NCTCS 2018

ε-Coresets for Clustering (with Outliers) in Doubling Metrics

Jian Li, Tsinghua University

Joint work with Lingxiao Huang (EPFL), Shaofeng Jiang (Weizmann), Xuan Wu (Tsinghua -> JHU)

Paper appeared in FOCS18

Big data era and Coresets

- Why Coreset?
- Turn BIG DATA to small data

Motivation

- Huge datasets
  - Store all data - expensive
  - How to analyze data efficiently
- Coreset: a small summary S of the full dataset
  - the objective computed from S approximates that computed from the full dataset.
- Benefits of coresets
  - Space: Save the storage space
  - Time: Since $|S|$ is small, computing the objective over S is much faster
  - Approximation: used for developing efficient approximation algorithm
- By now a very powerful technique to handle big data

Coreset: a powerful technique

- Shape fitting
- Clustering
- Matrix approximation
- Submodular functions
- Logistic regression
- Nonparametric learning
- Deep learning
- Multidimensional queries in database
- Extension to stochastic points
- Distributed computing (decomposable coresets)
- Deep connection to streaming/sketch/summary

Clustering

Consider a metric space $M(X,d)$ of n points

Definition ((k,z)-clustering)

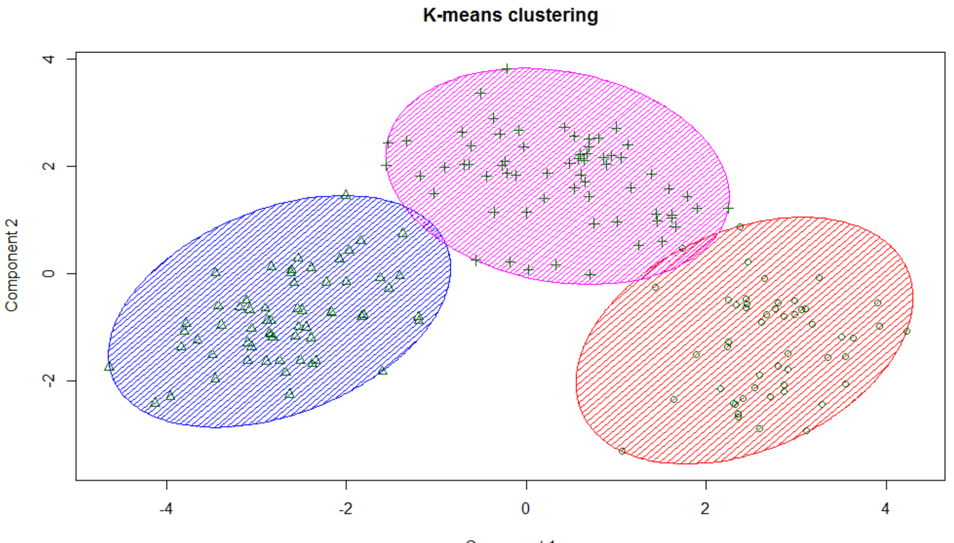
The (k,z)-clustering of M is to compute a k-subset $C \subseteq X$ such that

$$\mathcal{K}_z(X,C) \coloneqq \sum_{x\in X} d^z(x,C) = \sum_{x\in X} \min_{c\in C} d^z(x,c)$$

is minimized, where $\mathcal{K}_z$ is the clustering objective.

Special Cases

- k-median when $z = 1$
- k-means when $z = 2$
- k-center when $z = \infty$



Coreset for Clustering

Definition (ε-coreset for clustering)

A weighted subset $S \subseteq X$ with weight function $w: S \to \mathbb{R}_{\geq 0}$ is an **ε-coreset** for (k, z)-clustering of $M(X, d)$, if for any k-subset $C \subseteq X$,

$$\sum_{x \in S} w(x) \cdot d^z(x, C) \in (1 \pm \varepsilon) \cdot \mathcal{K}_z(X, C).$$

Goal: $|S|$ is independent of n for "bounded dimensional" metric spaces (depends on $k, \frac{1}{\varepsilon}, z$)

