Anastasios Sidiropoulos

Metric embeddings with outliers

We initiate the study of metric embeddings with *outliers*. Given some finite metric space we wish to remove a small set of points and to find either an isometric or a low-distortion embedding of the remaining points into some target metric space. This is a natural problem that captures scenarios where a small fraction of points in the input corresponds to noise.

More specifically, let $\mathcal{X} = (X, \rho)$, $\mathcal{Y} = (Y, \rho')$ be metric spaces of unit diameter. A (δ, k)-*outlier embedding* of $\mathcal{X}$ into $\mathcal{Y}$ is a mapping $\phi : X \setminus K \to Y$, with ℓ_∞-distortion at most δ, where $K \subset X$, with $|K| = k$. We say that an algorithm is a $(f(\delta), g(k))$-*relative outlier embedding*, for some functions f and g, if given $\delta, k \geq 0$, outputs some $(f(\delta), g(k))$-outlier embedding of $\mathcal{X}$ into $\mathcal{Y}$, assuming any (δ, k)-outlier embedding exists. For the case of isometries (i.e. when $\delta = 0$) we say that an algorithm is $g(k)$-*relative outlier embedding*, for some function g, if given $k \geq 0$, the algorithm outputs some $(0, g(k))$-outlier embedding of $\mathcal{X}$ into $\mathcal{Y}$, assuming any $(0, k)$-outlier embedding exists. We consider outlier embeddings into Euclidean space, ultrametrics, and trees. Our results are summarized as follows.

Embedding into Euclidean space: When the host space is the d-dimensional Euclidean space, we obtain a simple $2k$-relative outlier embedding with running time $O(n^{d+3})$. Using a randomized incremental approach, we obtain a $(3+\varepsilon)k$-relative outlier embedding with running time $\varepsilon^{-d} d^{O(1)} n^{O(1)}$, for any $\varepsilon > 0$. We also present a k-relative outlier embedding with running time $O(n^{d+3}) + 2^k n^{O(1)}$. These results are complemented by showing that for any $d \geq 2$ and for any $\varepsilon > 0$, there is no polynomial-time $(2 - \varepsilon)k$-relative outlier embedding, assuming the Unique Game Conjecture.

For non-isometric outlier embeddings, we obtain various trade-offs between the approximation factor and the running time: We give a $(O(\sqrt{\delta}), 2k)$-relative outlier embedding with running time $O(n^{d+3}) + 2^k n^{O(1)}$. With a small additional loss in the number of outliers, we obtain a $(O(\sqrt{\delta}), 2k)$-relative outlier embedding with running time $k^{O(d)} n^{O(1)}$. Finally, we derive a significantly faster $(O(\sqrt{\delta}), (2d + 2)k)$-relative outlier embedding with running time $2^{O(d)} n^{O(1)}$.

Embedding into trees and ultrametrics: We present a $3k$-relative outlier embedding into ultrametrics, and a $4k$-relative outlier embedding into trees. The running time of these algorithms is further improved to $O(n^2)$, which is optimal. We also show that for both of these problems it is NP-hard to obtain a $(2 - \varepsilon)k$-relative outlier embedding, for any $\varepsilon > 0$, assuming the Unique Games Conjecture.

Finally, for the case of non-isometries, we give a polynomial-time $(\delta \log n, 3k)$-approximation for embedding into ultrametrics.

We also discuss a brief experimental evaluation of our randomized non-isometric outlier embedding algorithms on synthetic and real-world data sets.

Joint work with Dingkang Wang and Yusu Wang.