Coresets for k-Means and k-Median Clustering and their Applications

SARIEL HAR-PELED   SOHAM MAZUMDAR

Tushar Marda | 2012CS10259 | COL870

Introduction

This paper discusses the construction of (k, ε)-coresets for the k-means and k-median clustering problems for a set of n points in d-dimensional space.

Definitions

- For a point set X, and a point p, both in $\mathbb{R}^d$, let $\boldsymbol{d}(p, X) = min_{x \in X} \|xp\|$ denote the *distance of p from X*.
- For a weighted point set P with points from $\mathbb{R}^d$, with a weight function $w: P \to \mathbb{Z}^+$ and any point set C, we define $\nu_c(P) = \Sigma_{p \in P} w(p) \boldsymbol{d}(p, C)$ as the *price* of the k-median clustering provided by C.
- Let $\nu_{opt}(P, k) = \min_{C \subseteq \mathcal{R}^d, |C| = k} \nu_C(P)$ denote the price of *optimal k-median* clustering for P.
- Similarly, $\mu_C(P) = \Sigma_{p \in P} w(p) \big(\boldsymbol{d}(p, C)\big)^2$ denotes the *price* of the k-means clustering of P provided by C, and $\mu_{opt}(P, k) = \min_{C \subseteq \mathcal{R}^d, |C| = k} \mu_C(P)$ denotes the price of the *optimal k-means* clustering for P.
- $R^{\nu}_{opt}(P, k) = \frac{\nu_{opt}(P,k)}{|P|}$ denotes the *average radius* of P, and $R^{\mu}_{opt}(P, k) = \sqrt{\frac{\mu_{opt}(P,k)}{|P|}}$ is the *average means radius* of P.
- For a weighted point set $P \subseteq \mathbb{R}^d$, a weighted set $S \subseteq \mathbb{R}^d$ is a (k, ϵ)-*coreset* of P for the k-median clustering, if for any set C of k points in $\mathbb{R}^d$, the following relation holds true $(1 - \epsilon)\nu_C(P) \leq \nu_C(S) \leq (1 + \epsilon)\nu_C(P)$.
- Similarly, $S \subseteq \mathbb{R}^d$ is a (k, ϵ)-*coreset* of P for the k-means clustering, if for any $C \subseteq \mathbb{R}^d$, we have $(1 - \epsilon)\mu_C(P) \leq \mu_C(S) \leq (1 + \epsilon)\mu_C(P)$.

Coreset for k-Median

Let P be a set of n points in $\mathbb{R}^d$, and $A = \{x_1, \dots, x_m\}$ be a point set, such that $\nu_A(P) \leq c\nu_{opt}(P, k)$, where c is a constant. Let P_i be the points of P having x_i as their nearest neighbor in A, for $i = 1, \dots, m$. Let $R = \frac{\nu_A(P)}{cn}$. For any $p \in P_i$, we have $\|px_i\| \leq \nu_A(P) = cnR$ for $i = 1, \dots, m$. Now, we construct an exponential grid around each x_i as follows. Let $Q_{i,j}$ be an axis parallel square with side length $2R2^j$ starting from $j = 0$ centered at x_i. Since $\|px_i\| \leq cnR$, $\max j = \max \left\lceil \lg\left(\frac{\|px_i\|}{2R}\right) \right\rceil = \left\lceil \lg\left(\frac{cn}{2}\right) \right\rceil$. So, $j \leq M$ where $M = \left\lceil \lg\left(\frac{cn}{2}\right) \right\rceil$. Next, let $V_{i,0} = Q_{i,0}$, and $V_{i,j} = Q_{i,j} \backslash Q_{i,j-1}$, for $j = 1, \dots, M$. Partition $V_{i,j}$ into a grid with side length $r_j = \frac{\epsilon R 2^j}{10cd}$, and let G_i be the resultant exponential grid for $V_{i,0}, \dots, V_{i,M}$. Next, for every point in P_i, compute the grid cell in G_i that contains it. For every non-empty grid cell in G_i, pick an arbitrary point of P_i inside it as the representative of all the points inside that cell, and set its weight equal to the number of points of P_i inside that cell. Let the resultant set be S_i for $i = 1, \ldots, m$, and let $S = \bigcup_i S_i$. Then, S is a (k, ϵ)-*coreset* of P for the k-median clustering.

