大数据算法-2025 春

Lecture 12: k-中值问题的次线性算法

2025.4.11

Lecturer: 丁虎 *Scribe:* 王运韬

我们首先分析 **k-中值（k-median）问题**。给定一个有限度量空间 (V, μ)，V 表示点集，μ 表示两个点间的距离函数。k-中值问题旨在从 V 中选择 k 个中心，使得所有点到其最近中心的距离之和最小。

定义：

- $C = \{c_1, \ldots, c_k\} \subseteq V$ 为选出的 k 个中心。
- $\mu(v, C) = \min_{1 \le i \le k} \mu(v, c_i)$ 为点 v 到最近中心的距离。
- 定义总代价为：

$$\text{medopt}(V, k) = \min_{C \subseteq V, |C| = k} \sum_{v \in V} \mu(v, C)$$

- 平均成本为：

$$\text{med}_{\text{avg}}(V, k) = \frac{1}{|V|} \cdot \text{medopt}(V, k)$$

- 对任意子集 $U \subseteq V$ 和中心集合 C，定义其平均代价为：

$$\text{cost}_{\text{avg}}(U, C) = \frac{1}{|U|} \sum_{v \in U} \mu(v, C)$$

1 方法概述

我们研究以下自然采样方案的近似保证：

步骤 1： 从数据集合 V 中独立均匀随机（i.u.r.）选取大小为 s 的多重子集 S。

步骤 2： 在样本 S 上运行 α-近似算法 $\mathbb{A}$，得到解 C^*。

步骤 3： 输出 C^* 作为原问题的近似解。

1.1 分析框架

- **第一步**：证明 $\text{cost}(S, C_{\text{opt}})$ 能以高概率近似 $\text{cost}(V, C_{\text{opt}})$，其中 C_{opt} 是 V 的最优解。
- **第二步**：由于 C_{opt} 可能不在 S 中（k-中值问题需中心点属于样本），需证明 S 中存在一个可行解，其成本不超过 $\frac{c}{\alpha} \cdot \text{cost}(S, C_{\text{opt}})$（$c \geq \alpha$ 为常数）。
- **关键引理**：若某个解 C 对 V 的成本高于 $c \cdot \text{cost}(V, C_{\text{opt}})$，则它要么不是 S 的可行解，要么在 S 上的成本高于 $c \cdot \text{cost}(S, C_{\text{opt}})$。
- **结论**：由于算法 $\mathbb{A}$ 返回的解 C^* 在 S 上的成本不超过 $c \cdot \text{cost}(S, C_{\text{opt}})$，因此 C^* 对 V 的成本也以高概率不超过 $c \cdot \text{cost}(V, C_{\text{opt}})$。

在分析具体问题之前，我们先介绍本文中所采用的算法思路：

采样方案 (V, A, s)：

1. 从输入集合 V 中独立均匀地采样一个大小为 s 的多重集合 S；
2. 对样本集 S，运行某个 α-近似算法 A 得到解 C^*；
3. 返回解 C^* 作为最终结果（例如，k 个中心点）；

我们希望分析的问题是：当我们在 S 上运行近似算法 A 时，得到的解 C^* 对原始集合 V 的质量如何?

分析策略分两步：

1. 我们首先证明，对于给定问题的最优解 C_{opt}，其在 S 上的成本 $\text{cost}(S, C_{\text{opt}})$ 与在 V 上的成本 $\text{cost}(V, C_{\text{opt}})$ 是接近的，具有高概率保证。
2. 由于 C_{opt} 不一定是 S 的可行解（例如，C_{opt} 可能不在 S 中），我们进一步证明：存在一个 S 中的可行解，其成本最多为 $c \cdot \text{cost}(S, C_{\text{opt}})$（$c \geq \alpha$ 为某个常数）。

