Clustering under Perturbation Resilience

Maria Florina Balcan, Yingyu Liang
Georgia Institute of Technology

October 4, 2012

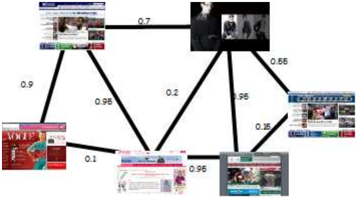
Clustering Comes Up Everywhere

- Cluster news articles or web pages by topic

- Cluster images by who is in them

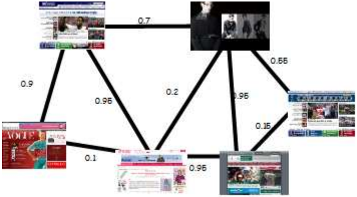
Standard Theoretical Approach

- View objects as nodes in weighted graph based on the distances



Standard Theoretical Approach

- View objects as nodes in weighted graph based on the distances



- Pick some objective to optimize
  - k-median: find centers $\{c_1, \ldots, c_k\}$ to minimize $\sum_i \sum_{p \in C_i} d(p, c_i)$
  - Min-sum: find partition $\{C_1, \ldots, C_k\}$ to minimize $\sum_i \sum_{p,q \in C_i} d(p, q)$

Standard Theoretical Approach

- Pick some objective to optimize
  - k-median: find centers $\{c_1, \ldots, c_k\}$ to minimize $\sum_i \sum_{p \in C_i} d(p, c_i)$
  - Min-sum: find partition $\{C_1, \ldots, C_k\}$ to minimize $\sum_i \sum_{p,q \in C_i} d(p, q)$
- k-median: NP-hard to approximate within a factor of $(1 + 1/e)$; can be approximated within a $(3 + \epsilon)$ factor
- Min-sum: NP-hard to optimize; can be approximated within a $\log n$ factor

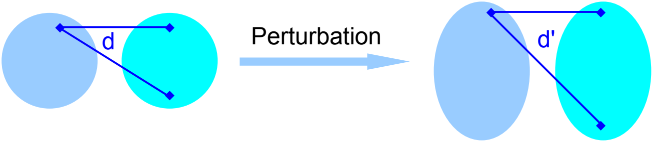
Standard Theoretical Approach

- Pick some objective to optimize
  - k-median: find centers $\{c_1, \ldots, c_k\}$ to minimize $\sum_i \sum_{p \in C_i} d(p, c_i)$
  - Min-sum: find partition $\{C_1, \ldots, C_k\}$ to minimize $\sum_i \sum_{p,q \in C_i} d(p, q)$
- k-median: NP-hard to approximate within a factor of $(1 + 1/e)$; can be approximated within a $(3 + \epsilon)$ factor
- Min-sum: NP-hard to optimize; can be approximated within a $\log n$ factor
- Cool new direction: exploit additional properties of the data to circumvent lower bounds

α-Perturbation Resilience

α-PR [Bilu and Linial, 2010, Awasthi et al., 2012]

A clustering instance (S, d) is α-perturbation resilient to a given objective function Φ if for any function $d' : S \times S \to R_{\geq 0}$ s.t. $\forall p, q \in S, d(p, q) \leq d'(p, q) \leq \alpha d(p, q)$, there is a unique optimal clustering $\mathcal{OPT}'$ for Φ under d' and this clustering is equal to the optimal clustering $\mathcal{OPT}$ for Φ under d.



Main Results

- Polynomial time algorithm for finding $\mathcal{OPT}$ for α-PR k-median instances when $\alpha \geq 1 + \sqrt{2}$
  - It works for any center-based objective function, e.g. k-means
- Polynomial time algorithm for a generalization (α, ϵ)-PR
- Polynomial time algorithm for finding $\mathcal{OPT}$ for α-PR min-sum instances when $\alpha \geq 3\frac{\max_i |C_i|}{\min_i |C_i| - 1}$

Structure Properties of α-PR k-Median Instance

Claim

α-PR for k-median implies that $\forall p \in C_i, \alpha d(p, c_i) < d(p, c_j)$.

- Blow up all the pairwise distances within the optimal clusters by α
- The $\mathcal{OPT}$ does not change, so $\forall p \in C_i, d'(p, c_i) < d'(p, c_j)$
- $d'(p, c_i) = \alpha d(p, c_i) < d'(p, c_j) = d(p, c_j)$

Structure Properties of α-PR k-Median Instance

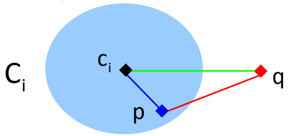
Claim

α-PR for k-median implies that $\forall p \in C_i, \alpha d(p, c_i) < d(p, c_j)$.

