *k-means++: The advantages of careful seeding*. Tech. rept. Stanford.

BRUZZESE, DARIO, & VISTOCCO, DOMENICO. 2015. DESPOTA: DEndrogram slicing through a pemutation test approach. *Journal of classification*, **32**, 285–304.

GREENACRE, MICHAEL. 2011. A simple permutation test for clusteredness.

KAUFMAN, LEONARD, & ROUSSEEUW, PETER J. 2009. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

ROUX, MAURICE. 2018. A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, **35**, 345–366.