   由此我们得到：算法 A 会返回成本 $\leq c \cdot \text{cost}(S, C_{\text{opt}})$ 的解 C^*。

   再结合上述第一步，我们最终得到：

$$\text{cost}(V, C^*) \leq c \cdot \text{cost}(V, C_{\text{opt}})$$

从而保证了该采样算法在全体输入上是 c-近似的。

2 主要定理

定理 1

设 (V, μ) 是一个度量空间。令 $0 < \delta < 1$、$\alpha \geq 1$、$0 < \beta \leq 1$ 且 $\epsilon > 0$ 为近似参数。设 Λ 是一个在度量空间中解决 k-中值问题的 α-近似算法。如果我们选择一个样本集 $S \subseteq V$，其大小满足

$$s \geq \frac{c\alpha}{\beta}\left(k + \frac{\Delta}{\epsilon\beta}\left(\alpha \ln\left(\frac{1}{\delta}\right) + k \ln\left(\frac{k\alpha}{\epsilon\beta^2}\right)\right)\right),$$

其中 c 是某个正的常数，并对样本集 S 运行算法 Λ，那么对算法得到的解 C^*，以下式子以至少 $1 - \delta$ 的概率成立：

$$\text{cost}_{\text{avg}}(V, C^*) \leq 2(\alpha + \beta) \cdot \text{med}_{\text{avg}}(V, k) + \epsilon$$

为了开始分析解 C^* 的近似质量和定理 1 的证明，我们先介绍一些基本符号。

若集合 C 的平均代价满足

$$\text{cost}_{\text{avg}}(V, C) > (\alpha + \beta) \cdot \text{med}_{\text{avg}}(V, k)$$

则称 C 为 β-坏的 α-近似（β-bad α-approximation）。否则称其为 β-好的（good）α-近似。

Lemma 2.1. 设 S 是一个大小为

$$s \geq \frac{3\Delta\alpha(1 + \alpha/\beta)\ln(1/\delta)}{\beta \cdot \mathit{med}_{\mathit{avg}}(V, k)}$$

的多重集，从 V 中独立均匀随机选择。

如果一个 α-近似算法 $\mathcal{A}$ 在输入 S 上运行，那么对于 $\mathcal{A}$ 返回的解 C^*，以下结论成立：

$$\Pr\left[\mathit{cost}_{\mathit{avg}}(S, C^*) \leq 2(\alpha + \beta) \cdot \mathit{med}_{\mathit{avg}}(V, k)\right] \geq 1 - \delta$$

Lemma 2.2. 设 S 是从 V 中选出的 s 个点组成的多重集，并满足：

$$s \geq c\left((1+\alpha/\beta)k + \frac{(\alpha+\beta)\Delta\left(\ln(1/\delta) + k\ln\left(\frac{k(\alpha+\beta)\Delta}{\beta^2 \cdot med_{avg}(V,k)}\right)\right)}{\beta^2 \cdot \boldsymbol{med}_{avg}(V,k)}\right),$$

其中 c 是某个正的常数。

设 $\mathcal{C}$ 是 V 的 k-中值问题中所有 $(2\alpha+6\beta)$-坏解的集合。则

$$\Pr\left[\exists \mathcal{C}_b \in \mathcal{C} \text{ 满足} \mathcal{C}_b \subseteq S \text{ 且} \boldsymbol{cost}_{avg}(S,\mathcal{C}_b) \leq 2(\alpha+\beta)\cdot \boldsymbol{med}_{avg}(V,k)\right] \leq \delta$$

上述两个引理的证明参见 [1] 的 section 2. 引理 2.1 保证算法返回的解 C^* 在样本上表现良好，引理 2.2 保证这样的解 C^* 不会是一个在原问题中表现很差的 “伪优解”。

因此，定理 1 能够断言：在适当的采样规模下，样本上运行的近似算法会返回一个在全体数据集上也表现良好的聚类解，而且概率高达 $1-2\delta$. 接下来我们给出定理 1 的证明。