- Blow up all the pairwise distances within the optimal clusters by α
- The $\mathcal{OPT}$ does not change, so $\forall p \in C_i, d'(p, c_i) < d'(p, c_j)$
- $d'(p, c_i) = \alpha d(p, c_i) < d'(p, c_j) = d(p, c_j)$

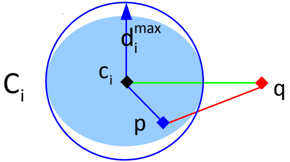
Implication:

- if $\alpha \geq 1 + \sqrt{2}, \forall p \in C_i, q \notin C_i$,
  $d(c_i, p) < d(c_i, q)$ and $d(c_i, p) < d(p, q)$



Structure Properties of α-PR k-Median Instance

- Let $d_i^{max} = \max_{p \in C_i} d(p, c_i)$. Construct a ball $B(c_i, d_i^{max})$
  - the ball covers exactly C_i
  - points inside are closer to the center than to points outside, i.e. $\forall p \in B(c_i, d_i^{max}), q \notin B(c_i, d_i^{max}), d(p, c_i) < d(p, q)$

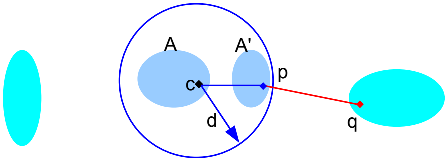


Closure Distance

Closure Distance

The closure distance $d_S(A, A')$ between two subsets A and A' is the minimum d, such that there exists a point $c \in A \cup A'$ satisfying:

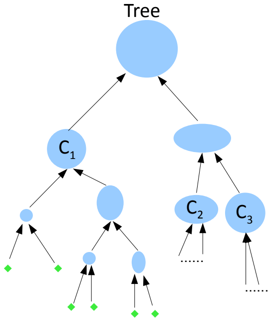
- **coverage condition**: the ball $B(c, d)$ covers $A \cup A'$;
- **margin condition**: points inside are closer to the center than to points outside, i.e. $\forall p \in B(c, d), q \notin B(c, d), d(c, p) < d(p, q)$.



Algorithm for α-PR k-median

Closure Linkage

- Begin with each point being a cluster
- Repeat until one cluster remains:
  merge the two clusters with
  minimum closure distance
- Output the tree with points as leaves and merges as internal nodes



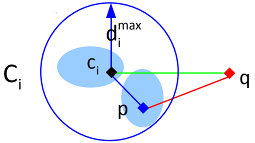
Theorem

If $\alpha \geq 1 + \sqrt{2}$, the tree output contains $\mathcal{OPT}$ as a pruning.

Proof

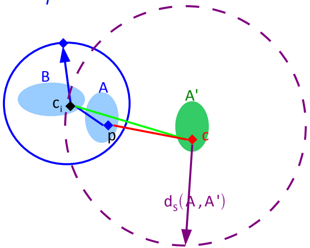
By induction, we show that the algorithm will not merge a strict subset $A \subset C_i$ with a subset A' outside C_i.

- Pick $B \subset C_i \setminus A$ such that $c_i \in A \cup B$
- $d_S(A, B) \leq d_i^{max} = \max_{p \in C_i} d(p, c_i)$
  - d_i^{max} and $c_i \in A \cup B$ satisfy the two conditions of closure distance



Proof

- $d_S(A, A') > d_i^{max}$
  - Suppose the center c for the ball defining $d_S(A, A')$ is from A'
  - Since $c \notin C_i$, $d(c_i, p) < d(p, c)$ for arbitrary $p \in A$.
    By margin condition, $c_i \in B(c, d_S(A, A'))$, $i.e.$ $d_S(A, A') \geq d(c_i, c)$
  - Since $c \notin C_i$, $d(c_i, c) > d_i^{max}$



  - A similar argument holds for the case $c \in A$

(α, ϵ)-Perturbation Resilience

- α-PR imposes a strong restriction that the $\mathcal{OPT}$ does not change after perturbation
- We propose a more realistic relaxation

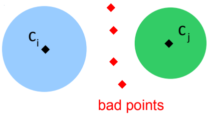
(α, ϵ)-Perturbation Resilience

A clustering instance (S, d) is (α, ϵ)-perturbation resilient to a given objective function Φ if for any function $d' : S \times S \to R_{\geq 0}$ s.t. $\forall p, q \in S, d(p, q) \leq d'(p, q) \leq \alpha d(p, q)$, the optimal clustering $\mathcal{OPT}'$ for Φ under d' is ϵ-close to the optimal clustering $\mathcal{OPT}$ for Φ under d.

Structure Property of (α, ϵ)-PR k-median instance

Theorem

Assume $\min_i |C_i| > c\epsilon n$. Except for at most ϵn bad points, any other point is α times closer to its own center than to other centers.



