中国科学 : 信息科学 2025年 第 55 卷 第 1 期: 32–45
SCIENTIA SINICA Informationis
论文

多样性公平 k-中位问题的 $(1+\varepsilon)$-近似算法

张震[1,2,3], 陈晓红[1,2*], 刘利枚[1,2*], 任剑[1,2], 姜林[1,2], 冯启龙[4,2]

1. 湖南工商大学前沿交叉学院, 长沙 410205
2. 湘江实验室, 长沙 410205
3. 中南大学商学院, 长沙 410083
4. 中南大学计算机学院, 长沙 410083
* 通信作者. E-mail: csu_cxh@163.com, seagullm@163.com

收稿日期: 2024–04–09; 修回日期: 2024–06–14; 接受日期: 2024–07–30; 网络出版日期: 2024–12–26

国家自然科学基金基础科学中心项目 (批准号: 72088101)、国家自然科学基金 (批准号: 62202161, 62376092, 62172446)、科技部重点研发计划 (批准号: 2021YFC3300603)、湖南省自然科学基金 (批准号: 2023JJ40240) 和湖南省教育厅科学研究 (批准号: 23B0597) 资助

摘要 多样性公平 k-中位问题在数据摘要等对聚类中心选取方式的公平性要求较高的聚类应用领域发挥重要作用. 给定一个用户集合、ℓ 个设施集合以及正整数 k, 该问题的目标是在每个设施集合中开设一个规模受限的子集, 使得开设设施数量不超过 k, 且每个用户与距离最近的开设设施之间具有较高的相似度. 本文将多样性公平 k-中位问题实例映射为低维空间中的小规模实例, 并围绕实例中的点划分空间以估计最优解中开设设施的位置. 基于这一思路, 本文在 d-维欧氏空间中为多样性公平 k-中位问题提出了时间复杂度为 $O(ndk\log n)+2^{\ell k+(k\varepsilon^{-1})^{O(1)}}n^{O(1)}$ 的 $(1+\varepsilon)$-近似算法, 其中, n 为设施与用户数量之和. 该结果改进了此前人们在更一般化的度量空间中利用相近的固定参数时间得到的 $(1+2\mathrm{e}^{-1}+\varepsilon)$-近似比.

关键词 固定参数算法, 近似算法, k-中位问题, 设施选址, 采样

1 引言

k-中位 (k-median) 问题是一个被广泛研究的聚类问题. 给定度量空间中的一个用户集合和一个设施集合以及正整数 k, k-中位问题要求选取最多 k 个开设设施 (或称为聚类中心) 并将每个用户连接到距离最近的开设设施, 使得用户的连接费用之和最小, 其中, 每个用户的连接费用为该用户与对应设施之间的距离. 大量研究工作致力于 k-中位问题求解算法的设计和分析. 目前, 关于该问题的最好近似结果是 Gowda 等 [1] 通过调整实例的亚可行解得到的 $(2.613+\varepsilon)$-近似比. 当实例中的点位于更特殊的高维欧氏空间时, Cohen-Addad 等 [2] 利用空间性质得到了 $(2.406+\varepsilon)$-近似比. 此外, 在维度为常数的低维欧氏空间中, 人们为 k-中位问题提出了一系列 $(1+\varepsilon)$-近似算法 [3~5].

引用格式: 张震, 陈晓红, 刘利枚, 等. 多样性公平 k-中位问题的 $(1+\varepsilon)$-近似算法. 中国科学: 信息科学, 2025, 55: 32–45, doi: 10.1360/SSI-2024-0108
Zhang Z, Chen X H, Liu L M, et al. A $(1+\varepsilon)$-approximation algorithm for diversity-aware k-median. Sci Sin Inform, 2025, 55: 32–45, doi: 10.1360/SSI-2024-0108

在 k-中位问题的解中, 用户与对应的开设设施之间具有较高的相似度. 这些开设设施可以作为用户集合的代表性数据点. 因此, k-中位问题的求解算法被广泛应用在数据摘要 (data summarization) 领域. 然而, 通过求解 k-中位问题得到的摘要在很多情况下缺乏公平性. 例如: 在构造职业图像搜索结果时, 人们基于数据摘要方法在数据库中选取代表性图像; k-中位问题旨在使图像数据对应的用户集合有最小的连接费用之和, 无法保证所选取的开设设施集合能公平地反映数据中性别、年龄等属性的分布情况 [6]. 鉴于此, Thejaswi 等 [7] 提出了多样性公平 k-中位 (diversity-aware k-median) 问题. 给定 ℓ 个设施集合, 该问题要求在每个集合中开设数量不小于给定下限的设施, 以保证每个集合在开设设施数量上的公平性.