SIZE OF CORESET

Every cell in G_i contributes at most one point to S_i, so $|S_i| \le |G_i| = \sum_j |V_{i,j}|$. To calculate this value-

$$|V_{i,0}| = \left(\frac{2R}{r_0}\right)^d = 2^d \left(\frac{10cd}{\epsilon}\right)^d$$

$$|V_{i,j}| = \left(\frac{2R2^j}{r_j}\right)^d - \left(\frac{2R2^{j-1}}{r_j}\right)^d = (2^d - 1)\left(\frac{R2^j}{\frac{\epsilon R 2^j}{10cd}}\right)^d = (2^d - 1)\left(\frac{10cd}{\epsilon}\right)^d \quad j = 1, \ldots, M$$

$$|G_i| = (M(2^d - 1) + 2^d)\left(\frac{10cd}{\epsilon}\right)^d = O(M\epsilon^{-d}) = O(\lg(n)\,\epsilon^{-d})$$

$$|\boldsymbol{S}| = \textstyle\sum_i |\boldsymbol{S_i}| = \boldsymbol{O}\left(|\boldsymbol{A}|\,\mathbf{lg}(\boldsymbol{n})\,\boldsymbol{\epsilon}^{-\boldsymbol{d}}\right)$$

PROOF OF CORRECTNESS

Let Y be an arbitrary set of k points in $\mathbb{R}^d$. We need to show that $(1-\epsilon)\nu_Y(P) \le \nu_Y(S) \le (1+\epsilon)\nu_Y(P)$ or,

$$\mathrm{E} = |\nu_Y(P) - \nu_Y(S)| \le \epsilon \nu_Y(P)$$

For any $p \in P_i$, let p' denote the image of p in S_i, that is, the point in P_i that was chosen as the representative of all points inside the same cell of G_i as p.

Lemma 1: $\boldsymbol{d}(p, Y) \le \|pp'\| + \boldsymbol{d}(p', Y)$ *and* $\boldsymbol{d}(p', Y) \le \|pp'\| + \boldsymbol{d}(p, Y)$

Proof: Let $\alpha, \beta \in Y$ such that α is the closest point to p in Y and β be the closest point to p' in Y. So, $\boldsymbol{d}(p, Y) = \|p\alpha\|$ and $\boldsymbol{d}(p', Y) = \|p'\beta\|$. For contradiction, let $\boldsymbol{d}(p, Y) > \|pp'\| + \boldsymbol{d}(p', Y)$. Now consider the triangle formed by p, p' and β. By triangle inequality, $\|p\beta\| \le \|pp'\| + \|p'\beta\|$. Or,

$$\|p\beta\| \le \|pp'\| + \boldsymbol{d}(p', Y) < \boldsymbol{d}(p, Y) = \|p\alpha\|$$

As $\|p\beta\| < \|p\alpha\|$, and $\boldsymbol{d}(p, Y) = \min_{y \in Y} \|py\|$, hence $\|p\alpha\| \ne \boldsymbol{d}(p, Y)$.

So, using proof by contradiction, $\boldsymbol{d}(p, Y) \le \|pp'\| + \boldsymbol{d}(p', Y)$ and $\boldsymbol{d}(p', Y) \le \|pp'\| + \boldsymbol{d}(p, Y)$.

Using the above lemma, we get $|\boldsymbol{d}(p, Y) - \boldsymbol{d}(p', Y)| \le \|pp'\|$. Now,

$$\mathrm{E} = |\nu_Y(P) - \nu_Y(S)| = \textstyle\sum_{p \in P} |\boldsymbol{d}(p, Y) - \boldsymbol{d}(p', Y)| \le \sum_{p \in P} \|pp'\|$$

For all the points p that lie in $Q_{i,0}$, $\|pp'\| \le r_0\sqrt{d} = \frac{\epsilon R}{10c\sqrt{d}} \le \frac{\epsilon}{10c} R$

So, for all points p such that $\boldsymbol{d}(p, A) \le R$, $\|pp'\| \le \frac{\epsilon}{10c} R$, since all such points will lie in $Q_{i,0}$.