Keypoint of the Proof

- Carefully construct a perturbation that forces all the bad points move
- By (α, ϵ)-PR, there could be at most ϵn bad points

Algorithm for (α, ϵ)-PR k-median instance

A robust version of Closure Linkage algorithm can be used to show:

Theorem

Assume $\min_i |C_i| \geq c\epsilon n$. If $\alpha \geq 2 + \sqrt{7}$, then the tree output contains a pruning that is ϵ-close to the optimal clustering. Moreover, the cost of this pruning is $(1 + O(\epsilon/\rho))$-approximation where $\rho = \min_i |C_i|/n$.

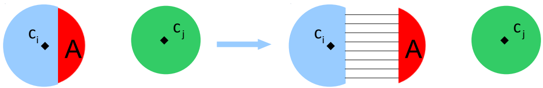
α-PR Min-Sum Instance

- Connect each point with its $\min_i |C_i|/2$ nearest neighbors
- Perform average linkage on the components

Theorem

If $\alpha \geq 3\frac{\max_i |C_i|}{\min_i |C_i| - 1}$, then the tree output contains $\mathcal{OPT}$ as a pruning.

- α-PR implies $\forall A \subseteq C_i, \alpha d(A, C_i \setminus A) < d(A, C_j)$
  - Consider blowing up the distances between A and $C_i \setminus A$ by α



α-PR Min-Sum Instance

- Connect each point with its $\min_i |C_i|/2$ nearest neighbors
- Perform average linkage on the components

Theorem

If $\alpha \geq 3\frac{\max_i |C_i|}{\min_i |C_i|-1}$, then the tree output contains $\mathcal{OPT}$ as a pruning.

- α-PR implies $\forall A \subseteq C_i, \alpha d(A, C_i \setminus A) < d(A, C_j)$
- The property guarantees
  - the components are pure
  - no strict subset of an optimal cluster will be merged with a subset outside the cluster

Conclusion

- Polynomial time algorithm for finding (nearly) optimal solutions for perturbation resilient instances.
- Also consider a more realistic relaxation (α, ϵ)-PR

Open Questions

- Design alg for (α, ϵ)-PR min-sum

Thanks!

Awasthi, P., Blum, A., and Sheffet, O. (2012).
Center-based clustering under perturbation stability.
Inf. Process. Lett., 112(1-2):49–54.

Bilu, Y. and Linial, N. (2010).
Are stable instances easy?
In *Innovations in Computer Science*.

Proof of Property of (α, ϵ)-PR: the perturbation

- For technical reasons, for each i select $\min(|B_i|, \epsilon n + 1)$ bad points from B_i
- Blow up all pairwise distances by α, except
  - between the bad points and their second nearest centers
  - between the other points and their own centers
- Intuition: ideally, after the perturbation, all bad points are assigned to their second nearest center, all the other points stay

Proof of Property of (α, ϵ)-PR: centers after perturbation

Let c_i' be the new center for the new i-th cluster C_i'.
Sufficient to show: $c_i' \neq c_i$ leads to a contradiction.

- C_i' differs from C_i on at most ϵn points
- c_i' is close to c_i
- $d(c_i', C_i' \cap C_i) \approx d(c_i, C_i' \cap C_i)$
- $d'(c_i', C_i' \cap C_i) = \alpha d(c_i', C_i' \cap C_i) \gg d'(c_i, C_i' \cap C_i) = d(c_i, C_i' \cap C_i)$
- $d'(c_i', C_i') > d'(c_i, C_i')$, a contradiction

Structure Property of α-PR Min-Sum Instance

Claim

α-PR for min-sum implies that $\forall A \subseteq C_i, \alpha d(A, C_i \setminus A) < d(A, C_j)$.

- Proof: blow up the distances between A and $C_i \setminus A$ by α
- Implication: by triangle inequality, if $\alpha \geq 3\frac{\max_i |C_i|}{\min_i |C_i| - 1}$,
  1. $\forall A_i \subseteq C_i, A_j \subseteq C_j$ s.t. $\min(|C_i \setminus A_i|, |C_j \setminus A_j|) > \min_i |C_i|/2$, $d_{avg}(A_i, A_j) > \min(d_{avg}(A_i, C_i \setminus A_i), d_{avg}(A_j, C_j \setminus A_j))$
  2. $\forall p \in C_i, q \notin C_i$, $2d_{avg}(p, C_i) < d(p, q)$

Algorithm for α-PR Min-Sum Instance

- Connect each point with its $\min_i |C_i|/2$ nearest neighbors
- Begin with each connected component being a cluster
- Repeatedly merge the two clusters with minimum average distance
- Output the tree with components as leaves and merges as internal nodes

Theorem

If $\alpha \geq 3\frac{\max_i |C_i|}{\min_i |C_i| - 1}$, then the tree output contains $\mathcal{OPT}$ as a pruning.

Keypoint of the Proof

- Implication 2 guarantees that the components are pure
- Implication 1 guarantees that no strict subset of an optimal cluster will be merged with a subset outside the cluster