Nearest Neighbors Query Performance for Unstable Distributions

Uri Shaft
Jonathan Goldstein
Kevin Beyer

Technical Report #1388

October 1998

Nearest Neighbors Query Performance for Unstable Distributions

Uri Shaft Jonathan Goldstein Kevin Beyer

October 7, 1998

Abstract

This technical report supplements our results about the instability of some data and query distributions in high dimensions. We show that under very common conditions, if the data and query distributions are not stable, any index structure that uses convex shapes as summary of data points will perform badly.

The report uses many of the conventions and terminology described in [1]. It also relies on the main result in [1]. It is recommended to be familiar with [1] before reading this report.

1 Instability Theorem

This section repeats the main definitions and claims in [1] without the proofs.

Definition 1 *A nearest neighbor query is* **unstable** *for a given ε if the distance from the query point to most data points is less than $(1+\varepsilon)$ times the distance from the query point to its nearest neighbor.*

Definition 2:
m is the variable that our distance distributions may converge under (m ranges over all positive integers).

$F_{data_1}, F_{data_2}, \ldots$ is a sequence of data distributions.

$F_{query_1}, F_{query_2}, \ldots$ is a sequence of query distributions.

n is the (fixed) number of samples (data points) from each distribution.

$\forall m \quad P_{m,1}, \ldots, P_{m,n}$ are n independent data points per m such that $P_{m,i} \sim F_{data_m}$.

$Q_m \sim F_{query_m}$ is a query point chosen independently from all $P_{m,i}$.

$0 < p < \infty$ is a constant.

$\forall m, d_m$ is a function that takes a data point from the domain of F_{data_m} and a query point from the domain of F_{query_m} and returns a non-negative real number as a result.

$DMIN_m = \min\{d_m(P_{m,i}, Q_m) \,|\, 1 \le i \le n\}$.

$DMAX_m = \max\{d_m(P_{m,i}, Q_m) \,|\, 1 \le i \le n\}$.

Theorem 1 *Under the conditions in Definition 2, if*

$$\lim_{m\to\infty} var\left(\frac{(d_m(P_{m,1}, Q_m))^p}{\mathbf{E}\left[(d_m(P_{m,1}, Q_m))^p\right]}\right) = 0 \tag{1}$$

Then for every $\varepsilon > 0$

$$\lim_{m\to\infty} P\left[DMAX_m \le (1+\varepsilon)DMIN_m\right] = 1$$

2 Convex Index Structures

Any index structure uses the following principle:

- Data points are divided into sets (not necessarily disjoint).
- Each set has some information associated with it. The information summarizes some common quality of the points in that set.
- Query processing involves looking at the information associated with a set. We decide based on that information if any of the points in the set might be a valid answer. Only then do we fetch the set and check all the points in it.

We may build another index structure for the information associated with all the sets. This yields a hierarchical indexing structure. For example, the leaf nodes of a B-tree contain sets of points. The information for each leaf node is an interval s.t. all points in the leaf are inside that interval. Each level in the tree is an index into the level below it.

We call an indexing structure *convex* if the following hold:

- The information stored for a set is a convex region of space.
- Given only the convex region associated with a set and a point that belongs to that region we can't exclude the possibility that the point is a data point in that set. (I.e., if the query region overlaps the convex region associated with a set we must fetch the points within that set.)

3 Performance Theorem

We show that when using the Euclidian distance metric and assuming that the data distribution and query distribution are the same, instability means that the performance of any convex indexing structure degenerates into scanning the entire data set for NN queries.

Theorem 2 *Suppose the conditions in Definition 2 are satisfied and for any* $\varepsilon > 0$

$$\lim_{m\to\infty} P\left[\left|\frac{DMAX_m}{DMIN_m} - 1\right| \leq \varepsilon\right] = 1. \tag{2}$$

If $F_{data} = F_{query}$ *and* d_m *is the Euclidian distance metric then the probability that the number of points fetched using any convex indexing structure is* n *converges to 1 as* m *goes to* ∞.

Proof Instead of d_m we use d as the Euclidian distance metric and d^2 as the square distance. We denote the query point $\vec{Q}_m$ by $\vec{P}_{m,0}$. For all $0 \leq i \leq n$ define the folowing random variables:

$$\mathrm{DMAX}_{mi} = \max\left\{d(\vec{P}_{m,i}, \vec{P}_{m,j}) \,|0 \leq j \leq n \text{ and } i \neq j\right\}$$

$$\mathrm{DMIN}_{mi} = \min\left\{d(\vec{P}_{m,i}, \vec{P}_{m,j}) \,|0 \leq j \leq n \text{ and } i \neq j\right\}$$

Since the points in S_m are iid, we can treat any $\vec{P}_{m,i}$ as the query point and the rest of the set as the data points and get that for all $0 \leq i \leq n$ and for all $\varepsilon > 0$

$$\lim_{m\to\infty} P\left[\left|\frac{\mathrm{DMAX}_{mi}}{\mathrm{DMIN}_{mi}} - 1\right| \leq \varepsilon\right] = 1$$

Defined the minima and maxima of all these distances as:

$$\text{RMAX}_m = \max\{\text{DMAX}_{mi} | 0 \leq i \leq n\}$$

$$\text{RMIN}_m = \min\{\text{DMIN}_{mi} | 0 \leq i \leq n\}$$

Part 1:
We'll now show that for all $\varepsilon > 0$

$$\lim_{m\to\infty} P\left[\left|\frac{\text{RMAX}_m}{\text{RMIN}_m} - 1\right| \leq \varepsilon\right] = 1$$

For each $0 \leq i \leq n$ define $r_i = \text{DMAX}_{mi}/\text{DMIN}_{mi}$. We know that $r_i \to_p 1$. (This is the same as saying that for all $\varepsilon > 0$ holds $\lim_{m\to\infty} P[|r_i - 1| \leq \varepsilon] = 1$.) Using Slutsky's theorem we have that for all $0 \leq i, j \leq n$ holds $r_i r_j \to_p 1$.

