Abstract

Near neighbor problems are fundamental in algorithms for high-dimensional Euclidean spaces. While classical approaches suffer from the curse of dimensionality, locality sensitive hashing (LSH) can effectively solve α-approximate r-near neighbor problem, and has been proven to be optimal in the worst case. However, for real-world data sets, LSH can naturally benefit from well-dispersed data and low doubling dimension, leading to significantly improved performance.

In this paper, we address this issue and propose a refined analyses for running time of approximating near neighbors queries via LSH. We characterize dispersion of data using N_β, the number of βr-near pairs among the data points. Combined with optimal data-oblivious LSH scheme, we get a $O\left(\left(1+\frac{4\sqrt{2}\alpha}{\beta}\right)^{\frac{d}{2\alpha^2}}(n+N_\beta)^{\frac{1}{2\alpha^2}}\right)$ bound for expected query time. For many natural scenarios where points are well-dispersed or lying in a low-doubling-dimension space, our result leads to sharper performance than existing worst-case analysis. This paper not only presents the *first* rigorous proof on how LSHs make use of the structure of data points, but also provides important insights into parameter setting in the practice of LSH beyond worst case. Besides, the techniques in our analysis involve a generalized version of sphere packing problem, which might be of some independent interest.