Approximate Clustering via Core-Sets

Mihai Badoiu, Sariel Har-Peled, Piotr Indyk

STOC, 2002

Contribution

- Several results on clustering P in R^d using coresets, where d is large
- Corsets size independent of d
- Previous work had coresets sizes dependend polynomially or exponentially on d

Outline

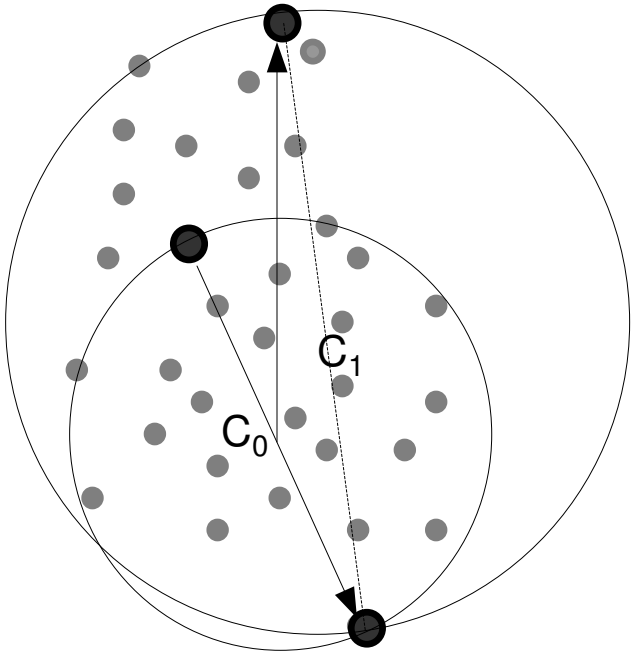
- the core-set result
- k-center clustering
- k-median clustering

An incremental algorithm for minimum enclosing ball

- an ε–approximation to the minimum-enclosing ball of a set of n points $P \subset R^d$, with a coreset of size $O(1/\varepsilon^2)$
- Number of iterations: $O(1/\varepsilon^2)$
- The bound was later improved to $O(1/\varepsilon)$

An incremental algorithm for minimum enclosing ball

- Initial set S_0 contains an arbitrary point of P
- In iteration i, compute the minimum enclosing ball of S_{i-1}, $B(c_{i-1}, r_{i-1})$
  - the $(1+\varepsilon)$-expansion of the ball contains P, terminate;
  - Otherwise, $S_{i+1} = S_i \cup \{p\}$, where point $p \in P$ is the farthest from c_{i-1}



Claim: $r_{i+1} \geq \left(1 + \frac{\varepsilon^2}{16}\right) r_i$

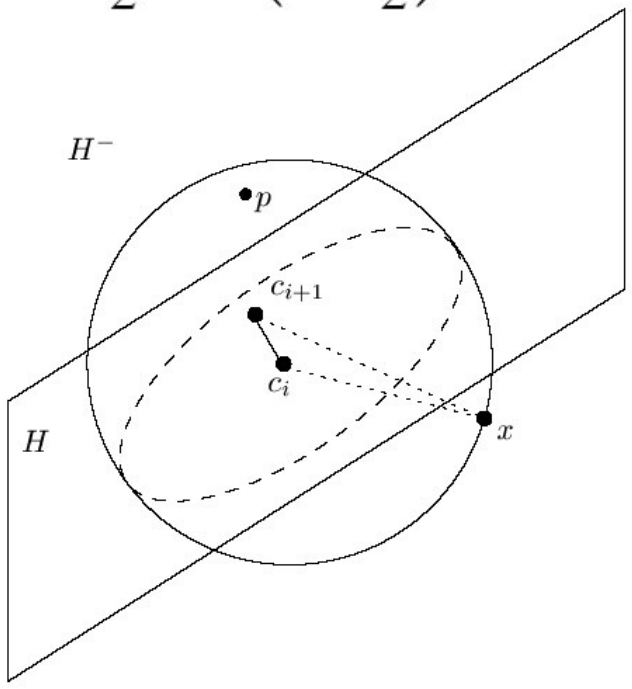
1) If $\|c_i - c_{i+1}\| < (\varepsilon/2) r_i$,

$$\|p - c_{i+1}\| \geq \|p - c_i\| - \|c_i - c_{i+1}\| \geq (1+\varepsilon) r_i - \frac{\varepsilon}{2} r_i = \left(1 + \frac{\varepsilon}{2}\right) r_i.$$

done.

