Perturbation-Resilient Clustering Problems

2023 年 12 月 1 日

Problem

Def. 1(k-Clustering with the l_p Objective)

An instance of k-clustering with the l_p objective($p \geq 1$)consists of a metric space (X, d) and a natural number k. The goal is to partition X into k disjoint clusters $C_1, \cdots, C_k$ and assign a center c_i to each cluster C_i so as to minimize the following objective function:

$$\sum_{i=1}^{k} \sum_{u \in C_i} d^p(u, c_i).$$

- For $p = \infty$, the objective function is $\max_{i \in \{1, \cdots, k\}, u \in C_i} |d(u, c_i)|$.

Problem

Def. 2(λ-Center Proximity)

Let (X, d) be an instance of the k-clustering problem with the l_p objective. Consider an optimal solution $C_1, \cdots, C_k$ with centers $c_1, \cdots, c_k$.

- We say that $c_1, \cdots, c_k$ satisfies the λ-**center proximity condition** (where $\lambda \geq 1$) if $\forall u \in C_i$ **and** $j \neq i$**, we have** $\lambda d(u, c_i) < d(u, c_j)$.
- We say that (X, d) has an optimal solution satisfying the λ-center proximity condition if there exists an optimal solution $C_1, \cdots, C_k$ with centers $c_1, \cdots, c_k$ satisfying the λ-center proximity condition.

Restatement of Perturnation Resilience

Def. 3(Perturbations and Metric Perturbations)

Consider a metric space (X, d).

- We say that **a symmetric function** $d' : X \times X \to \mathbb{R}^+$ is **a** γ**-perturbation of** d if for all $u, v \in X$ we have $\frac{1}{\gamma} d(u, v) \leq d'(u, v) \leq d(u, v)$.
- We say that d' is **a metric** γ**-perturbation of** d if d' is **a** γ**-perturbation of** d and a **metric** itself; i.e., d' satisfies the **triangle inequality**.

Restatement of Perturnation Resilience

Def. 4(Perturbation Resilience)

Consider an instance (X, d) of the k-clustering problem with the l_p objective. Let $C_1, \cdots, C_k$ be the optimal clustering. Then,

- (X, d) is γ-**perturbation resilient** if for **every** γ-perturbation of d, the unique optimal clustering of (X, d) is $C_1, \cdots, C_k$.
- Similarly, (X, d) is **metric** γ**-perturbation-resilient** if for **every** metric γ-perturbation of d, the unique optimal clustering of (X, d) is $C_1, \cdots, C_k$.

Observation

Thm. 5(Awasthi et al., 2012, and Angelidakis et al., 2017)

Let (X, d) be **a metric** γ**-perturbation-resilient** instance of the k-clustering problem with the l_p objective $(p \geq 1)$. Consider the unique optimal solution $C = (C_1, \cdots, C_k)$ and an optimal set of centers $\{c_1, \cdots, c_k\}$ (which is not necessarily unique). Then, **centers** $c_1, \cdots, c_k$ **satisfy the** γ**-center proximity property**.

Observation

Pf. 考虑 X 内任意一点 p. Let c_i be the closest center to p in $\{c_1, \cdots, c_k\}(p \in C_i)$ and c_j be another center. We need to show that $d(p, c_j) > \gamma d(p, c_i)$.
使用反证法．假设 $d(p, c_j) \leq \gamma d(p, c_i)$. 令 $r^* = d(p, c_i)$. 下面我们设法定义有助于我们解决问题的 d'.
考虑 X 上的完全图 $G = (X, E)$．对于每一条边 (u, v), 使 $len(u, v) = d(u, v)$，保持其它边的 len 不动，仅缩短 $len(p, c_j)$，即定义:

$$len'(u, v) = \begin{cases} r^*, & (u, v) = (p, c_j) \\ d(u, v), & (u, v) \neq (p, c_j) \end{cases}$$

现在我们在完全图 G 上考虑 len'，并据此定义 d' 为最短路的长度，易知:

$$d'(u, v) = min(d(u, v), d(u, p) + r^* + d(c_j, v), d(u, c_j) + r^* + d(p, v)).$$

