Stairstep-like dendrogram cut: a permutation test approach

Dario Bruzzese[1] **, Domenico Vistocco**[2*]

1. Dip.to di Scienze Mediche Preventive, University of Naples "Federico II"
2. Dip.to di Scienze Economiche, University of Cassino

* Contact author: vistocco@unicas.it

Keywords: Hierarchical Clustering, Permutation Test, Significant Clusters

The output of hierarchical clustering methods is typically displayed as a dendrogram describing a family of partitions indexed by an ultrametric distance. Actually, after the tree structure of the dendrogram has been set up, the most tricky problem is that of cutting the tree with a suitable threshold in order to take out a sub-optimal classification. Several (more or less) objective criteria may be used to achieve this goal, e.g. the deepest step, but most often the partition relies on a subjective choice leaded by interpretation issues. Additionally, whatever the chosen criterion is, only one solution can be obtained for each desired granularity, i.e. the one where clusters are joined at consecutive heights starting from the adopted threshold.

We propose an algorithm, exploiting the methodological framework of permutation test, allowing to find out automatically a sub-optimal partition where clusters do not necessarily obey to the afore-mentioned principle.

Starting from the root node of the dendrogram, a *partial threshold* is moved down the tree until a link joining two clusters is encountered. A permutation test is thus performed in order to verify whether the two clusters must be accounted as a unique group (the null hypothesis) or not (the alternative one). If the null cannot be rejected, the corresponding branch will become a cluster of the final partition and none of its sub-branches will be longer processed. Otherwise each of them will be further visited in the course of the procedure. In fact, in both cases, the *partial threshold* will continue its path and the next branch of the dendrogram will be processed. The algorithm stops when there are no more branches that stand the test (i.e. the null cannot be rejected any more).

The permutation test on which the whole procedure is based can be summarized in this way. Under the *Null*, if all the units belonging to each of the two clusters are mixed up together and then randomly split up, with the only constraint of the group cardinality, the distance among the shuffled clusters should not be very different from the original one. Repeating the shuffling m times, a montecarlo p-value can be computed as the number of permuted distances at least as extreme as the original one.

The algorithm allows us to explore partitions which are not directly achievable using a standard cut-level approach. The obtained partition will be evaluated using several criteria proposed in literature.

References

Good, P. (1994). *Permutation tests : a practical guide to resampling methods for testing hypotheses.* Springer, New York.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. (2005). Cluster Analysis Basics and Extensions. *unpublished.*

Rand, W.M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, December 1971, 66, 336, 846–850.