2) if $\|c_i - c_{i+1}\| \geq (\varepsilon/2) r_i$

$$r_{i+1} \geq \|c_{i+1} - x\| \geq \sqrt{r_i^2 + \frac{\varepsilon^2}{4} r_i^2} \geq \left(1 + \frac{\varepsilon^2}{16}\right) r_i$$



H is the (d-1)-dimentional hyperplane that passes through c_i and $H \perp c_i c_{i+1}$. H^- is the open half-space having p inside it.

k-center clustering

Def: to find a set of k points in R^d such that the maximum distance of points in $P \subset R^d$ to their closest center is minimized.

1-center clustering

For 1-center, it is the minimum enclosing ball problem.

Running time:

this requires computing $O(1/\varepsilon^2)$ times $(1+\varepsilon)$-approximate enclosing ball of at most $O(1/\varepsilon^2)$ points in $O(1/\varepsilon^2)$ dimensions. By known convex optimization techniques, we have $O(dn/\varepsilon^2 + (1/\varepsilon)^{10}\log(1/\varepsilon))$.

2-center clustering

- Start from two empty sets of points + an oracle, get the same running time as that in 1-center.
- To remove the oracle, enumerate all possible guesses. It amounts to running the algorithm $2^{O(1/\varepsilon^2)}$ (all possible guesses) times. Overall: $dn2^{O(1/\varepsilon^2)}$

k-center clustering

Running time: $dn2^{O((k\log k)/\varepsilon^2)}$, where now each guess is a number between 1 and k, and we have to enumerate $O(k/\varepsilon^2)$ guesses.

Note

- Q is a weaker coreset than its lower dimensional counterpart: it is not necessarily true that the (1+ε)-expansion of *any* ball containing Q contains P.
- The *smallest* ball containing Q, when (1+ε)-expanded, contains P.

k-median clustering: definitions

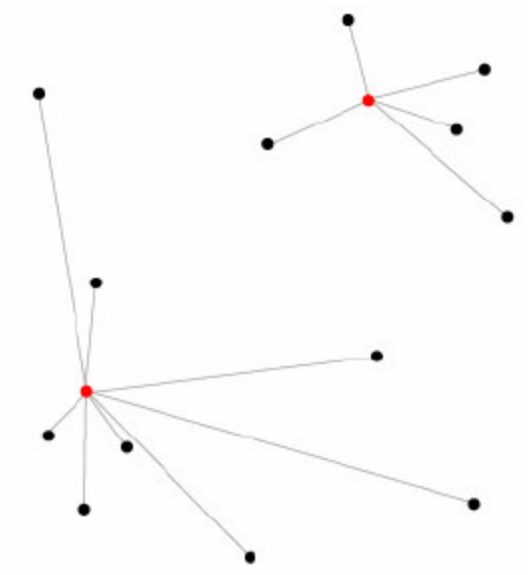
let $\mathrm{dist}(X,p) = \min_{x \in X} ||x - p||$

the optimal price of the k-median problem

$$\mathrm{med}_{\mathrm{opt}}(P,k) = \min_{K \subseteq \mathbb{R}^d, |K|=k} \sum_{p \in P} \mathrm{dist}(K,p)$$

the average radius of the k-median clustering

$$\mathrm{AvgMed}(P,k) = \mathrm{med}_{\mathrm{opt}}(P,k)/|P|$$



Sketch

- For a random sample X from P of size $O(1/\varepsilon^3 \log 1/\varepsilon)$, with a constant probability:
  - The flat span(X) contains an $(1+\varepsilon)$-approximate 1-median for P
  - X contains a point close to the center of a 1-median of P
- Thus, generate a small number of candidate points on span(X), s.t. one of them is an $(1+\varepsilon)$-approximate 1-median for P
- For k-median clustering
  - Guess the average radius and cardinality of the heaviest cluster, generate a candidate set for centers for this cluster using random sampling
  - Recurs on the remaining points
  - Running time: $2^{(k/\varepsilon)^{O(1)}} d^{O(1)} n \log^{O(k)} n$, correct with high probability

Existence proof

1. Randomly sample P. Partition the random samples into rounds: a round continues till a point that has some required property. The first round continues till s_i, s.t. $\| s_i - c_{opt} \| \leq 2R$ (with prob. $\geq ½$)
2. Let $F_i = \text{span}(s_1, s_2, \ldots, s_i)$. If $\text{dist}(F_i, c_{opt}) \leq \varepsilon R$, done. $\text{Proj}(c_{opt}, F_i)$ is the approximation; else, start another round.

