Abstract

We introduce a new variant of the nearest neighbor search problem, which allows for some coordinates of the dataset to be arbitrarily corrupted or unknown. Formally, given a dataset of n points $P = P_1, ..., P_n$ in high-dimensions, and a parameter k, the goal is to preprocess the dataset, such that given a query point q, one can compute quickly a point $p \in P$, such that the distance of the query to the point p is minimized, when ignoring the "optimal" k coordinates. Note, that the coordinates being ignored are a function of both the query point and the point returned. We present a general reduction from this problem to answering ANN queries, which is similar in spirit to LSH (locality sensitive hashing) [IM98]. Specifically, we give a sampling technique which achieves a bi-criterion approximation for this problem. If the distance to the nearest neighbor after ignoring k coordinates is r, the data-structure returns a point that is within a distance of $O(r)$ after ignoring $O(k)$ coordinates. We also present other applications and further extensions and refinements of the above result. The new data-structures are simple and (arguably) elegant, and should be practical – specifically, all bounds are polynomial in all relevant parameters (including the dimension of the space, and the robustness parameter k).