Related Work

- Euclidean space $\mathbb{R}^d$
  - An ε-coreset for (k, z)-clustering of size $\tilde{O}(dk/\varepsilon^{2z})$ can be constructed in $\tilde{O}(nk)$ time [Feldman and Langberg, 2011]
  - [Braverman et al., 2016] improved the size to $\tilde{O}(\mathrm{k}\varepsilon^{-2}\min\{\mathrm{d}, \mathrm{k}/\varepsilon\})$ for k-means
  - [Sohler and Woodruff, 2018] removed the size dependence of d for k-median (and subspace approximation)
  - For k-center ($z = \infty$), size $\mathrm{O}(k/\varepsilon^d)$ in $\mathrm{O}(n + k/\varepsilon^d)$ time [Agarwal and Procopiuc, 2002; Har-Peled, 2004]

- For general metrics, an ε-coreset for (k, z)-clustering of size $\tilde{O}(k \log n\ /\varepsilon^{2z})$ can be constructed in $\tilde{O}(nk)$ time [Feldman and Langberg, 2011]
  - In general, we can't get rid of the dimensionality $\log n$

Related Work

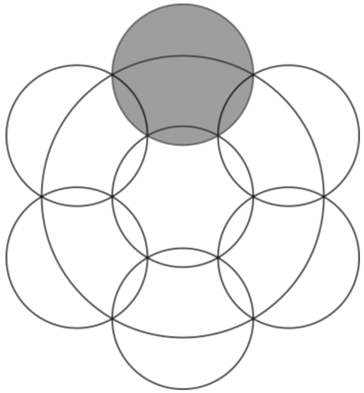
- Coreset in the streaming or distributed model (e.g., [Feldman and Langberg, 2011; Ackermann et al., 2012; Feldman and Schulman, 2012; Feldman et al., 2013; Balcan et al., 2013; Braverman et al., 2016; Braverman et al., 2017])
- Coreset for stochastic data
  - Stochastic minimum enclosing ball (1-center) [Munteanu et al., 2014]
  - Stochastic k-center [Huang and Li, 2017]

Coreset for Clustering in Doubling Metrics

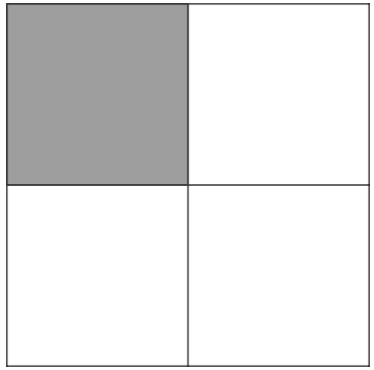
Doubling Dimension

Definition (doubling dimension)

The doubling dimension of $M(X, d)$, denoted as $ddim(M)$, is the smallest integer t such that any ball can be covered by at most 2^t balls of half the radius.



$(a)\ ddim(l_2^2) = 3$



$(b)\ ddim(l_\infty^2) = 2$

- In general, metric l_p in $\mathbb{R}^d$ has doubling dimension $O(d)$ [Assouad, 1983]

Why Doubling Metrics?

- Doubling metrics extensively studied
  - Spanners [Cole and Gottlieb, 2006; Chan et al., 2016; etc]
  - Metric embedding [Gupta et al., 2003; Abraham et al., 2006; Chan et al., 2010]
  - Nearest neighbor search [Clarkson 1999; Har-Peled and Mendel, 2005; etc]
  - Approximation algorithms [Chan and Elbassioni, 2011; Friggstad et al., 2016]
  - Machine learning [Bshouty et al., 2009; Gottlieb et al., 2014]
- Force us to forget about the coordinate and think about the metric space per se
- Some metric data lives in high dimensional Euclidean space, but may inherently have low doubling dimension $ddim$
- Other examples: Earthmover distance (EMD), Edit distance with real penalty (ERP) [Gottlieb et al., 2014], machine learning classifiers [Bshouty et al., 2009]
- Natural attempt: embed doubling metrics to Euclidean spaces
  - There exists $O(1)$-distortion embedding to l_2 which leads to $O(1)$-coreset. However, there is also $\Omega(1)$-distortion lower bound [Gupta eta al., 2013].
  - Constant size ε-coreset?