定义1 (多样性公平 k-中位问题)　多样性公平 k-中位问题的一个实例 $(\ell, \{\mathcal{F}_1, \ldots, \mathcal{F}_\ell\}, \mathcal{C}, k, \boldsymbol{r})$ 包含正整数 ℓ、度量空间中的 ℓ 个设施集合 $\mathcal{F}_1, \ldots, \mathcal{F}_\ell$ 和一个用户集合 $\mathcal{C}$、不超过 $|\bigcup_{i=1}^{\ell} \mathcal{F}_i|$ 的正整数 k 和 ℓ 个非负整数组成的向量 $\boldsymbol{r} = (r_1, \ldots, r_\ell)$. 该实例的一个可行解是一个满足 $|\mathcal{S}| \leqslant k$, $\mathcal{S} \subseteq \bigcup_{i=1}^{\ell} \mathcal{F}_i$ 和 $|\mathcal{S} \cap \mathcal{F}_i| \geqslant r_i \, \forall i \in \{1, \ldots, \ell\}$ 的设施集合 $\mathcal{S}$, 其费用为 $\sum_{c \in \mathcal{C}} \min_{f \in \mathcal{S}} \Delta(c, f)$, 其中, $\Delta(c, f)$ 是 c 与 f 之间的距离. 多样性公平 k-中位问题的目标是找到费用最低的可行解.

人们基于动态规划、随机舍入等技术为多样性公平 k-中位问题提出了一系列启发式算法 [8]. 然而, 这些算法缺乏可被证明的近似保证. 在实例中的设施集合互不相交的假设下, Thejaswi 等 [7] 证明了在每轮迭代中交换 $O(1)$ 个设施的局部搜索算法在 $\ell = 2$ 时是关于多样性公平 k-中位问题的 $O(1)$-近似算法. 在此基础上, Zhang 等 [9] 进一步证明了在每轮迭代中交换 $O(\ell)$ 个设施的局部搜索算法有不高于 $O(\ell)$ 的近似比. Hotegni 等 [10] 基于相同假设为多样性公平 k-中位问题提出了具有常数近似保证的线性规划舍入算法. 然而, 在不满足这一假设条件的情况下, 多样性公平 k-中位问题的求解难度明显提升: Thejaswi 等 [7] 基于支配集 (dominating set) 问题实例给出的归约结果说明, 即使是判定多样性公平 k-中位问题的给定实例是否有可行解也是 NP-难的.

在涉及多样性公平 k-中位问题的实际应用中, 开设设施数量上限 k 和设施集合数量 ℓ 通常明显小于实例规模. 因此, 假设 k 和 ℓ 取值较小是松弛该问题的可行手段. 在这一松弛条件下, Thejaswi 等 [8] 以 k 和 ℓ 作为固定参数, 基于次模最大化方法为多样性公平 k-中位问题提出了固定参数时间 (即 $h(k, \ell, \varepsilon) n^{O(1)}$ 时间, 其中, h 为任意正值函数, n 为用户与设施数量之和) 的 $(1 + 2\mathrm{e}^{-1} + \varepsilon)$-近似算法.

1.1　主要结果

关于多样性公平 k-中位问题的参数复杂性结果表明, 人们无法基于现有技术改进 Thejaswi 等 [8] 提出的固定参数时间近似比: 当 k 和 ℓ 为固定参数时, Thejaswi 等 [8] 证明了多样性公平 k-中位问题是 W[2]-难问题, 这说明不存在关于该问题的固定参数时间精确算法; Cohen-Addad 等 [11] 以间隙指数时间假设 [12] 成立为条件, 证明了即使在 k 为固定参数且 $\ell = 1$ 的情况下, 多样性公平 k-中位问题固定参数时间算法的近似比也不可能小于 $1 + 2\mathrm{e}^{-1} - \varepsilon$. 然而, 这些复杂性结果只在一般化的度量空间中成立. 由于聚类问题在实际应用中的大部分实例分布在更特殊的欧氏空间中, 在该类空间中为多样性公平 k-中位问题提出更好的近似结果是一个值得探索的方向.

