DESPOTA: an algorithm to automatically detect a reliable partition on a dendrogram

Dario Bruzzese[1], Davide Passaretti[2], Domenico Vistocco[2]

{dbruzzes@unina.it, passarettidav@gmail.com, vistocco@unicas.it}

[1] Department of Public Health, University of Naples "Federico II", Italy

[2] Department of Economics and Law, University of Cassino, Italy

The output of hierarchical clustering methods is typically displayed as a dendrogram describing a family of nested partitions. However, the typical approach, horizontally cutting the dendrogram at a given distance level, explores only a restricted subset of the whole set of partitions.

We proposed an algorithm, DESPOTA - DEndrogram Slicing through a PermutatiOn Test Approach (Bruzzese and Vistocco, 2015), exploiting the methodological framework of permutation tests (Pesarin and Salmaso, 2010), that permits a partition to be automatically found where clusters do not necessarily obey the above principle.

DESPOTA offers a validated partition to the final user and it adapts to every choice of the distance metric and agglomeration criterion used to grow the tree. The algorithm retraces the tree downward, starting from the root of the dendrogram, where all objects are classified in a unique cluster, and moving down a partial threshold until a link joining two clusters is encountered. A permutation test is then performed in order to verify whether the two clusters should be considered a single group (the null hypothesis) or not (the alternative one). If the Null cannot be rejected, the corresponding branch will become an element of the final partition and none of its sub-branches will be processed any longer. Otherwise each of them will be further visited in the course of the procedure.

DESPOTA is shown in action both on real and synthetic datasets through a comparison with competitive methods (Gurrutxaga, 2010), (Milligan, 1981) (Tibshirani, 2001). The results obtained both on synthetic and real datasets show that DESPOTA performs well in situations characterized by different data and cluster structures.

Main References

BRUZZESE, D., and VISTOCCO, D. (2015), DESPOTA: DEndrogram Slicing through a PemutatiOn Test Approach, Journal of Classification 32

GURRUTXAGA, I., ALBISUA, I., ARBELAITZ, O., MARTN, J.I., MUGUERZA, J., PREZ, J.M., and PERONA, I. (2010), “SEP/COP: An Efficient Method to Find the Best Partition in Hierarchical Clustering Based on a New Cluster Validity Index”, Pattern Recognition, 43(10), 3364–3373.

MILLIGAN, G.W. (1981), “A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis”, Psychometrika, 46(2), 187–199.

PESARIN, F., and SALMASO, L. (2010), Permutation Tests for Complex Data. Theory, Applications and Software, Chichester: John Wiley and Sons.

TIBSHIRANI, R., WALTHER, G., and HASTIE, T. (2001), “Estimating the Number of Clusters in a Data Set via the Gap Statistic, Journal of Royal Statistical Society B, 83(2), 411–423