And for all points $p \in Q_{i,0}$ such that $\boldsymbol{d}(p, A) > R$, $\|pp'\| \leq \frac{\epsilon}{10c}\boldsymbol{d}(p, A)$

For all the points p that lie in $V_{i,j}$ $(j = 1, \ldots, M)$, $\boldsymbol{d}(p, A) \geq 2R2^{j-1}$ (as they are outside $Q_{i,j-1}$). So,

$$\|pp'\| \leq r_j\sqrt{d} = \frac{\epsilon}{10c\sqrt{d}}R2^j \leq \frac{\epsilon}{10c}\boldsymbol{d}(p, A)$$

Now,

$$\sum_{p\in P}\|pp'\| \leq \sum_i\left(\sum_{p\in P,\boldsymbol{d}(p,A)\leq R}\left(\frac{\epsilon}{10c}R\right) + \sum_{p\in P,\boldsymbol{d}(p,A)>R}\left(\frac{\epsilon}{10c}\boldsymbol{d}(p, A)\right)\right)$$

$$\leq \frac{\epsilon}{10c}nR + \frac{\epsilon}{10c}\textstyle\sum_{p\in P}\boldsymbol{d}(p, A) \leq \frac{2\epsilon}{10c}\nu_A(P) \leq \epsilon\nu_{opt}(P, k) \leq \epsilon\nu_Y(P)$$

$$|\nu_Y(P) - \nu_Y(S)| \leq \epsilon\nu_Y(P)$$

Hence, S is a (k, ϵ)-*coreset* of P. Also, the above algorithm can be easily extended for weighted point sets.

Theorem: *Given point sets P and A with n and m points, respectively, such that $\nu_A(P) \leq c\nu_{opt}(P, k)$, where c is a constant, one can compute a weighted set S which is a (k, ϵ)-coreset for P under k-median clustering, and $|S| = O(m\epsilon^{-d}\log(n))$. If P is weighted, then $|S| = O(m\epsilon^{-d}\log(W))$, where W is the total weight of P.*

Coreset for k-Means

The construction of the coreset is the same as that for k-median but for a few changes. Let P be a set of n points in $\mathbb{R}^d$, and $A = \{x_1, \ldots, x_m\}$ be a point set such that $\mu_A(P) \leq c\mu_{opt}(P, k)$. For constructing a k-means coreset, let $R = \sqrt{\frac{\mu_A(P)}{cn}}$. For any point $p \in P_i$, $\|px_i\|^2 \leq \mu_A(P) = \sqrt{cn}R$. The set S constructed using this R and the method described earlier is a (k, ϵ)-coreset of P for k-means clustering.

PROOF OF CORRECTNESS

Consider an arbitrary set Y of k points in $\mathbb{R}^d$. Let p' be the image of p in S. Using **Lemma 1**, we get-

$$|\boldsymbol{d}(p, Y) - \boldsymbol{d}(p', Y)| \leq \|pp'\|$$

And,

$$\boldsymbol{d}(p, Y) + \boldsymbol{d}(p', Y) \leq 2\boldsymbol{d}(p, Y) + \|pp'\|$$

Now, we need to show that $\mathrm{E} = |\mu_Y(P) - \mu_Y(S)| \leq \epsilon\mu_Y(P)$

$$\mathrm{E} = |\mu_Y(P) - \mu_Y(S)| \leq \textstyle\sum_{p\in P}|\boldsymbol{d}(p, Y)^2 - \boldsymbol{d}(p', Y)^2| \leq \sum_{p\in P}\left|\left(\boldsymbol{d}(p, Y) - \boldsymbol{d}(p', Y)\right)\left(\boldsymbol{d}(p, Y) + \boldsymbol{d}(p', Y)\right)\right|$$

$$\mathrm{E} \leq \textstyle\sum_{p\in P}\|pp'\|(2\boldsymbol{d}(p, Y) + \|pp'\|)$$

We divide P in three sets-

$$P_R = \{p \in P | \boldsymbol{d}(p, Y) \leq R, \boldsymbol{d}(p, A) \leq R\}, \qquad \mathrm{E}_R = \textstyle\sum_{p \in P_R} \|pp'\|(2\boldsymbol{d}(p, Y) + \|pp'\|)$$

$$P_A = \{p \in P \backslash P_R | \boldsymbol{d}(p, Y) \leq \boldsymbol{d}(p, A)\}, \qquad \mathrm{E}_A = \textstyle\sum_{p \in P_A} \|pp'\|(2\boldsymbol{d}(p, Y) + \|pp'\|)$$

$$P_Y = P \backslash (P_R \cup P_A), \qquad \mathrm{E}_Y = \textstyle\sum_{p \in P_Y} \|pp'\|(2\boldsymbol{d}(p, Y) + \|pp'\|)$$