实际上, 当实例位于欧氏空间中且设施可以被开设在空间中的任意位置时 (即设施集合等同于 $\mathbb{R}^d$), 人们已经为 k-中位问题提出了一系列时间复杂度为 $(nd)^{O(1)} h(k, \varepsilon)$ 的 $(1+\varepsilon)$-近似算法 [13~16]. 给定用户子集 $\mathcal{C}' \subset \mathbb{R}^d$, 这些算法将满足 $\sum_{c \in \mathcal{C}'} ||c - f|| = \min_{f' \in \mathbb{R}^d} \sum_{c \in \mathcal{C}'} ||c - f'||$ 的集合中位点 f 作为与 $\mathcal{C}'$ 对应的最优开设设施. 然而, 在设施开设位置受限的多样性公平 k-中位问题中, 由于集合中位点无法作为可行解中的开设设施, 如何有效利用欧氏空间的性质提出比 $1 + 2\mathrm{e}^{-1} + \varepsilon$ 更好的固定参数时间近似结果还是未知的. 本文基于不同的思路求解多样性公平 k-中位问题, 在欧氏空间中提出了近似比为 $1 + \varepsilon$ 的固定参数时间近似算法, 如定理 1 所述.

定理1　给定常数 $\varepsilon \in (0,1)$ 以及满足 $\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C} \subset \mathbb{R}^d$ 和 $|\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C}| = n$ 的多样性公平 k-中位问题实例 $\mathcal{I} = (\ell, \{\mathcal{F}_1, \ldots, \mathcal{F}_\ell\}, \mathcal{C}, k, \boldsymbol{r})$, 我们可以在不超过 $O(2^{\ell k} nk)$ 的时间内判定 $\mathcal{I}$ 是否有可行解. 此外, 如果 $\mathcal{I}$ 有可行解, 则存在时间复杂度为 $O(nd \log n) + 2^{\ell k + (k\varepsilon^{-1})^{O(1)}} n^{O(1)}$ 且近似比为 $1+\varepsilon$ 的随机近似算法.

1.2　基本定义及引理

令 ϵ 表示 $(0, \frac{1}{2})$ 内的一个常数. 给定整数 $i \geqslant 1$, 令 $[i] = \{1, \ldots, i\}$. 给定欧氏空间中的两个点 x 和 y 以及一个集合 $\mathcal{Z}$, 令 $\Delta(x,y) = ||x-y||$ 表示 x 和 y 之间的距离, 并令 $\Delta(x, \mathcal{Z}) = \min_{z \in \mathcal{Z}} \Delta(x,z)$ 表示 $\mathcal{Z}$ 中的点与 x 之间的最小距离. 令 $\mathcal{I} = (\ell, \{\mathcal{F}_1, \ldots, \mathcal{F}_\ell\}, \mathcal{C}, k, \boldsymbol{r})$ 表示多样性公平 k-中位问题的一个实例, 其中, $\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C} \subset \mathbb{R}^d$ 且 $|\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C}| = n$. 给定设施 $f \in \bigcup_{i=1}^{\ell} \mathcal{F}_i$, 本文将满足 $f \in \mathcal{F}_i$ 的整数 i 称为 f 的类别.

本文基于以下引理分析算法的时间复杂度.

引理1　给定大于 1 的实数 i 和 j, 不等式 $\log^j i \leqslant \max\{i, j^{O(j)}\}$ 成立.

证明　当 $j \geqslant \frac{\log i}{\log\log i}$ 时, 不等式 $\log^j i \leqslant j^{O(j)}$ 成立, 而当 $j < \frac{\log i}{\log\log i}$ 时, 不等式 $\log^j i < \log^{\frac{\log i}{\log\log i}} i = i$ 成立. 由此可知, 引理 1 正确.

本文基于约翰逊 – 林登施特劳斯变换 (Johnson-Lindenstrauss transform) 将高维欧氏空间中的实例映射到低维空间中.

