Approximate Nearest Neighbor Search in Euclidean Space(s)

Piotr Indyk

Stanford University

PLEB

Point Location in Equal Balls:

Given: metric space $D = (X, d)$, n-set $P \subset X$, $r > 0$

Goal: a data structure which for any query $q \in X$

- if there is $p \in P$ s.t. $d(q, p) \leq r$: return p
- otherwise: return NONE

ϵ-PLEB:

Goal: a data structure which for any query $q \in X$

- if $d(q, P) \leq r$: return p s.t. $d(q, p) \leq r(1 + \epsilon)$
- if $d(q, P) > (1 + \epsilon)r$: return NONE
- otherwise: return anything

Reductions of ϵ-NN to ϵ-PLEB

- binary search:
  - create PLEBs with $r = r_0, r_0(1+\epsilon), r_0(1+\epsilon)^2 \ldots$
  - binary search for smallest r such that $d(q, P) \leq r$
  - overhead: $O(\log R)$ (space) and $O(\log \log R)$ (query time)

- Ring/cover-tree [IM'98]
  - storage/query time overhead: $O(\log^{O(1)} n)$
  - deterministic

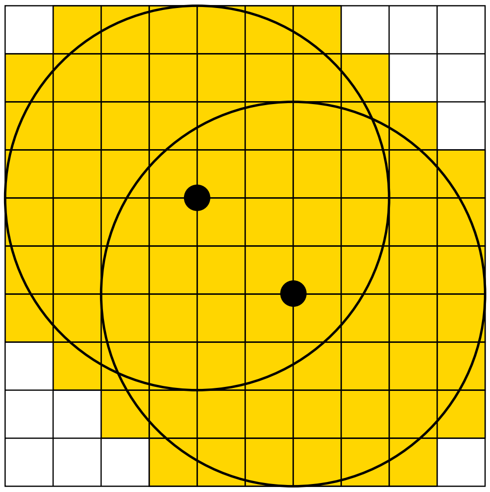
- other reductions [KOR'98,IM'98]

R : largest/smallest distance ratio

ϵ-PLEB: results

Query	Storage	Technique
d	$\epsilon^{-d}n$	partitioning
$d\log n$	$n^{O(\log(1/\epsilon)/\epsilon^2)}$ $n^{O(1/\epsilon^2)}$	JL Lemma $\rightarrow$ $d = O(\log n)$ [IM'98] via Hamming space [KOR'98]
$n^{1/(1+\epsilon)}\log n$	$dn + n^{1+1/(1+\epsilon)}$	Locality Sensitive Hashing [IM'98]

PLEB: method



Wanted:

- small cube diameter
- small storage

PLEB: analysis

Radius$= 1$

Side $= \epsilon/\sqrt{d}$

$$\begin{aligned}
\text{Cube diameter} &= \sqrt{d \cdot (\epsilon/\sqrt{d})^2} \\
&= \epsilon \\
\text{Number of cubes} &\approx V^d(\sqrt{d}/\epsilon) \\
&= \frac{2\pi^{d/2}}{d\Gamma(d/2)}(\sqrt{d}/\epsilon)^d \\
&\approx \epsilon^{-d}\frac{(\pi^{1/2}d^{1/2})^d}{(d/2)^{d/2}/e^{d/2}} \\
&= \left(\frac{(2e\pi)^{1/2}}{\epsilon}\right)^d \\
&= O(\epsilon)^{-d}
\end{aligned}$$

Johnson-Lindenstrauss lemma

Lemma[Frankl-Maehara]: For

- any $0 < \epsilon < \frac{1}{2}$
- any (sufficiently large) set S of points in $\Re^d$
- $k = \lceil 9(\epsilon^2 - 2\epsilon^3/3)^{-1} \ln |S| \rceil + 1$

there exists a linear map $f : S \to \Re^k$ such that for all $u, v \in S$

$$(1-\epsilon)||u-v||^2 < ||f(u) - f(v)||^2 < (1+\epsilon)||u-v||^2.$$

In fact, random projection is good with constant probability.