Our Result

Main Theorem (informal)

Given a metric space $M(X, d)$ of n points, there exists a poly-time algorithm that constructs an ε-coreset of size

$$\mathrm{poly}(k, ddim(M), 1/\varepsilon)$$

for the (k, z)-clustering problem, with probability at least 0.99.

High-Level Sketch of Our Technique

Main Approach: Importance Sampling

Importance Sampling Framework [Langberg and Schulman, 2010; Feldman and Langberg, 2011]

- Sensitivity: $\sigma(x) := \max_{C \subseteq X : |C| = k} \frac{d^z(x,C)}{\mathcal{K}_z(X,C)}$
  - Sensitivity measures the "importance" of each point
- Approx. compute sensitivities of all points
- Sample points from a distribution proportional to sensitivities $\sigma(x)$, each sample has a weight $1/\sigma(x)$ for unbiased estimation.

Importance Sampling -> Coreset

Theorem [Feldman and Langberg, 2011]

An ε-coreset can be constructed in poly-time with size

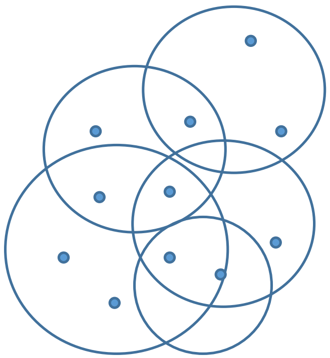
$$\left(\frac{\sigma}{\epsilon}\right)^2 (dim + \log 1/\delta)$$

$\sigma = \sum_x \sigma(x) = O(2^{2z} k)$ [Varadarajan and Xiao, 2012]

Definition (shattering dimension)

For $x \in X, r \geq 0$, define ball $B(x,r) \coloneqq \{y \in X: d(x,y) \leq r\}$. The shattering dimension $\mathbf{dim}(\boldsymbol{M})$ is the least integer t such that for any $A \subseteq X$ of size ≥ 2, the number of different subsets of A intersected by balls

$$|\{A \cap B(x,r): x \in X, r \geq 0\}| \leq |A|^t$$



Shattering dimension plays a similar role as VC dimension:

$$\dim(M) \leq VC\text{-}dim(M) \leq \dim(M) \log \dim(M)$$

Doubling Dimension $v.s.$ Shattering Dimension

Does $ddim(M) = O(1)$ imply $dim(M) = O(1)$?

- If M is Euclidean, then $ddim(M)$=$O(1)$ implies $dim(M)$=$O(1)$.
- How about general metric spaces?

The answer is unfortunately **NO**.

- Example: $ddim(M) = O(1)$ but $dim(M) = \Omega(\frac{\log n}{\log \log n})$
  - Point set: $M = \{u_1, \ldots, u_m, v_0, \ldots, v_{2^m-1}\}$ where $m \approx \log n$
  - Distance: $d(u_i, u_j) = |i-j|$; $d(v_i, v_j) = |i-j|$; $d(u_i, v_j) = 2^m$ if digit i of j's binary representation is 0 and otherwise $d(u_i, v_j) = 2^m$+1
- Difficulty: how to relate $ddim(M)$ and $dim(M)$?

Main Idea: Distortion

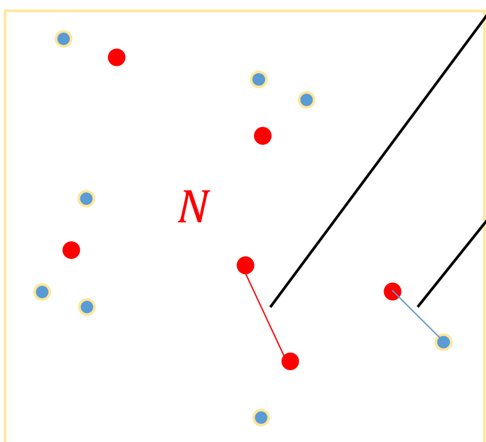
- We want to "distort" the distance $d(\cdot,\cdot)$ such that the shattering dimension is bounded by the doubling dimension:
  - Low distortion: for any $x, y \in X$, $\delta(x, y) \in (1 \pm \varepsilon) \cdot d(x, y)$
  - Objective: For the "smoothed metric space" $M(X, \delta)$, we have

$$\dim(M(X, \delta)) \leq f(ddim(M), \frac{1}{\varepsilon}).$$

- Next step: construct coresets via $M(X, \delta)$.
  - Since δ is a low distortion, an ε-coreset of $M(X, \delta)$ is a 3ε-coreset of $M(X, d)$.