Observation

(续) 由 $d(p, c_j) \le \gamma d(p, c_i)$，可得

$$\frac{1}{\gamma} d(u, v) \le d'(u, v) \le d(u, v),$$

故 d' 是 γ-perturbation.
又 d' 满足三角不等式（分类讨论容易验证），故 d' 是 metric γ-perturbation.
Thus, 由于 (X, d) 是 metric γ-perturbation-resilient，$C = (C_1, \cdots, C_k)$ 也是对于 d' 的唯一最优解, 故

$$d'(p, c_i) < d'(p, c_j).$$

而

$$d'(p, c_i) = d(p, c_i), d'(p, c_j) = d(p, c_i).$$

This leads to a contradiction!

Algorithm for 2-Perturbation-Resilient Instances

Thm. 6(Angelidakis et al., 2017)

There exists a polynomial-time algorithm that given an instance (X, d) of k-clustering with the l_p objective outputs an optimal solution if (X, d) has an optimal solution satisfying the 2-center proximity condition.

Algorithm. Consider the complete graph G on X , in which every edge (u,v) has length d(u,v).

1. Construct the minimum spanning tree (MST) T in G
2. Cluster T using dynamic programming (later(*))

Algorithm for 2-Perturbation-Resilient Instances

Thm. 7

Consider an instance (X, d) of k-clustering with the l_p objective.

- Let $C_1, \cdots, C_k$ be an optimal clustering with centers $c_1, \cdots, c_k$ satisfying the 2-center proximity condition
- Let $T = (X, E)$ be the minimum spanning tree (MST) in the complete graph on X with edge lengths $d(u, v)$.

Then, each cluster C_i is a subtree of T (i.e., for every two vertices $u, v \in C_i$, the unique shortest path from u to v in T completely lies within C_i).

Algorithm for 2-Perturbation-Resilient Instances

Lemma 8

Consider an instance (X, d) of the k-clustering problem with the l_p objective. Suppose that $C_1, \cdots, C_k$ is an optimal clustering for (X, d) and $c_1, \cdots, c_k$ is an optimal set of centers.

- If $c_1, \cdots, c_k$ satisfy the 2-center proximity property, then for every two distinct clusters C_i and C_j and all points $u \in C_i$ and $v \in C_j$, we have $d(u, c_i) < d(u, v)$.

Pf.

1. Since $c_1, \cdots, c_k$ satisfy the 2-center proximity property and by the triangle inequality, we have

$$2d(u, c_i) < d(u, c_j) \leq d(u, v) + d(v, c_j), \;\; 2d(v, c_j) < d(v, c_i) \leq d(u, v) + d(v, c_i).$$

2. Thus, we can sum the left inequality$\times \frac{2}{3}$ and right inequality$\times \frac{1}{3}$ to get the proof.

Algorithm for 2-Perturbation-Resilient Instances

Pf. of Thm. 7.

1. 我们只要证，在最小生成树 T 中，for every two vertices $u, v \in C_i$, the unique shortest path from u to v in T completely lies within C_i.
2. 从 C_i 中任取一点 u，u 到 c_i 在最小生成树 T 中有唯一的 path，途经的点设为 $u_1, \cdots, u_M$, 其中 $u_1 = u, u_M = c_i$. 下面我们只需要证明 $u_m \in C_i$, $m = 2, \cdots, M-1$. 由数学归纳法, 我们只需证明 $u_m \in C_i$ 可以推出 $u_{m+1} \in C_i$.
3. 由 MST 的 cycle poverty，在圈

$$u_m \to u_{m+1} \to \cdots \to u_M \to u_m$$

中，有

$$d(u_m, u_M) \geq d(u_m, u_{m+1}).$$

4. 由 Lemma 8，$u_{m+1} \in C_i$.

Algorithm for 2-Perturbation-Resilient Instances

(*)**Dynamic Program** Let us choose an arbitrary vertex r in X as a root for the MST T. Denote by T_u the subtree rooted at vertex u.