引理2 (约翰逊 – 林登施特劳斯变换 [17])　给定集合 $\mathcal{X} \subset \mathbb{R}^d$ 和常数 $\epsilon \in (0, \frac{1}{2})$, 可以在不超过 $O(d|\mathcal{X}|\log|\mathcal{X}|)$ 的时间内构造满足 $\tilde{d} = O(\epsilon^{-2}\log|\mathcal{X}|)$ 的映射 $g: \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$, 使得任意 $x, y \in \mathcal{X}$ 都满足 $\Delta(g(x), g(y)) \in [1, 1+\epsilon]\Delta(x,y)$.

以下引理是 Narayanan 和 Nelson [18] 针对约翰逊 – 林登施特劳斯变换提出的加强版.

引理3 ([18])　给定集合 $\mathcal{X} \subset \mathbb{R}^d$ 和常数 $\epsilon \in (0, \frac{1}{2})$, 可以在不超过 $(d|\mathcal{X}|)^{O(1)}$ 的时间内构造满足 $\tilde{d} = O(\epsilon^{-2}\log|\mathcal{X}|)$ 的映射 $g: \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$, 使得任意 $x \in \mathcal{X}$ 和 $y \in \mathbb{R}^d$ 都满足 $\Delta(g(x), g(y)) \in [1, 1+\epsilon]\Delta(x,y)$.

与约翰逊 – 林登施特劳斯变换 (引理 2) 相比, 引理 3 中算法的时间复杂度更高, 但该引理在更大范围内保证了原空间与其低维映射之间的距离相似性. 具体来说, 当以集合 $\mathcal{X} \subset \mathbb{R}^d$ 为输入时, 引理 2 只能保证 $\mathcal{X}$ 中任意两点之间的距离在映射前后的相似性, 而引理 3 能保证 $\mathcal{X}$ 中任意一点与 $\mathbb{R}^d$ 中任意一点之间的距离在映射前后的相似性. 本文假设引理 2 和 3 构造的映射 $g: \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ 都是单映射. 这一假设不失一般性: 本文可以通过复制 $\mathbb{R}^{\tilde{d}}$ 中有多个原像的点区分 $\mathbb{R}^d$ 中每个点的像.

本文还将基于 Chen [13] 提出的核心集构造方法压缩实例规模. 以下引理是该方法的性能保证.

引理4 ([13])　给定集合 $\mathcal{X} \subset \mathbb{R}^d$、常数 $\epsilon \in (0, \frac{1}{2})$ 和正整数 k, 可以在 $O(|\mathcal{X}|dk)$ 时间内构造带有权重函数 $w: \mathcal{X}' \to [1, +\infty)$ 且满足 $\sum_{c \in \mathcal{X}'} w(c) = |\mathcal{X}|$ 和 $|\mathcal{X}'| \leqslant d(k\epsilon^{-1}\log|\mathcal{X}|)^{O(1)}$ 的加权子集 $\mathcal{X}' \subseteq \mathcal{X}$, 使得任意规模不超过 k 的集合 $\mathcal{S} \subset \mathbb{R}^d$ 都满足 $\sum_{c \in \mathcal{X}'} w(c)\Delta(c, \mathcal{S}) \in [1-\epsilon, 1+\epsilon] \sum_{c \in \mathcal{X}} \Delta(c, \mathcal{S})$.

本文基于算法 1 判定实例 $\mathcal{I}$ 是否有可行解. 算法 1 首先生成设施类别集合 $[\ell]$ 的幂集 $\mathbb{L}$ 及其 k 次笛卡尔乘积 (Cartesian product) $[\mathbb{L}]^k$, 其中, $[\mathbb{L}]^k$ 中的每个子集组合都对应 k 个开设设施的类别. 对于每个子集组合, 算法 1 在第 5 步验证其对应的设施类别是否满足实例的公平性约束条件, 并在第 9 步检查实例中是否存在满足其类别要求的设施集合. 如果 $[\mathbb{L}]^k$ 中存在满足所有条件的子集组合, 则算法 1 返回 `Ture`, 否则, 该算法返回 `False`. 以下引理给出了算法 1 的时间复杂度和正确性.