- Problem: how to construct such a distorted distance δ?

Notations: Packing, Covering and Net



$M(X, d)$

N is a ρ-packing, if $\forall u, v \in N, d(u, v) \geq \rho$.

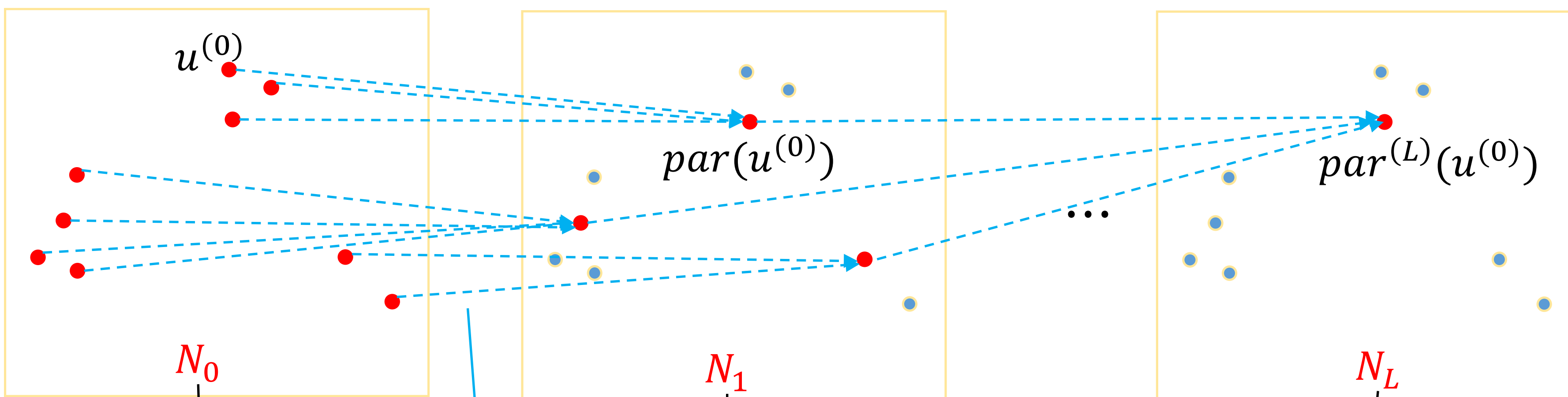
Packing property: $|\mathrm{N}| \leq \left(\frac{2Diam(N)}{\rho}\right)^{ddim(M)}$.

N is a ρ-covering, if $\forall x \in X$, there exists $u \in N$ such that $d(u, x) \leq \rho$.

N is a ρ-net, if N is both a ρ-packing and a ρ-covering.

Notations: Hierarchical Net and Net Tree

Scale the metric such that the minimum intra point distance is 1



$\{N_0, N_1, \ldots, N_L\}$ is a *hierarchical net*, where N_i is a 2^i-net of N_{i-1}.

- Useful concept in doubling metrics [Talwar 2004; etc]

Net tree: node set $\bigcup_i N_i$. The parent $par(u^{(i)})$ of $u^{(i)} \in N_i$ is its nearest point in N_{i+1}.

