Band-based similarity indices for gene expression clustering and classification.

Aurora Torrente (etorrent@est-econ.uc3m.es)
Universidad Carlos III de Madrid

Methodology article

Keywords: data depth, similarity index, clustering, classification; gene expression

Posted Date: April 22nd, 2019

DOI: https://doi.org/10.21203/rs.2.4296/v1

Abstract

Background: The concept of depth induces an ordering from centre outwards in multivariate data. Most depth definitions are unfeasible for dimensions larger than three or four, but the Modified Band Depth (MBD) is a notable exception that has proven to be a valuable tool in the analysis of gene expression data. However, given a notion of depth, there exists no straight forward method to derive a depth-based similarity or dissimilarity measure between observations to be used in standard tasks such as clustering or classification. Results: We propose a methodology to assess a data-driven (dis)similarity between two observations, taking advantage of the bands used in the computation of the MBD. To that end, we build binary vectors associated to each observation to compute the number of times each coordinate is located between the limits of the intervals defined by all possible bands in the set. Those vectors and their Boolean products are used to derive contingency tables from which standard similarity indices can be calculated. Our approach is computationally efficient and can be applied to bands formed by any number of observations from the data set. Conclusions: We have evaluated the performance of several similarity indices with respect to that of the Euclidean distance, used as benchmark, in standard clustering and classification techniques in a variety of simulated and real data sets. Our experiments show that the technique for deriving such measures is very promising, with some of the selected indices outperforming the Euclidean distance. The use of the method is not restricted to these, the extension to other similarity coefficients being straight-forward.

Manuscript Access

Due to technical limitations, the manuscript text cannot easily be converted to HTML. Please use the PDF link in the upper right corner of the page to download the manuscript. The supplemental material associated with this manuscript is available in the section below.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.pdf
- supplement2.pdf
- supplement3.pdf
- supplement4.pdf
- supplement5.pdf