Main tool

Lemma[IM'98,Sch'99,DG'99,AV'99]: Assume

- u: a unit vector in $\Re^d$
- $U = (U_1, \ldots, U_k)^T$: k random d-dimensional vectors with coordinates from $N(0,1)$ (k even)
- $X = (X_1, \ldots, X_k) = U \cdot u$
- $L = L(u) = ||X||^2$

Then, for any $\beta > 1$,

1. $E(L) = k$
2. $\Pr[L \geq \beta k] < O(k) \times \exp(-\frac{k}{2}(\beta - (1 + \ln \beta)))$
3. $\Pr[L \leq k/\beta] < O(k) \times \exp(-\frac{k}{2}(\beta^{-1} - (1 - \ln \beta)))$

Properties

- By the spherical symmetry of $N^d(0,1)$ each X_i is distributed as $N(0,1)$

- if $Y_i = X_{2i-1}^2 + X_{2i}^2$, then Y_i has exponential distribution with $\lambda = \frac{1}{2}$

- $L = \sum Y_i$ has Γ distribution with $\alpha = \frac{1}{2}$ and $v = k/2$

- $\Pr[L \geq \beta k] = \Pr[P_{\beta k}^{1/2} \leq v - 1]$ where P_t^α has Poisson distribution with parameter αt, i.e.

$$\Pr[P_t^\alpha \leq v - 1] = \sum_{i=0}^{v-1} e^{-\alpha t} \frac{(\alpha t)^i}{i!}$$

Proof

Part 1: $E(L) = \sum_{i=1}^{k/2} E(Y_i) = (k/2) \times 2 = k$

Part 2:

$$
\begin{aligned}
\Pr[L \geq \beta k] &= \sum_{i=0}^{v-1} e^{-\beta v} \frac{(\beta v)^i}{i!} \\
&\leq v e^{-\beta v} \frac{(\beta v)^v}{v!} \\
&\leq v e^{-\beta v} \frac{(\beta v)^v}{\frac{v^v}{e^v}} \\
&= v(e^{-\beta} \beta e)^v \\
&= v e^{-v(\beta - (1 + \ln \beta))}
\end{aligned}
$$

Part 3: similar.

Algorithmic Implications

- for fixed q, setting $k = O(\log n/\epsilon^2)$ approx. preserves the distances from q to P with high probability

- can replicate data structures d times for correctness for all queries (see also [KOR'98])

- don't know if correct (Monte Carlo)

Extensions[I2'99]:

- $(1 + \epsilon)$-approximate Las Vegas

- $(2 + \epsilon)$-approximate deterministic (for l_1)

Locality-Sensitive Hashing

Fix: distance measure D

Family $\mathcal{H} = \{h : X \to Y\}$ is (r_1, r_2, P_1, P_2)-*sensitive* for D if for all $q, p \in X$

- if $d(p,q) \leq r_1$ then $\Pr_{\mathcal{H}}[h(q) = h(p)] \geq P_1$,
- if $d(p,q) > r_2$ then $\Pr_{\mathcal{H}}[h(q) = h(p)] \leq P_2$.

Assume: $P_1 > P_2$ and $r_1 < r_2$.

Example: Hamming metric

- $\mathcal{H} = \{h(b_1 \ldots b_d) = b_i,$ for $i = 1, \ldots, d\}$
  (i.e., sample one bit at random)
- $\Pr_{\mathcal{H}}[h(q) = h(p)] = 1 - \dfrac{d(p,q)}{d}$

Using LSH for PLEB

To solve ϵ-PLEB :

- Set $r_1 \leftarrow r$
- Set $r_2 \leftarrow (1+\epsilon)r$

Define $\mathcal{G} = \{g \mid g(p) = h_1(p) \cdot h_2(p) \cdot \ldots \cdot h_k(p)\}$

(Hamming metric $\Rightarrow$ sample k random bits)

Preproc: $\forall p \in P$, store p at $g_1(p)$, $g_2(p)$, $\ldots$, $g_l(p)$

$\Rightarrow$ Storage: $O(dn + nl)$

Query q: check at $g_1(q)$, $g_2(q)$, $\ldots$, $g_l(q)$

stop after $O(l)$ steps $\Rightarrow$ Time: $O(dl)$

LSH: Theorem

Question: How many indices l do we need for correctness?

