Probabilistic k-Median Clustering in Data Streams

WAOA 2012

Christiane Lammersen, Melanie Schmidt, Christian Sohler

13.09.2012

Clustering

$\rightarrow$ Partition a set of given objects into subsets of similar objects
$\rightarrow$ Similarity or Dissimilarity is measured by a distance function

Clustering

$\rightarrow$ Partition a set of given objects into subsets of similar objects
$\rightarrow$ Similarity or Dissimilarity is measured by a distance function

Metric k-median clustering

Clustering

$\rightarrow$ Partition a set of given objects into subsets of similar objects
$\rightarrow$ Similarity or Dissimilarity is measured by a distance function

Metric k-median clustering

Given a set of points P from a metric space $M = (X, D)$, find

- a set $C := \{c_1, \ldots, c_k\} \subseteq X$ minimizing

Clustering

$\rightarrow$ Partition a set of given objects into subsets of similar objects
$\rightarrow$ Similarity or Dissimilarity is measured by a distance function

Metric k-median clustering

Given a set of points P from a metric space $M = (X, D)$, find

- a set $C := \{c_1, \ldots, c_k\} \subseteq X$ minimizing

$$\text{cost}(P, C) := \sum_{i=1}^{n} \min_{c \in C} \mathrm{D}(p_i, c).$$

Metric k-median clustering

Given a set of points P from a metric space $M = (X, D)$, find

- a set $C := \{c_1, \ldots, c_k\} \subseteq X$ minimizing

$$\mathrm{cost}(P, C) := \sum_{i=1}^{n} \min_{c \in C} \mathrm{D}(p_i, c).$$

Metric k-median clustering

Given a set of points P from a metric space $M = (X, D)$, find

- a set $C := \{c_1, \ldots, c_k\} \subseteq X$ minimizing

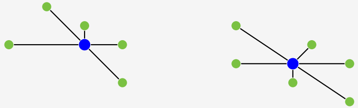
$$\operatorname{cost}(P, C) := \sum_{i=1}^{n} \min_{c \in C} \mathrm{D}(p_i, c).$$

Metric k-median clustering

Given a set of points P from a metric space $M = (X, D)$, find

- a set $C := \{c_1, \ldots, c_k\} \subseteq X$ minimizing

$$\text{cost}(P, C) := \sum_{i=1}^{n} \min_{c \in C} \mathrm{D}(p_i, c).$$



Metric k-median clustering

Given a set of points P from a metric space $M = (X, D)$, find

- a set $C := \{c_1, \ldots, c_k\} \subseteq X$ minimizing

$$\mathrm{cost}(P, C) := \sum_{i=1}^{n} \min_{c \in C} \mathrm{D}(p_i, c).$$

Probabilistic Data

- Sensor data
- Database joins
- Movement data

Probabilistic Data

- Sensor data
- Database joins
- Movement data

Probabilistic points

For us, a probabilistic point is a discrete probability distribution

The probabilistic k-median problem

Given

The probabilistic k-median problem

Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,

The probabilistic k-median problem

Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$

The probabilistic k-median problem

Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node v_i, given by realization probabilities p_{ij} for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,

The probabilistic k-median problem

Given

- finite set $\mathcal{X} := \{x_1, \ldots, x_m\}$ from metric space $M = (X, D)$,
- set of nodes $V : \{v_1, \ldots, v_n\}$
- probability distribution $\mathcal{D}_i$ for each node v_i, given by realization probabilities p_{ij} for all $j \in [m]$, $\sum_{j=1}^{m} p_{ij} \leq 1$,

find a set $C := \{c_1, \ldots, c_k\} \subseteq X$ that minimizes

$$\mathbf{E}_{\mathcal{D}}\left[\text{cost}(V, C)\right] := \min_{\rho : V \to C} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \cdot \mathrm{D}(x_j, \rho(v_i)).$$

Related work: Clustering probabilistic Data

Cormode, McGregor (PODS 2008)

- $(1+\varepsilon)$-approximation for a variant of the above problem
- $(1+\varepsilon)$-approximation for uncertain *k*-means
- Constant approximation for (assigned) metric *k*-median
- Bicriteria approximations for uncertain metric *k*-center

Guha and Munagala (PODS 2009)

- Constant approximation for uncertain metric *k*-center

Data Streams

- large amounts of data
- data arrives in a stream
- only one pass over the data allowed
- limited storage capacity

Data Streams

- large amounts of data
- data arrives in a stream
- only one pass over the data allowed
- limited storage capacity

