Direct Reading Algorithm for Hierarchical Clustering

Fionn Murtagh

Department of Computing and Mathematics
University of Derby
Derby, UK
fmurtagh@acm.org

Pedro Contreras

Thinking Safe Ltd.
Egham, UK
pedro.contreras@acm.org

Abstract

Reading the clusters from a data set such that the overall computational complexity is linear in both data dimensionality and in the number of data elements includes the following. In [3], direct reading of clusters is carried out, through filtering the data in wavelet transform space. Then in [4], this approach is carried out after an initial transforming of the data to a canonically order. Including high dimensional, high cardinality data, such a canonical order is provided by row and column permutations of the data matrix [2].

In [6,5] we induce a hierarchical clustering from seriation (cf. [1]) through unidimensional representation of our observations. This linear time hierachical classification is directly derived from the use of the Baire metric, which is simultaneously an ultrametric. In [7] the linear time construction of a hierarchical clustering is studied from following viewpoint: representing the hierarchy initially in an m-adic, m =

10, tree representation, followed by decreasing m to smaller valued representations that include p-adic representations, where p is prime and m is a non-prime positive integer. This has the advantage of facilitating a more direct visualization and hence reading of the hierarchy. In this work we present further case studies and examples of how this approach is very advantageous for such an ultrametric topological data mapping.

Keywords: Analytics, ultrametric topology, p-adic number representation, linear time computational complexity.

References

[1] F. Critchley and W. Heiser, Hierarchical trees can be perfectly scaled in one dimension, Journal of Classification, 5, 5–20, 1988.

[2] , Seriation and matrix reordering methods: An historical overview, Statistical Analysis and Data Mining: The ASA Data Science Journal, 3(2), 70–91, 2010.

[3] , F. Murtagh and J.L. Starck, Pattern clustering based on noise modelling in wavelet space, Pattern Recognition, 31, 847–855, 1998.

[4] , F. Murtagh, J.L. Starck and M.W. Berry, Overcoming the curse of dimensionality in clustering by means of the wavelet transform, Computer Journal, 43,107–121, 2000.

[5] F. Murtagh and P. Contreras, Random projection towards the Baire metric for high dimensional clustering, in A. Gammerman et al., Eds., Proceedings SLDS 2015, Symposium on Learning and Data Sciences, Lecture Notes in Artificial Intelligence Volume 9047, pp. 424–431, 2015.

[6] F. Murtagh and P. Contreras, Clustering through high dimensional data scaling: applications and implementation, Archives of Data Science, Proceedings of ECDA 2015, European Conference on Data Analysis 2015, submitted, 2015.

[7] , F. Murtagh, Constant time search and retrieval in massive data with linear time and space set-up, through randomly projected piling and sparse p-adic coding, preprint, 27 pp, in preparation, 2015.