- $par^{(j)}(u)$: the ancestor of u in N_j

Distortion: Smoothed Distance Function

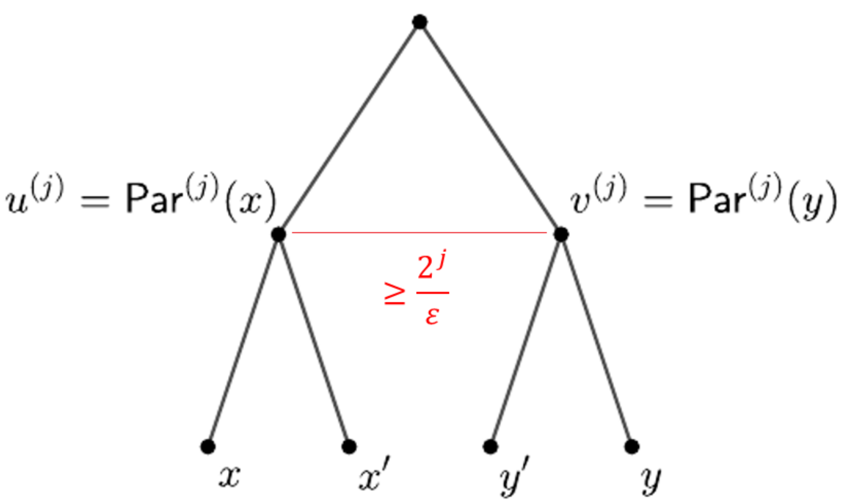
Definition (smoothed distance function)

Given a net tree T, for $x, y \in X$, let j be the largest integer such that

$$d\left(par^{(j)}(x), par^{(j)}(y)\right) \geq \frac{2^j}{\varepsilon}.$$

The ε-smoothed distance function is defined by

$$\delta(x,y) := d\left(par^{(j)}(x), par^{(j)}(y)\right)$$



Lemma

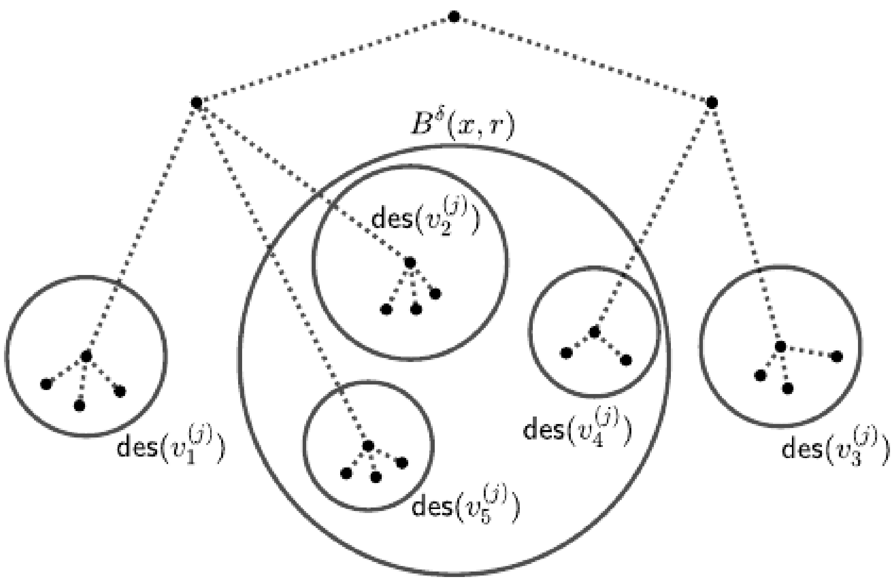
$\forall x, y \in X$,

$d(x,y) \in (1 \pm 4\varepsilon)\delta(x,y)$.

Smooth Property: Cross-Free

Lemma (cross-free property)

Consider $0 < \varepsilon \le \frac{1}{8}$ and an integer j. Suppose $r \ge 100 \cdot \frac{2^j}{\varepsilon}$. Then for any $x \in X$ and $v^{(j)} \in N_j$, either none or all descendants of $v^{(j)}$ are contained in $B^\delta(x, r)$.



Smooth Property → Bounded Shattering Dimension

Idea: Fix $A \subseteq X$.