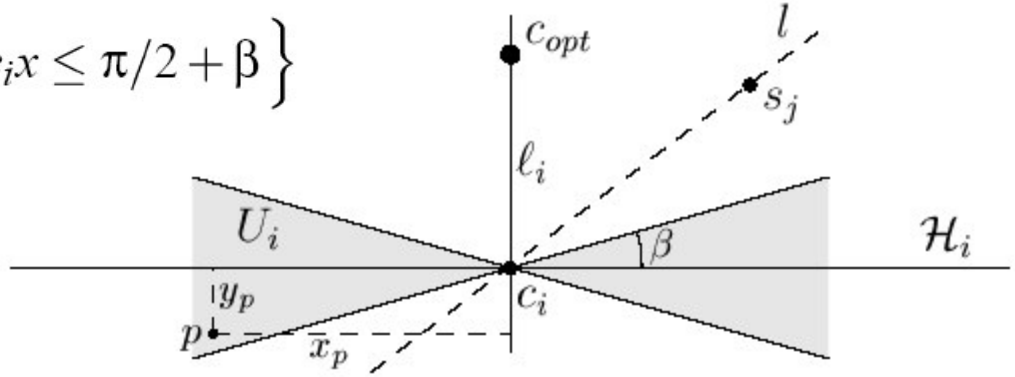
Note: $d_{i+1} = \text{dist}(F_{i+1}, c_{opt}) \leq d_i = \text{dist}(F_i, c_{opt})$

Existence proof

$$U_i = \left\{ x \,\middle|\, x \in \mathbb{R}^d \text{ s.t. } \pi/2 - \beta \le \angle c_{opt} c_i x \le \pi/2 + \beta \right\}$$

$$c_i = \text{proj}(c_{opt}, F_i)$$

$$\beta = \varepsilon/16$$



$\mathcal{H}_i$ be the $(d-1)$-dimensional hyperplane passing through c_i and perpendicular $c_i c_{opt}$.

If $p \in U_i$,

$$\|p - c_i\| \le x_p + y_p \le (1 + \varepsilon/4) \left\| p - c_{opt} \right\|.$$

Thus, if the number of points in $Q_i = P \setminus U_i$ is smaller than $n\varepsilon/4$, done.

perform a round of random sampling till we pick a point that is in Q_i , $s_j \in Q_i$

$$\text{dist}(F_j, c_{opt}) \quad \le \quad (1 - \beta^2/4)\, \text{dist}(F_i, c_{opt}).$$

Running time

as long as $|Q_i| \geq \varepsilon n/4$, the probability of success is at least $\varepsilon/4$.

The number of rounds we need is

$$M = \left\lceil \log_{1-\beta^2/4} \frac{\varepsilon}{2} \right\rceil = \left\lceil \frac{\log(\varepsilon/2)}{\log(1-\beta^2/4)} \right\rceil = O\left(\frac{1}{\varepsilon^2} \log \frac{2}{\varepsilon} \right).$$

then $E[X] = O(1/\varepsilon^3 \log 1/\varepsilon)$

X is the number of random samples till get M successes.

by the Markov inequality, that with constant probability, ...