引理5　给定多样性公平 k-中位问题的实例 $\mathcal{I} = (\ell, \{\mathcal{F}_1, \ldots, \mathcal{F}_\ell\}, \mathcal{C}, k, \boldsymbol{r})$, 算法 1 的时间复杂度为 $O(2^{\ell k}|\bigcup_{i=1}^{\ell} \mathcal{F}_i|k)$. 此外, 当且仅当算法 1 返回 `True` 时, 实例 $\mathcal{I}$ 有可行解.

证明　令 $\mathbb{L}$ 为集合 $[\ell]$ 的幂集, 并令 $[\mathbb{L}]^k$ 表示 $\mathbb{L}$ 的 k 次笛卡尔乘积. 可以得出, $|[\mathbb{L}]^k| = |\mathbb{L}|^k = 2^{\ell k}$. 对于 $[\mathbb{L}]^k$ 中的每个子集组合, 算法 1 在第 5 步花费 $O(\ell k)$ 时间验证其中的设施类别是否满足公平性

算法 1 多样性公平 k-中位问题的可行解存在性判定算法.

输入: 多样性公平 k-中位问题的实例 $\mathcal{I}=(\ell,\{\mathcal{F}_1,\ldots,\mathcal{F}_\ell\},\mathcal{C},k,\boldsymbol{r})$;
输出: 布尔值 `bool` $\in\{$`True`,`False`$\}$;

```
1: bool ← False;
2: 令 𝕃 为 [ℓ] 的幂集;
3: 令 [𝕃]^k 为笛卡尔乘积 𝕃 × ... × 𝕃 (k 个);
4: for each (ℒ_1, ..., ℒ_k) ∈ [𝕃]^k do
5:    if |{i ∈ [k] : j ∈ ℒ_i}| ⩾ r_j ∀ j ∈ [ℓ] then
6:       for each i ∈ [k] do
7:          令 γ(i) 为 (ℒ_1, ..., ℒ_k) 中与 ℒ_i 相等的集合数量;
8:       end for
9:       if |(⋂_{j∈ℒ_i} ℱ_j)\(⋃_{j∈[ℓ]\ℒ_i} ℱ_j)| ⩾ γ(i) ∀ i ∈ [k] then
10:         bool ← True;
11:      end if
12:   end if
13: end for
14: return bool.
```

约束条件, 并在第 9 步花费 $O(|\bigcup_{i=1}^{\ell}\mathcal{F}_i|k)$ 时间验证是否存在满足其类别要求的设施集合. 由此可知, 算法 1 的时间复杂度为 $O(2^{\ell k}|\bigcup_{i=1}^{\ell}\mathcal{F}_i|k)$.

下面分析算法 1 的正确性. 在算法 1 返回 `True` 的情况下, 令 $(\mathcal{L}_1,\ldots,\mathcal{L}_k)\in[\mathbb{L}]^k$ 为满足算法 1 中两个判定条件的一个子集组合. 由算法 1 第 9 步中的判定条件可知, 存在一个满足

$$\{j\in[\ell]:f_i\in\mathcal{F}_j\}=\mathcal{L}_i\,\forall\, i\in[k] \tag{1}$$

且不包含重复元素的设施集合 $\{f_1,\ldots,f_k\}\subseteq\bigcup_{i=1}^{\ell}\mathcal{F}_i$. 可以得出, 每个整数 $j\in[\ell]$ 都满足

$$|\{f_1,\ldots,f_k\}\cap\mathcal{F}_j|=|\{i\in[k]:j\in\mathcal{L}_i\}|\geqslant r_j, \tag{2}$$

其中, 第 1 步根据等式 (1) 得出, 第 2 步根据算法 1 第 5 步中的判定条件得出. 不等式 (2) 说明 $\{f_1,\ldots,f_k\}$ 是实例 $\mathcal{I}$ 的可行解. 由此可知, 当算法 1 返回 `True` 时, 实例 $\mathcal{I}$ 有可行解.

