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 $\left(3 + \epsilon\right)$ 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
Introduction

\(\alpha\)-Perturbation Resilience for \(k\)-median

\(\alpha, \epsilon\)-Perturbation Resilience for \(k\)-median

\(\alpha\)-Perturbation Resilience for Min-Sum

\(\alpha\)-Perturbation Resilience

\(\alpha\)-PR [Bilu and Linial, 2010, Awasthi et al., 2012]

A clustering instance \((S, d)\) is \(\alpha\)-perturbation resilient to a given objective function \(\Phi\) if for any function \(d' : S \times S \rightarrow R_{\geq 0}\) 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 \(OPT'\) for \(\Phi\) under \(d'\) and this clustering is equal to the optimal clustering \(OPT\) for \(\Phi\) under \(d\).
Main Results

- Polynomial time algorithm for finding $\text{OPT}$ for $\alpha$-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 $(\alpha, \epsilon)$-PR

- Polynomial time algorithm for finding $\text{OPT}$ for $\alpha$-PR min-sum instances when $\alpha \geq 3 \frac{\max_i |C_i|}{\min_i |C_i| - 1}$
Structure Properties of $\alpha$-PR $k$-Median Instance

Claim

$\alpha$-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 $\alpha$
- The $OPT$ 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 \( \alpha \)-PR \( k \)-Median Instance**

**Claim**

\( \alpha \)-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 \( \alpha \)
- The \( OPT \) 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 \( \alpha \geq 1 + \sqrt{2}, \forall p \in C_i, q \not\in C_i, d(c_i, p) < d(c_i, q) \) and \( d(c_i, p) < d(p, q) \)
Structure Properties of $\alpha$-PR $k$-Median Instance

Let $d_{i\text{max}} = \max_{p \in C_i} d(p, c_i)$. Construct a ball $B(c_i, d_{i\text{max}})$

- the ball covers exactly $C_i$
- points inside are closer to the center than to points outside, i.e.
  $\forall p \in B(c_i, d_{i\text{max}}), q \not\in B(c_i, d_{i\text{max}}), d(p, c_i) < d(p, q)$
**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. $\forall p \in B(c, d), q \notin B(c, d), d(c, p) < d(p, q)$. 

![Diagram of closure distance](image)
Algorithm for $\alpha$-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 $\text{OPT}$ as a pruning.

Maria Florina Balcan, Yingyu Liang
Clustering under Perturbation Resilience
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 $c_i \in A \cup B$
- $d_S(A, B) \leq d_i^{\text{max}} = \max_{p \in C_i} d(p, c_i)$
  - $d_i^{\text{max}}$ and $c_i \in A \cup B$ satisfy the two conditions of closure distance
Proof

- $d_S(A, A') > d_{i}^{\text{max}}$
  - Suppose the center $c$ for the ball defining $d_S(A, A')$ is from $A'$
  - Since $c \not\in C_i$, $d(c_i, p) < d(p, c)$ 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 $c \not\in C_i$, $d(c_i, c) > d_{i}^{\text{max}}$

- A similar argument holds for the case $c \in A$
(\alpha, \epsilon)-Perturbation Resilience

- \alpha-PR imposes a strong restriction that the OPT does not change after perturbation.
- We propose a more realistic relaxation.

\[(\alpha, \epsilon)-Perturbation\ Resilience\]

A clustering instance \((S, d)\) is \((\alpha, \epsilon)\)-perturbation resilient to a given objective function \(\Phi\) 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 \(OPT'\) for \(\Phi\) under \(d'\) is \(\epsilon\)-close to the optimal clustering \(OPT\) for \(\Phi\) under \(d\).
Structure Property of \((\alpha, \epsilon)-\text{PR} k\text{-median} \) instance

**Theorem**

Assume \(\min_i |C_i| > c\epsilon n\). Except for at most \(\epsilon n\) bad points, any other point is \(\alpha\) 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 \((\alpha, \epsilon)-\text{PR}\), there could be at most \(\epsilon n\) bad points
Algorithm for \((\alpha, \epsilon)\)-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 \(\epsilon\)-close to the optimal clustering. Moreover, the cost of this pruning is \((1 + O(\epsilon/\rho))\)-approximation where \(\rho = \min_i |C_i|/n\).
**$\alpha$-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 $OPT$ as a pruning.

- $\alpha$-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 $\alpha$
Introduction

\(\alpha\)-Perturbation Resilience for \(k\)-median

\((\alpha, \epsilon)\)-Perturbation Resilience for \(k\)-median

\(\alpha\)-Perturbation Resilience for Min-Sum

\(\alpha\)-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 \(OPT\) as a pruning.

- \(\alpha\)-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 $(\alpha, \epsilon)$-PR

Open Questions

- Design alg for $(\alpha, \epsilon)$-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 \((\alpha, \epsilon)\)-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 \(\alpha\), 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 \((\alpha, \epsilon)\)-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 \(\epsilon n\) 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 $\alpha$-PR Min-Sum Instance

Claim

$\alpha$-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 $\alpha$

- 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. $\min(|C_i \setminus A_i|, |C_j \setminus A_j|) > \min_i |C_i|/2$, $d_{\text{avg}}(A_i, A_j) > \min(d_{\text{avg}}(A_i, C_i \setminus A_i), d_{\text{avg}}(A_j, C_j \setminus A_j))$

  2. $\forall p \in C_i, q \not\in C_i$, $2d_{\text{avg}}(p, C_i) < d(p, q)$
**Algorithm for $\alpha$-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 $OPT$ 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