A Kolmogorov Complexity for Multidisciplinary Domains †

João S. Resende [1], Marco Almeida [2], Rolando Martins [2] and Luís Antunes [2]

[1] University of Porto/CRACS, Porto, Portugal

[2] Faculty of Sciences University of Porto, Porto, Portugal

† Presented at the Entropy 2021: The Scientific Tool of the 21st Century, 5–7 May 2021; Available online: https://sciforum.net/conference/Entropy2021/.

Kolmogorov complexity, or algorithmic information theory, measures information in an individual object as its smallest possible representation. These metrics have been applied in Computer Science and in several other scientific disciplines. It is known that this measure is not computable, however, we can approach it from above using standard compressors.

In the scope of statistical or clustering methods, it is important to measure the absolute information distance between individual objects. The Normalized Information Distance (NID) measures the minimal amount of information needed to translate between two objects, however, it is uncomputable. There is a set of metrics used to approximate the NID, such as Normalized Compression Distance (NCD), Compression-Based Dissimilarity Measure (CDM) and Lempel–Ziv Jaccard Distance (LZJD). These methods, unlike other approaches, do not require any specific background knowledge of the dataset, that is, the user of this method only needs to have some knowledge about data mining techniques or data visualization.

It is utmost important to improve current implementations in order to create a central, simply usable repository that supports multiple metrics, ensuring that a researcher can use some of the most important past work techniques to extract the most accurate information with the approach of NID. Current implementations are poorly document and lake the support to some enhancements already presented in past work literature. An example is the need to replace the compressed size with the percentage of compression in each of the files to achieve comparison with different file sizes.