One way to deal with data streams: Coresets

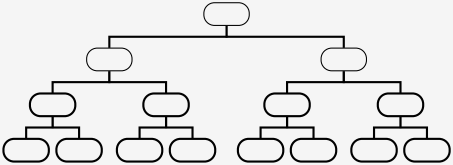
Coresets

- small summary of given data
- typically of constant or polylogarithmic size
- can be used to approximate the cost of the original data

Coresets

- small summary of given data
- typically of constant or polylogarithmic size
- can be used to approximate the cost of the original data

Merge & Reduce



- read data in blocks
- compute a coreset for each block $\to s$
- merge coresets in a tree fashion
- $\rightsquigarrow$ space $s \cdot \log n$

Related work: Coreset constructions

- '01: Agarwal, Har-Peled and Varadarajan: Coreset concept
- '02: Bădoiu, Har-Peled and Indyk: First coreset construction for clustering problems
- '04: Har-Peled and Mazumdar, Coreset of size $\mathcal{O}(k\varepsilon^{-d} \log n)$ for Euclidean k-median, maintainable in data streams
- '05: Har-Peled, Kushal: Coreset of size $\mathcal{O}(k^2\varepsilon^{-d})$ for Euclidean k-median
- '05: Frahling and Sohler: Coreset of size $\mathcal{O}(k\varepsilon^{-d} \log n)$ for Euclidean k-median, insertion-deletion data streams
- '06: Chen: Coresets for metric and Euclidean k-median and k-means, polynomial in d, $\log n$ and ε^{-1}
- '10: Langberg, Schulman: $\tilde{\mathcal{O}}(d^2k^3/\varepsilon^2)$
- '11: Feldman, Langberg: $\mathcal{O}(dk/\varepsilon^2)$

Our goal

Compute a coreset for the probabilistic k-median problem

Our goal

Compute a coreset for the probabilistic k-median problem

Coresets

Given a set of probabilistic points V, a weighted subset U is a (k,ε)-coreset if for all sets C of k centers it holds

$$|\mathbf{E}_{\mathcal{D}'}[\mathrm{cost}_w(U,C)] - \mathbf{E}_{\mathcal{D}}[\mathrm{cost}(V,C)]| \leq \varepsilon \mathbf{E}_{\mathcal{D}}[\mathrm{cost}(V,C)]$$

where $\mathbf{E}_{\mathcal{D}'}[\mathrm{cost}_w(U,C)] := \min_{\rho: U\to C} \sum_{v_i \in U} \sum_{j=1}^{m} p'_{ij} w(v_i) \mathrm{D}(x_j, \rho(v_i))$.

Our goal

Compute a coreset for the probabilistic k-median problem

Coresets

Given a set of probabilistic points V, a weighted subset U is a (k, ε)-coreset if for all sets C of k centers it holds

$$|\mathbf{E}_{\mathcal{D}'}[\mathrm{cost}_w(U, C)] - \mathbf{E}_{\mathcal{D}}[\mathrm{cost}(V, C)]| \leq \varepsilon \mathbf{E}_{\mathcal{D}}[\mathrm{cost}(V, C)]$$

where $\mathbf{E}_{\mathcal{D}'}[\mathrm{cost}_w(U, C)] := \min_{\rho: U \to C} \sum_{v_i \in U} \sum_{j=1}^{m} p'_{ij} w(v_i) \mathrm{D}(x_j, \rho(v_i)).$

$|U|$ and support of probability distributions should be small

Metric k-median

Idea

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?

- Expected distance?

Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?

- Expected distance?
- Expected distance between two copies of the same probabilistic node is not zero

Metric k-median

Idea

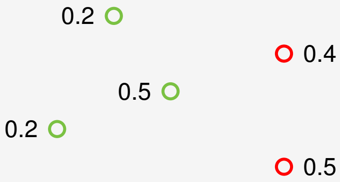
- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?

- Expected distance?
- Expected distance between two copies of the same probabilistic node is not zero
- $\rightsquigarrow$ expected distance is not a metric

Metric k-median

Idea

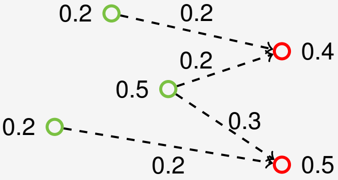
- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?



Metric k-median

Idea

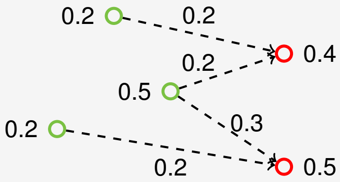
- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?



Metric k-median

Idea

- Extend cost function to a metric
- (so far only defined for a tuple of a node and a center)
- Point $c \in X \rightsquigarrow$ node with all probability at c
- Generalization of cost function to distance between nodes?