Sketch

- For a random sample X from P of size $O(1/\varepsilon^3 \log 1/\varepsilon)$, with a constant probability:
  - The flat span(X) contains an $(1+\varepsilon)$-approximate 1-median for P
  - X contains a point close to the center of a 1-median of P
- **Generate a small number of candidate points on span(X), s.t. one of them is an $(1+\varepsilon)$-approximate 1-median for P**
- For k-median clustering
  - Guess the average radius and cardinality of the heaviest cluster, generate a candidate set for centers for this cluster using random sampling
  - Recurs on the remaining points
  - Running time: $2^{(k/\varepsilon)O(1)} d^{O(1)} n \log^{O(k)} n$, correct with high probability

k-median clustering

Theorem 3.4 *Let P be a normalized set of n points in $\mathbb{R}^d$, $1 > \varepsilon > 0$, and let R be a random sample of $O(1/\varepsilon^3 \log 1/\varepsilon)$ points from P. Then one can compute, in $O\left(d2^{O(1/\varepsilon^4)} \log n\right)$ time, a point-set $S(P,R)$ of cardinality $O\left(2^{O(1/\varepsilon^4)} \log n\right)$, such that with constant probability (over the choice of R), there is a point $q \in S(P,R)$ such that $cost(q,P) \leq (1+\varepsilon)\mathrm{med}_{\mathrm{opt}}(P,1)$.*

Proof: Let's assume that we had found a t such that $t/2 \leq \mathrm{AvgMed}(P,1) \leq t$. Clearly, we can find such a t by checking all possible values of $t = 2^i$, for $i = 0, \ldots, O(\log n)$, as P is a normalized point-set (see Lemma 3.3).

For each point of $p \in R$, we construct a grid $G_p(t)$ of side length $\varepsilon t/(10|R|) = O(t\varepsilon^4 \log(1/\varepsilon))$ centered at p on H, and let $B(p,3t)$ be a ball of radius $2t$ centered at p. Finally, let $S'(p,t) = G_p(t) \cap B(p,3t)$. Clearly, if $t/2 \leq \mathrm{AvgMed}(P,1) \leq t$, and $\left\|p - c_{opt}\right\| \leq 2t$, then there is a point $q \in S'(p,t)$ such that $\mathrm{cost}(q,P) \leq (1+\varepsilon)\mathrm{med}_{\mathrm{opt}}(P,1)$.

Let $S(P,R) = \bigcup_{i=0}^{O(\log n)} \bigcup_{p \in R} S'(p,2^i)$. Clearly, $S(P,R)$ is the required point-set

Sketch

- For a random sample X from P of size $O(1/\varepsilon^3 \log 1/\varepsilon)$, with a constant probability:
  - The flat span(X) contains an $(1+\varepsilon)$-approximate 1-median for P
  - X contains a point close to the center of a 1-median of P
- Thus, generate a small number of candidate points on span(X), s.t. one of them is an $(1+\varepsilon)$-approximate 1-median for P
- **For k-median clustering**
  - **Guess the average radius and cardinality of the heaviest cluster, generate a candidate set for centers for this cluster using random sampling**
  - **Recurs on the remaining points**
  - **Running time: $2^{(k/\varepsilon)^{O(1)}} d^{O(1)} n \log^{O(k)} n$, correct with high probability**

2-median clustering

Theorem 3.5 *For any point-set $P \subset \mathbb{R}^d$ and $0 < \varepsilon < 1$, a $(1+\varepsilon)$-approximate 2-median for P can be found in*

$$O(2^{(1/\varepsilon)^{O(1)}} d^{O(1)} n \log^{O(1)} n)$$

expected time, the results are correct with high-probability.

Case 1: $|p_1| \geq |p_2| \geq |p_1|\varepsilon$

Sample a random set of points R of cardinality $O(1/\varepsilon^4 \log 1/\varepsilon)$.

Check all possible partitions of R into R_1 and R_2.

Generate S_1, S_2 that with constant probability contains the two centers.

Checking each pair of centers takes O(nd) time, $O(|S_1||S_{2|})$ pairs.

$O(nd2^{O(1/\varepsilon^4 \log 1/\varepsilon)} \log^2 n)$

k-median clustering

Case 2: $|p_1|\varepsilon \geq |p_2|$

Generate a set C_1 of candidates for cluster P_1, then get $c`_1$.

To get $c``_2$, remove some elements from P_1 to facilitate random sampling from P_2 (Details in the paper).

Add V (guessed $|P_2|$) copies of $c`_1$ to P_1.

Apply case 1.

Running time: $\mathbf{2^{(k/\varepsilon)O(1)}d^{O(1)}nlog^{O(k)}n}$