Practical Data-Dependent Metric Compression with Provable Guarantees

Ilya Razenshteyn
Columbia University

10/11/2017 at 11:00 am
CoRE 431

Abstract

How well can one compress a dataset of points from a high-dimensional space while preserving pairwise distances? Indyk and Wagner have recently obtained almost optimal bounds for this problem, but their construction (based on hierarchical clustering) is not practical. In this talk, I will show a new practical, quadtree-based compression scheme, whose provable performance essentially matches that of the result of Indyk and Wagner.

In additional to the theoretical results, we will see experimental comparison of the new scheme and Product Quantization (PQ)–one of the most popular heuristics for distance-preserving compression–on several datasets. Unlike PQ and other heuristics that rely on the clusterability of the dataset, the new algorithm ends up being more robust.

The talk is based on a joint work with Piotr Indyk and Tal Wagner.

Organizer(s): Rutgers/DIMACS Theory of Computing