For each m we have three possible cases:

Case 1: $\text{RMAX}_m/\text{RMIN}_m = 1$.

Case 2: Exist distinct points $\vec{P}_{m,i}, \vec{P}_{m,j}, \vec{P}_{m,k}$ s.t.

$$\frac{\text{RMAX}_m}{\text{RMIN}_m} = \frac{d(\vec{P}_{m,i}, \vec{P}_{m,j})}{d(\vec{P}_{m,i}, \vec{P}_{m,k})}$$

In this case we have that $\text{RMAX}_m/\text{RMIN}_m \leq r_i \leq r_i r_i$.

Case 3: Exist distinct points $\vec{P}_{m,i}, \vec{P}_{m,j}, \vec{P}_{m,k}, \vec{P}_{m,l}$ s.t.

$$\frac{\text{RMAX}_m}{\text{RMIN}_m} = \frac{d(\vec{P}_{m,i}, \vec{P}_{m,j})}{d(\vec{P}_{m,k}, \vec{P}_{m,l})}$$

Therefore,

$$\frac{\text{RMAX}_m}{\text{RMIN}_m} = \frac{d(\vec{P}_{m,i}, \vec{P}_{m,j})}{d(\vec{P}_{m,i}, \vec{P}_{m,k})} \frac{d(\vec{P}_{m,i}, \vec{P}_{m,k})}{d(\vec{P}_{m,k}, \vec{P}_{m,l})} \leq r_i r_k$$

Of course, in all cases $P[\text{RMAX}_m/\text{RMIN}_m \geq 1] = 1$. Therefore, for all m holds

$$1 \leq \frac{\text{RMAX}_m}{\text{RMIN}_m} \leq \max\{r_i r_j | 0 \leq i, j \leq n\}$$

Using Slutsky's theorem we have $\max\{r_i r_j | 0 \leq i, j \leq n\} \to_p 1$. Therefore $\text{RMAX}_m/\text{RMIN}_m \to_p 1$. In other words, for all $\varepsilon > 0$

$$\lim_{m\to\infty} P\left[\left|\frac{\text{RMAX}_m}{\text{RMIN}_m} - 1\right| \leq \varepsilon\right] = 1$$

Part 2:
Consider the case when $\text{RMAX}_m \leq \frac{\sqrt{5}}{2}\text{RMIN}_m$ (i.e., $\varepsilon = (\sqrt{5}/2) - 1$). We'll show that in this case the distance from the query point to any convex region that includes at least two data points is at most RMIN_m. Considering that the distance from the query point to its nearest neighbor is at least RMIN_m we conclude that any convex region that includes at least as close to the query point as its nearest neighbor. This means that for any convex indexing structure, processing the query involves fetching all the data points.

To show that the distance of the query point to a convex region is at most RMIN_m we'll show that the distance of the query point to a specific point in the convex region is at most RMIN_m. Take two distinct data points $\vec{X}, \vec{Y}$ in the convex region. We'll show that $d(\vec{Q}, (\vec{X} + \vec{Y})/2) \leq \text{RMIN}_m$. Note that the point $(\vec{X} + \vec{Y})/2$ is the midpoint between $\vec{X}$ and $\vec{Y}$, so by definition of convex, it is in the convex region.

$$d^2\left(\vec{Q}, \frac{\vec{X}+\vec{Y}}{2}\right) =$$

$$= \sum_{i=1}^{k} \left[Q_i - \frac{X_i+Y_i}{2}\right]^2 =$$

$$= \sum_{i=1}^{k} \left[Q_i^2 - 2Q_i\frac{X_i+Y_i}{2} + \left(\frac{X_i+Y_i}{2}\right)^2\right] =$$

$$= \sum_{i=1}^{k} \left[\frac{1}{2}(Q_i - X_i)^2 + \frac{1}{2}(Q_i - Y_i)^2 - \frac{1}{4}(Y_i - X_i)^2\right] =$$

$$= \frac{1}{2}d^2(\vec{Q}, \vec{X}) + \frac{1}{2}d^2(\vec{Q}, \vec{Y}) - \frac{1}{4}d^2(\vec{Y}, \vec{X}) \leq$$

$$\leq \frac{1}{2}\mathrm{RMAX}_m{}^2 + \frac{1}{2}\mathrm{RMAX}_m{}^2 - \frac{1}{4}\mathrm{RMIN}_m{}^2 \leq$$

$$\leq \left(\frac{\sqrt{5}}{2}\mathrm{RMIN}_m\right)^2 - \frac{1}{4}\mathrm{RMIN}_m{}^2 =$$

$$= \mathrm{RMIN}_m{}^2$$

Therefore $d(\vec{Q}, (\vec{X}+\vec{Y})/2) \leq \mathrm{RMIN}_m$. Since the probability of the event $\mathrm{RMAX}_m \leq \frac{\sqrt{5}}{2}\mathrm{RMIN}_m$ converges to 1 as m goes to ∞ we get that the probability that a nearest neighbor query using a convex indexing structure will result in fetching all data points goes to 1 as m goes to ∞. ■

References

[1] Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When Is Nearest Neighbors Meaningful? Technical Report No. TR1377, Computer Sciences Dept., Univ. of Wisconsin-Madison, June 1998