- $OPT(u, m) :=$ the optimal(minimum) cost of partitioning subtree T_u into m clusters that are subtrees of T.
- $OPT_{AC}(u, m, c) :=$ the optimal(minimum) cost of partitioning subtree T_u into m clusters subject to the following constraint: vertex u and all points in its cluster must be assigned to the center c.

The cost of k-clustering X equals $OPT(r, k)$. For simplicity, we can assume that the T is a binary tree (the general case can be handled by transforming any tree to a binary tree by adding "dummy" vertices). Let $left(u)$ be the left child of u and $right(u)$ be the right child of u. And we have

$$OPT(u, m) = \min_{c \in X} OPT_{AC}(u, m, c).$$

Algorithm for 2-Perturbation-Resilient Instances

- To find $OPT_{AC}(u, m, c)$, we find the optimal solutions for the left and right subtrees and combine them.
- To this end, we need to guess the number of clusters m_L and m_R in the left and right subtrees.
- We present formulas for $OPT_{AC}(u, m, c)$ in the four possible cases.

$$\begin{cases} \min\limits_{\substack{m_L, m_R \in \mathbb{Z}^+ \\ m_L + m_R = m+1}} & d(c, u) + OPT_{AC}(\mathit{left}(u), c, m_L) + OPT_{AC}(\mathit{right}(u), c, m_R) \\ \min\limits_{\substack{m_L, m_R \in \mathbb{Z}^+ \\ m_L + m_R = m}} & d(c, u) + OPT_{AC}(\mathit{left}(u), c, m_L) + OPT(\mathit{right}(u), c, m_R) \\ \min\limits_{\substack{m_L, m_R \in \mathbb{Z}^+ \\ m_L + m_R = m}} & d(c, u) + OPT(\mathit{left}(u), c, m_L) + OPT_{AC}(\mathit{right}(u), c, m_R) \\ \min\limits_{\substack{m_L, m_R \in \mathbb{Z}^+ \\ m_L + m_R = m-1}} & d(c, u) + OPT(\mathit{left}(u), c, m_L) + OPT(\mathit{right}(u), c, m_R) \end{cases}$$

Algorithm for 2-Perturbation-Resilient Instances

- We compute the values of $OPT_{AC}(u, m, c)$ in the 4 cases above(e.g. $p = 1$) and choose the minimum among them.
- The sizes of the DP tables for OPT and OPT_{AC} are $O(nk)$ and $O(n^2k)$, respectively.
- It takes $O(n)$ and $O(k)$ time to compute each entry in the tables OPT and OPT_{AC} , respectively. Thus, the total running time of the DP algorithm is $O(n^2k^2)$.

Now, we have proved Thm. 6.

$(3+\epsilon)$-Certified Local Search Algorithm for k-Medians

Now We take the k-medians($p = 1$) as an instance.

- Consider an arbitrary set of centers $c_1, \cdots, c_k$. $u \in C_i$ if and only if c_i is the closest center to u. Denote by $cost(c_1, \cdots, c_k)$ its cost.
- We now describe a 1-local search algorithm. The algorithm maintains a set of k centers $c_1, \cdots, c_k$.
- It starts with an arbitrary set of centers $c_1, \cdots, c_k$.
- While ($\exists$ pair (c_i, u), $cost(c_1, \cdots, c_{i-1}, u, c_{i+1}, \cdots, c_k) < cost(c_1, \cdots, c_k)$)
  {
  Replace the center c_i with u. (the size of the swaps is 1)
  }
- We call the obtained set of centers 1-locally optimal and denote it by L.

ρ-local search algorithm

- ρ-local search algorithm is a more powerful (alas less practical) version of the local search algorithm considering swaps of size up to ρ instead of 1.
- Its running time is exponential in ρ.

Thm. 9(Cohen-Addad and Schwiegelshohn, 2017 and Balcan and White, 2017)

The ρ-local search algorithm for k-medians outputs the optimal solution on $(3 + O(\frac{1}{\rho}))$-perturbation-resilient instances.

End

Thanks for listening!