Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in DISCRETE MATHEMATICS

Electronic Notes in Discrete Mathematics 56 (2016) 37–42

www.elsevier.com/locate/endm

Direct Reading Algorithm for Hierarchical Clustering

Fionn Murtagh

Department of Computing and Mathematics
University of Derby, Derby DE22 1GB, UK
Email: fmurtagh@acm.org

Pedro Contreras

Thinking Safe Ltd., Egham TW20 0EX, UK
Email: 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 has been carried out through filtering the data in wavelet transform space. This objective is also carried out after an initial transforming of the data to a canonical order. Including high dimensional, high cardinality data, such a canonical order is provided by row and column permutations of the data matrix. In our recent work, we induce a hierarchical clustering from seriation through unidimensional representation of our observations. This linear time hierarchical classification is directly derived from the use of the Baire metric, which is simultaneously an ultrametric. In our previous work, the linear time construction of a hierarchical clustering is studied from the 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 interpretation of the hierarchy. In this work we present further case studies

http://dx.doi.org/10.1016/j.endm.2016.11.006

and examples of how this approach is very advantageous for such an ultrametric topological data mapping.

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

1 Introduction

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 [15], direct reading of clusters is carried out, through filtering the data in wavelet transform space. Then in [16], this approach is carried out after an initial transforming of the data to a canonical order. Including high dimensional, high cardinality data, such a canonical order is provided by row and column permutations of the data matrix [8,10]. A data matrix is identical to a data table or array.

The human visual system, and other sense systems including aural and haptic, are very capable of directly observing and detecting clusters, encompassing all relationship models such as proximity, adjacency, coverage and set inclusion. Such relationship models are both implicit and explicit. From [6], it is even seen how relevant such direct observation is for unconscious and subconscious reasoning processes. In that work, it is noted how human unconscious reasoning is vastly superior, computationally or processing-wise, to conscious reasoning. [6] point to how conscious thought can process between 10 and 60 bits per second. In reading, one processes about 45 bits per second, which corresponds to the time it takes to read a fairly short sentence. However the visual system alone processes about 10 million bits per second. In our work, reported on in this article, we seek to read off, and therefore observe visually, the cluster results. Setting up this representation has linear computational time, or $O(n)$, for n observations.

2 Inducing a Hierarchical Clustering from Seriation through Unidimensional Representation of Our Observations

In this section, work described in [12] is summarized.

The following is based on [5], which establishes the foundations for inducing a hierarchical clustering from a newly represented, or newly encoded, mapping of our data. This very important result allows us to seek a seriation in order