SAND2019-1061C

Reordering of Genomic Data for Improved Compression-Based Inference

We have previously shown that compression algorithms can be extended in a variety of ways for useful application in machine learning and data analytics, including deception detection in text, boundary detection in audio, and anomaly detection in network traffic [refs]. Compression-based analytics rely on the data to occur locally and sequentially in order to identify patterns, which can be applied towards effective decision making. Although genomic data is nominally read as a sequence of nucleotides, the information content is neither local nor sequential; long-range interactions and regulations of genes with similar biochemical functions exist.

We study how the dependencies among single nucleotide variants (SNVs) revealed from pairwise linkage disequilibrium calculations can be used to re-order the genomic sequence and improve the ability of a compression-based analytic to identify patterns and make inferences. In particular, we apply Louvain community detection, a graph-based algorithm to reorder the SNVs into sections of highly dependent SNVs. We use prediction by partial matching (PPM), an adaptive statistical data compression technique, to train local and global models on the re-ordered sequences. We demonstrate that the re-ordering by Louvain can improve a compression-based classifier's ability to infer a population attribute. Our results are compared to standard machine learning classifiers such as Random Forests. Ultimately, understanding how to improve inference can be used to understand how to improve genomic privacy.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.