Clustering of All that is Exceptional and Anomalous, Counterposed to Commonality, in Big Data Analytics

Fionn Murtagh, Professor Emeritus, University of Huddersfield, (central England, UK).

April 10, 2021

Abstract

This is often to be p-adics: for data analysis, to determine evolution and changes over time, and possibly also the context that can determine evolution.

Agglomerative hierarchical clustering obtains a binary tree. Here we use a ternary, 3-adic, tree, a divisive hierarchical clustering from the Euclidean metric endowed semantic factor space, mapped from qualitative and quantitative data by Correspondence Analysis, also termed Geometric Data Analysis. This hierarchical clustering can be of linear computational complexity.

Motivation for ternary hierarchical cluster analysis is clustering of all that is exceptional and anomalous, counterposed to commonality. Projections on the factors are used, with positive and negative projections on a factor determining exceptional and anomalous properties, and projections close to the origin being commonality.

We also want to examine this here: hierarchical clustering can be chronological (if time is a variable here) or other adjacency constrained agglomerative clustering. This is to be used for textual narrative, to determine the evolution of emotions.

From previous work [1], an important analytical issue is the resolution scale of the data and what can be the ethical issues relating to all data that is aggregated. A very important role in the analytics here now is to have the data re-encoded, such as using p-adic data encoding, rather than real-valued data encoding. For text mining, and also for medical and health analytics, the analysis determines a divisive, ternary (i.e. p-adic where p = 3) hierarchical clustering from factor space mapping. Hence the topology (i.e. ultrametric topology, here using a ternary hierarchical clustering), related to the geometry of the data (i.e. the Euclidean metric endowed factor space, semantic mapping, of the data, from Correspondence Analysis). Determined is the differentiation in Big Data analytics of what is both exceptional and quite unique relative to what is both common and shared, and predominant. The analytics seeks both the typical

and standard data characteristics, as well as orienting the analysis towards the exceptional and atypical data characteristics.

The principal applications here are text analysis and medical and health analytics. A future objective is to have application to repeated surveys, over time, and in relation to many applications.

Reference

F. Murtagh, Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics, Chapman and Hall, CRC Press, 2017.