When $\boldsymbol{d}(p, A) \leq R$, p lies in $Q_{i,0}$ by the construction. So, $\|pp'\| \leq \frac{\epsilon}{10} R$.

$$\mathrm{E}_R \leq \sum_{p \in P_R} \frac{\epsilon}{10} R \left(2R + \frac{\epsilon}{10} R\right) \leq \frac{\epsilon}{3} \sum_{p \in P_R} R^2 \leq \frac{\epsilon}{3} \mu_{opt}(P, k) \leq \frac{\epsilon}{3} \mu_B(P)$$

When $\boldsymbol{d}(p, A) > R$, $\|pp'\| \leq \frac{\epsilon}{10c} \boldsymbol{d}(p, A)$. So,

$$\mathrm{E}_A \leq \sum_{p \in P_A} \frac{\epsilon}{10c} \boldsymbol{d}(p, A) \left(2 + \frac{\epsilon}{10c}\right) \boldsymbol{d}(p, A) \leq \frac{\epsilon}{3c} \sum_{p \in P_A} \boldsymbol{d}(p, A)^2 \leq \frac{\epsilon}{3} \mu_{opt}(P, k) \leq \frac{\epsilon}{3} \mu_B(P)$$

For $p \in P_Y$, if $\boldsymbol{d}(p, A) \leq R$,

$$\|pp'\| \leq \frac{\epsilon}{10c} R \leq \frac{\epsilon}{10c} \boldsymbol{d}(p, Y)$$

Else,

$$\|pp'\| \leq \frac{\epsilon}{10c} \boldsymbol{d}(p, A) \leq \frac{\epsilon}{10c} \boldsymbol{d}(p, Y)$$

So,

$$\mathrm{E}_Y \leq \sum_{p \in P_Y} \frac{\epsilon}{10c} \boldsymbol{d}(p, Y) \left(2 + \frac{\epsilon}{10c}\right) \boldsymbol{d}(p, B) \leq \frac{\epsilon}{3} \sum_{p \in P_Y} \boldsymbol{d}(p, Y)^2 \leq \frac{\epsilon}{3} \mu_B(P)$$

$$\mathrm{E} \leq \mathrm{E}_R + \mathrm{E}_A + \mathrm{E}_Y \leq \epsilon \mu_B(P)$$

Hence, S is a (k, ϵ)-*coreset* of P. Also, the above algorithm can be easily extended for weighted point sets.

Theorem: *Given point sets P and A with n and m points, respectively, such that* $\mu_A(P) \leq c\mu_{opt}(P, k)$, *where c is a constant, one can compute a weighted set S which is a* (k, ϵ)*-coreset for P under k-means clustering, and* $|S| = O(m\epsilon^{-d} \log(n))$. *If P is weighted, then* $|S| = O(m\epsilon^{-d} \log(W))$, *where W is the total weight of P.*

COMPUTATIONAL TIME

To compute S, we need to calculate $\|px_i\|$ for all x_i for all p. This can be done naively in $O(mn)$ time. However, the authors suggest the use of a data-structure that answers constant approximate nearest neighbor queries in $O(\log m)$ per point in P after $O(m \log m)$ pre-processing. This data structure and algorithm is discussed in a paper given by S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman and A. Y. Wu in 1998. Next, we compute the exponential grids, and compute for each point of P_i its grid cell. This takes $O(1)$ time per point, if carefully implemented using hashing, log and floor functions.

Hence, the total time complexity for the algorithm becomes $O(m \log m + n \log m + n) = \boldsymbol{O(n\ log\ m)}$ in worst case, or $\boldsymbol{O(mn)}$ if implemented naively.

FAST CONSTANT FACTOR APPROXIMATION

To get the set $A = \{x_1, \dots, x_m\}$, the authors have applied algorithms previously given by Feder and Greene in 1988 and by S. Har-Peled in 2001 on the original point set P. They take the union of the resultant set with a randomly picked subset of P. The size of the resultant set is claimed to be $O(k \log^3 n)$, or $O(k \log^3 W)$ if weighted. The running time is $O(n \log(k \log n))$, or $O(n \log^2 W)$ if weighted.