Theorem: Setting $l = n^\rho$ for $\rho = \frac{\log 1/P_1}{\log 1/P_2}$ is sufficient with constant probability.

(Hamming metric $\Rightarrow \rho = 1/(1+\epsilon)$)

LSH: Analysis

Define:

- p^* - a point s.t. $d(q, p^*) \leq r$
- FAR(q) - all p with $d(q, p) > (1 + \epsilon)r$
- BUCKET$_j(q)$ - all p s.t. $g_j(p) = g_j(q)$

Events:

- $E_1 : \sum_{j=1}^{l} |\text{FAR}(q) \cap \text{BUCKET}_j(q)| \leq 3l$
- $E_2 : g_j(p^*) = g_j(q)$ for some g_j, $1 \leq j \leq l$

Will show: $\Pr[E_1] > x > \frac{1}{2}$ and $\Pr[E_2] > y > \frac{1}{2}$

Thus: $\Pr[E_1 \wedge E_2] > x + y - 1 > 0$

LSH: Bad collisions

Let $p \in \mathsf{FAR}(q)$. Then

$$\Pr[p \in \mathsf{BUCKET}_j(q)] \leq P_2^k$$

For $k = \log_{\frac{1}{P_2}} n$

$$\Pr[p \in \mathsf{BUCKET}_j(q)] \leq P_2^{\log_{\frac{1}{P_2}} n} = \frac{1}{n}$$

Thus

$$\begin{aligned} E[|\mathsf{FAR}(q) \cap \mathsf{BUCKET}_j(q)|] &\leq n \cdot \tfrac{1}{n} = 1 \\ E[\sum_{j=1}^{l} |\mathsf{FAR}(q) \cap \mathsf{BUCKET}_j(q)|] &\leq l \end{aligned}$$

By Markov inequality

$$\Pr[\sum_{j=1}^{l} |\mathsf{FAR}(q) \cap \mathsf{BUCKET}_j(q)| > 3l] \leq \frac{1}{3}$$

LSH: Good collision

For any g_j, $1 \leq j \leq l$

$$\Pr[g_j(p^*) = g_j(q)] \geq P_1^k = P_1^{\log_{1/P_2} n} = n^{-\frac{\log 1/P_1}{\log 1/P_2}} = n^{-\rho}$$

For $l = n^\rho$ we have

$$\begin{aligned}
\Pr[E_2] &\geq 1 - (1 - \Pr[g_j(p^*) = g_j(q)])^l \\
&\geq 1 - (1 - n^{-\rho})^{n^\rho} \\
&\geq 1 - 1/e > \frac{1}{2}
\end{aligned}$$

Another example of LSH

Let $D(A, B) = \frac{|A \cap B|}{|A \cup B|}$.

Fact: [Broder, Glassman, Manasse, Zweig '97]:

$$\mathcal{H} = \{h_\pi : h_\pi(A) = \max_{a \in A} \pi(a),\ \pi \text{ is a permutation of } X\}$$

is (r_1, r_2, r_1, r_2)-sensitive.

Dot product of boolean vectors.

Extension to l_p

1. Extension to l_1:

Embed $\{1 \dots M\}^d$ into $\{0,1\}^{dM}$ via unary mapping:

$$
\begin{aligned}
(x_1, \dots, x_d) &\rightarrow \underbrace{1 \dots 1}_{x_1}\underbrace{0 \dots 0}_{d-x_1} \dots \underbrace{1 \dots 1}_{x_d}\underbrace{0 \dots 0}_{d-x_d} \\
(y_1, \dots, y_d) &\rightarrow \underbrace{1 \dots 1}_{y_1}\underbrace{0 \dots 0}_{d-y_1} \dots \underbrace{1 \dots 1}_{y_d}\underbrace{0 \dots 0}_{d-y_d}
\end{aligned}
$$

2. Extension to l_p, $p \in [1,2]$:

   Embed l_p into l_1 using [Johnson, Schechtman '82].

Other extensions

- dynamic setting
- closest pair
  - hierarchical clustering
  - greedy matching
- bichromatic closest pair
  - approximate MST