DSSV 2023 Data Science, Statistics & Visualisation
ECDA 2023 European Conference on Data Analysis

Dendrogram slicing through a permutation test approach for mixed data

Lucio Palazzo[1], Domenico Vistocco[1], and Francesco Palumbo[1]

[1]University of Naples Federico II, Department of Political Sciences, Leopoldo Rodinò, 22 - 80133 Napoli, Italy

DESPOTA [1, DEndrogram Slicing through a PermutatiOn Test Approach] is a clustering technique that cuts the tree branches at various heterogeneity levels to find the optimal division among those feasible from a hierarchical clustering tree. According to the null hypothesis that two descending branches support only one cluster, DESPOTA does a permutation test at each node. The choice of the ideal number of clusters is based on separate permutation tests, taking into account the minimal cost necessary for combining two branches and the cost associated with the merging process. DESPOTA uses a data-driven methodology and does not rely on any distributional assumptions.

Mixed data comprises both numeric and categorical features, and mixed datasets occur frequently in many domains, such as biology, education, and healthcare, among others. Mixed datasets are frequently subjected to clustering to identify structures and collect similar individuals. However, it can be difficult to directly apply mathematical operations to mixed features, making clustering a tricky task.

This paper extends the applicability of DESPOTA to mixed-type data. To this aim, the original agglomerative-based procedure is questioned and a divisive approach is proposed. The presented approach only requires the distance matrices, and thus is well suited in the case of mixed data.

[1] D. Bruzzese and D. Vistocco, "Despota: Dendrogram slicing through a pemutation test approach," *Journal of classification*, vol. 32, pp. 285–304, 2015.