- Cross-free → $A \cap B^\delta(x, r)$ is a disjoint union of $A \cap des(v_i^{(j)})$
- Packing property → there are at most $O\left(\frac{r}{2^j}\right)^{O(ddim(M))}$ such $v_i^{(j)}$
- $\dim(M(X, \delta)) \le \varepsilon^{-O(ddim(M))}$

Weakness

- We have constructed a smoothed distance function δ such that
  - For any $x, y \in X$, $\delta(x, y) \in (1 \pm \varepsilon) \cdot d(x, y)$
  - $\dim(M(X, \delta)) \leq \varepsilon^{-O(ddim(M))}$

- Weakness
  - The exponential dependence on $ddim(M)$
  - Only works for an unweighted ground set X. However for coresets, we need to relate $ddim(M)$ and $dim(M)$ for weighted point sets.

An Improved Framework

Definition (probabilistic shattering dimension, informal)

Let $M(X, \delta)$ be a metric space where δ is a randomized distortion function.

The probabilistic shattering dimension $\mathrm{pdim}_\tau(M)$ is the least integer t such that for any $A \subseteq X$ of size ≥ 2, the number of different subsets of A intersected by balls

$$\left|\{A \cap B^\delta(x, r) : x \in X, r \geq 0\}\right| \leq |A|^t,$$

with probability at least $1 - \tau$.

An Improved Framework

Exponential Improvement via Randomness

Introducing randomness in the distortion δ,

$$\mathrm{pdim}_\tau(M) \leq O(ddim(M) \cdot \log\frac{1}{\varepsilon} + \log\log\frac{1}{\tau})$$

where $pdim(M)$ is probabilistic shattering dimension.

- Proof more involved.
- The randomized δ is constructed based on a randomized hierarchical decomposition [Abraham et al., 2006].

New framework: bounded $pdim(M)$ + importance sampling -> coreset

Application: Centroid Set

Definition (centroid set)

Given an ε-coreset $S \subseteq X$ with weights $\mathrm{w}(\mathrm{x})$, an (ε, k, z)-centroid set of (S, w) is a subset H such that

- $S \subseteq H \subseteq X$
- There exists a k-point set $C \subseteq H$ such that

$$\sum_{x \in S} w(x) \cdot d^z(x, C) \leq (1 + 2\varepsilon) \cdot \min_{C' \subseteq H : |C'| = k} \mathcal{K}_z\,(X, C').$$

Theorem (centroid set)

Given $S \subseteq X$ with weights $\mathrm{w}(\mathrm{x})$, there exists a poly-time algorithm that constructs an $(O(z, \varepsilon), k, z)$-centroid set of size

$$O(\varepsilon)^{-O(ddim(M))} \cdot |S|^2.$$

Application: Fast Local Search Algorithm

Local Search Yields a PTAS

As analyzed in [Friggstad et al., 2016; Cohen-Addad et al., 2016], the local search algorithm that swaps at most $\rho(\varepsilon, ddim(M), z)$ centers at each iteration satisfies

- $(1+\varepsilon)$-approx. for (k, z)-clustering
- Per-iteration running time: n^{ρ}
- The number of iterations is polynomial in the input size

Accelerating via Centroid Set

- As noted in [Friggstad et al., 2016], applying the centroid set for Euclidean spaces yields per-iteration running time $(k/\varepsilon)^{O(\rho)}$.
- Our results of coreset and centroid set can achieve a similar bound for doubling metrics.

Conclusion Remark

- (γ, ε)-robust coreset: allow outliers
  - Size $\tilde{O}(kd\gamma^{-2}\varepsilon^{-4})$ [Feldman and Langberg, 2011]
  - Improve to $\tilde{O}(kd\gamma^{-2}\varepsilon^{-2})$ [this paper]
- Is the probabilistic notion of dimension $pdim(M)$ necessary, i.e., does there exist a deterministic distortion δ such that $\dim(M, \delta) \approx O(ddim(M))$?
- We give the first coreset construction for doubling metrics
  - $\tilde{O}(dk/\varepsilon^{2z})$ size in Euclidean spaces [Feldman and Langberg, 2011]
  - Can we improve our coreset size to match the Euclidean bound?
- Multidimensional queries in database (C**L**RWWZ ICDT 17)
- Extension to stochastic points (H**L**PW ESA16, H**L** SODA 17)

Thank you! Questions?

Jian Li
lapordge@gmail.com

wechat: lapordge