$\rightarrow$ Earth Mover Distance (EMD)

- EMD is a metric

- EMD is a metric
- EMD is a generalization of the cost function

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset

– if we thin out the probability distributions and

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset
  - if we thin out the probability distributions and
  - handle non-uniform realization probabilities.

- EMD is a metric
- EMD is a generalization of the cost function
- for each $x \in \mathcal{C}$, create an artificial node $\rightsquigarrow \mathcal{C}'$
- A deterministic (k, ε)-coreset for V with center set $\mathcal{C}'$ and metric EMD is a probabilistic (k, ε)-coreset
  - if we thin out the probability distributions and
  - handle non-uniform realization probabilities.

- (Compute EMD efficiently!)

Partitioning nodes

Does the same approach work in the Euclidean case?

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)
- in the Euclidean case, one usually sets $\mathcal{C} = \mathbb{R}^d$.

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)
- in the Euclidean case, one usually sets $\mathcal{C} = \mathbb{R}^d$.

$\rightsquigarrow$ algorithms for the general case do not work here

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)
- in the Euclidean case, one usually sets $\mathcal{C} = \mathbb{R}^d$.

$\rightsquigarrow$ algorithms for the general case do not work here

$\rightsquigarrow$ even though probabilistic Euclidean k-median can be seen as deterministic metric k-median, we cannot use deterministic algorithms

Does the same approach work in the Euclidean case?

- in the general metric case, $\mathcal{C}$ is usually finite (e.g. P)
- in the Euclidean case, one usually sets $\mathcal{C} = \mathbb{R}^d$.

$\rightsquigarrow$ algorithms for the general case do not work here

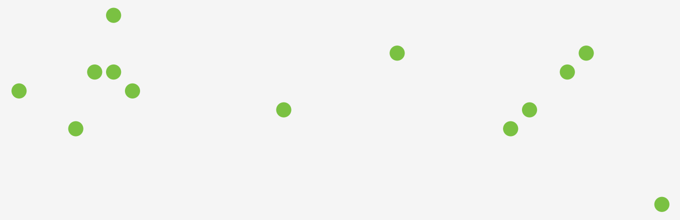
$\rightsquigarrow$ even though probabilistic Euclidean k-median can be seen as deterministic metric k-median, we cannot use deterministic algorithms

$\rightsquigarrow$ Develop coreset construction

$\rightsquigarrow$ Use deterministic coreset construction by Chen

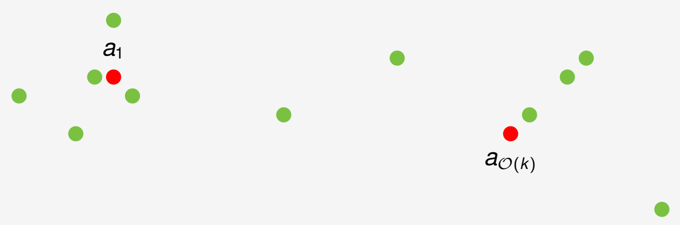
Partitioning nodes

Chen (2006)



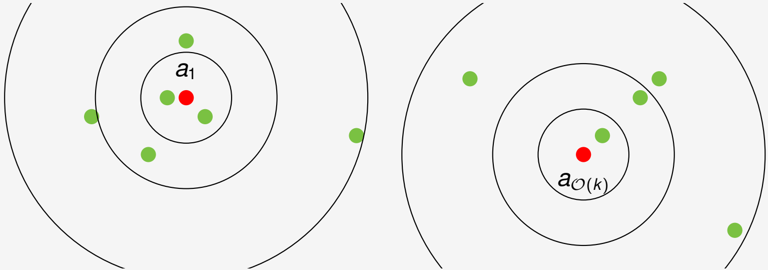
Chen (2006)

- compute bicriteria approximation



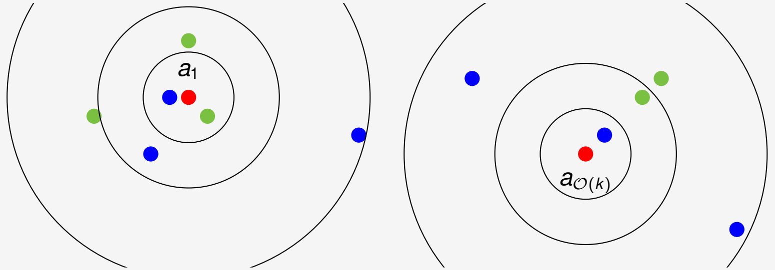
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost



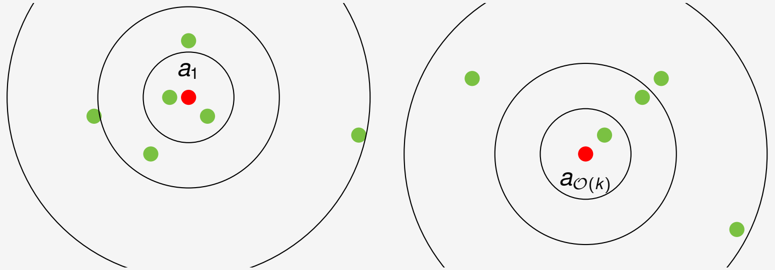
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



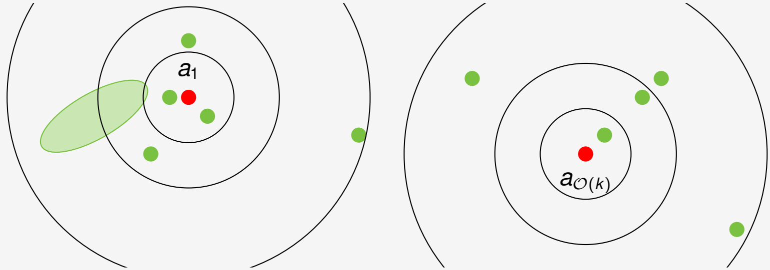
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



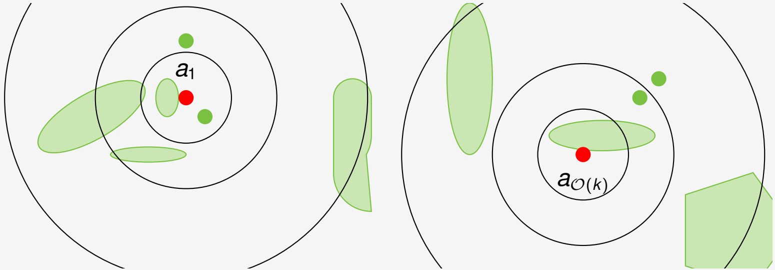
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



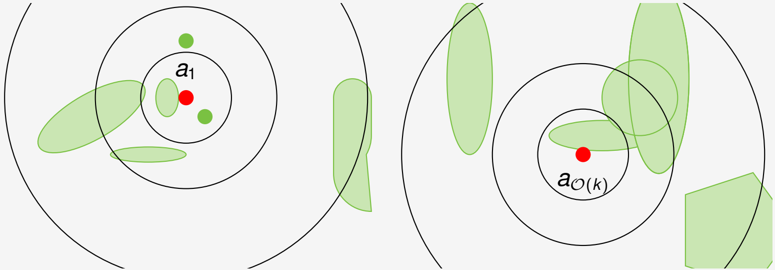
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



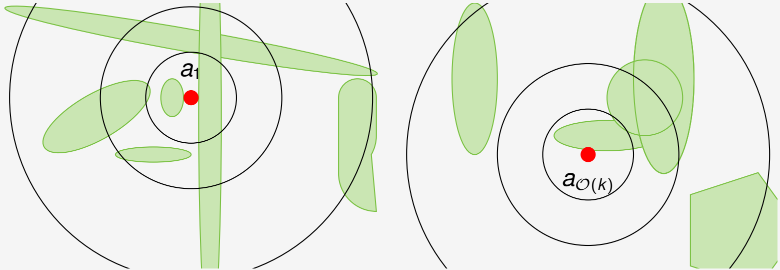
Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



Chen (2006)

- compute bicriteria approximation
- partition input points into subsets of points which are close to each other compared to the optimal clustering cost
- sample representatives from each subset



Theorem

We can compute a probabilistic (k, ε)-coreset of size

$$\mathcal{O}(k^2\varepsilon^{-3} \cdot \text{polylog}(|\mathcal{C}|, n, \delta, 1/p_{\min}))$$

for the probabilistic metric k-median problem and of size

$$\mathcal{O}(k^2\varepsilon^{-2}d \cdot \text{polylog}(n, \delta, \varepsilon^{-1}, 1/p_{\min}))$$

for the probabilistic Euclidean k-median problem.

Theorem

We can compute a probabilistic (k, ε)-coreset of size

$$\mathcal{O}(k^2 \varepsilon^{-3} \cdot \text{polylog}(|\mathcal{C}|, n, \delta, 1/p_{\text{min}}))$$

for the probabilistic metric k-median problem and of size

$$\mathcal{O}(k^2 \varepsilon^{-2} d \cdot \text{polylog}(n, \delta, \varepsilon^{-1}, 1/p_{\text{min}}))$$

for the probabilistic Euclidean k-median problem.

Thank you for your attention!