Proof. 设 β^* 是一个将在稍后证明中设定的正参数。设 s 被选择使得引理 2.2 和 2.3 的前提条件在 β 被 β^* 替换后成立，即

$$s \geq c(1+\alpha/\beta^*)\left(k + \frac{\Delta}{\beta^* med_{\text{avg}}(V,k)}\left(\alpha\ln(1/\delta) + k\ln\left(\frac{k(\alpha+\beta^*)\Delta}{(\beta^*)^2 med_{\text{avg}}(V,k)}\right)\right)\right) \quad (1)$$

对于某个常数 c。

设 S 是从 V 中以均匀分布随机选择的 s 个点的多重集。那么根据引理 2.3, 以至少 $1-\delta$ 的概率，没有集合 $C \subseteq S$ 是 V 的 k-中值问题的 $(2\alpha+6\beta^*)$-坏解，同时满足不等式

$$cost_{\text{avg}}(S,C) \leq 2(\alpha+\beta^*)med_{\text{avg}}(V,k)$$

另一方面，如果我们对集合 S 运行算法 A，则根据引理 2.2，结果集合 C^* 以至少 $1-\delta$ 的概率满足

$$cost_{\text{avg}}(S,C^*) \leq 2(\alpha+\beta^*)med_{\text{avg}}(V,k)$$

结合上述结论表明，以至少 $1-2\delta$ 的概率，集合 C^* 是 V 的 k-中值问题的 $(6\beta^*, 2\alpha)$-好解，即

$$\Pr\left[cost_{\text{avg}}(V,C^*) \leq (2\alpha+6\beta^*)med_{\text{avg}}(V,k)\right] \geq 1-2\delta$$

为了完成证明，我们只需要在公式 (5) 中设定参数 β 和 ϵ 并证明 $med_{\text{avg}}(V,k)$ 在 $1/\delta$ 的范围内。

首先我们考虑当 $\text{med}_{\text{avg}}(V,k) \leq \epsilon$ 的情况。我们使用公式 (5) 和 (6)，并令 $\beta^* = \frac{1}{6} \cdot \frac{\epsilon}{\text{med}_{\text{avg}}(V,k)}$。由于 $\beta^* \geq \frac{1}{6}$，那么如果满足条件

$$s \geq c(1+\alpha)\left(k + \frac{\Delta}{\epsilon}\left(\alpha \ln\left(\frac{1}{\delta}\right) + k \ln\left(\frac{k(\alpha+1)\Delta}{\epsilon}\right)\right)\right),$$

其中 c 是某个正的常数，则以至少 $1-2\delta$ 的概率，我们可以得到：

$$\text{cost}_{\text{avg}}(V, C^*) \leq (2\alpha + 6\beta^*) \cdot \text{med}_{\text{avg}}(V,k) = 2\alpha \cdot \text{med}_{\text{avg}}(V,k) + \epsilon$$

注意这个界限与参数 β 无关。

接下来我们考虑当 $\text{med}_{\text{avg}}(V,k) > \epsilon$ 的情况。此时，由公式 (5) 和 (6) 可知，对于某个正的常数 c，并令 $\beta = 3\beta^*$，如果满足

$$s \geq c\left(1 + \frac{\alpha}{\beta}\right)\left(k + \frac{\Delta}{\beta\epsilon}\left(\alpha \ln\left(\frac{1}{\delta}\right) + k \ln\left(\frac{k\Delta(1+\alpha/\beta)}{\beta^2\epsilon}\right)\right)\right),$$

那么以至少 $1-2\delta$ 的概率，我们可以得到：

$$\text{cost}_{\text{avg}}(V, C^*) \leq 2(\alpha+\beta) \cdot \text{med}_{\text{avg}}(V,k)$$

结合上述两个情况的界限，得证。 □

References

[1] A. Czumaj and C. Sohler. Sublinear-time approximation algorithms for clustering via random sampling. *Random Structures & Algorithms*, 30(1-2):226–256, 2007.