在实例 $\mathcal{I}$ 有可行解的情况下, 不等式 $k\leqslant|\bigcup_{i=1}^{\ell}\mathcal{F}_i|$ 说明实例 $\mathcal{I}$ 有一个规模为 k 的可行解 $\{f_1,\ldots,f_k\}\subseteq\bigcup_{i=1}^{\ell}\mathcal{F}_i$. 给定整数 $i\in[k]$, 令 $\mathcal{L}_i=\{j\in[\ell]:f_i\in\mathcal{F}_j\}$, 并令 $\gamma(i)$ 表示 $(\mathcal{L}_1,\ldots,\mathcal{L}_k)$ 中与 $\mathcal{L}_i$ 相等的集合数量. 由 $\mathcal{L}_i$ 的定义以及 $\{f_1,\ldots,f_k\}$ 的可行性可知, 每个整数 $j\in[\ell]$ 都满足

$$|\{i\in[k]:j\in\mathcal{L}_i\}|=|\{f_1,\ldots,f_k\}\cap\mathcal{F}_j|\geqslant r_j. \tag{3}$$

此外, 给定整数 $i\in[k]$, $\mathcal{L}_i$ 和 $\gamma(i)$ 的定义说明

$$\left|\left(\bigcap_{j\in\mathcal{L}_i}\mathcal{F}_j\right)\setminus\left(\bigcup_{j\in[\ell]\setminus\mathcal{L}_i}\mathcal{F}_j\right)\right|\geqslant\left|\left(\bigcap_{j\in\mathcal{L}_i}\mathcal{F}_j\right)\setminus\left(\bigcup_{j\in[\ell]\setminus\mathcal{L}_i}\mathcal{F}_j\right)\cap\{f_1,\ldots,f_k\}\right|=\gamma(i). \tag{4}$$

结合不等式 (3) 和 (4) 可知, 子集组合 $(\mathcal{L}_1,\ldots,\mathcal{L}_k)$ 满足算法 1 第 5 步和第 9 步中的判定条件. 因此, 当实例 $\mathcal{I}$ 有可行解, 算法 1 返回 `True`.

综上所述, 当且仅当算法 1 返回 `True` 时, 实例 $\mathcal{I}$ 有可行解. 因此, 引理 5 成立.

2　求解思路

本文基于 Cohen-Addad 等 [11] 提出的算法设计框架求解多样性公平 k-中位问题. 该框架首先在用户集合中搜索与最优解中的开设设施距离较近的一组引导点 (leaders), 然后在以引导点为中心、半

径接近于引导点与对应设施之间距离的一组环形空间中选取候选开设设施. 人们已经利用这一框架为一系列 k-中位相关问题设计了固定参数时间的求解算法 [8,19~22], 其中包括 Thejaswi 等 [8] 为多样性公平 k-中位问题提出的 $(1+2\mathrm{e}^{-1}+\varepsilon)$-近似算法. 本文在求解多样性公平 k-中位问题时, 结合欧氏空间的性质进一步压缩候选开设设施的选取范围, 在固定参数时间内得到了更好的近似比.

在为多样性公平 k-中位问题构造近似解时, Thejaswi 等 [8] 在以引导点为中心的整组环形空间中基于次模最大化方法选取开设设施. 通过这一方式构造的近似解与最优解之间费用的差值可能与环形空间的直径相关. 本文提出的算法改进策略是将环形空间划分为一组规模较小的网格单元; 对于最优解中的每个开设设施, 本文将其所在网格单元的中心点作为近似解中的开设设施. 在该过程中, 本文构造以下定义中的数据网 (data net).

定义2 (数据网 [23])　给定集合 $\mathcal{X}\subset\mathbb{R}^d$、距离参数 $\tau>0$ 和子集 $\mathcal{N}\subseteq\mathcal{X}$, 如果任意 $x\in\mathcal{N}$ 都满足 $\Delta(x,\mathcal{N}\backslash\{x\})>\tau$ 且任意 $x\in\mathcal{X}$ 都满足 $\Delta(x,\mathcal{N})<\tau$, 则称 $\mathcal{N}$ 为 $\mathcal{X}$ 的 τ-数据网.

Har-Peled 和 Mendel [24] 证明了当给定集合位于低维空间时, 可以在近线性时间内为其构造数据网, 如引理 6 所述.

引理6 ([24])　给定集合 $\mathcal{X}\subset\mathbb{R}^d$ 和距离参数 $\tau>0$, 可以在 $|\mathcal{X}|\log|\mathcal{X}|2^{O(d)}$ 时间内为 $\mathcal{X}$ 构造规模不超过 $\max\{|\mathcal{X}|,\left(\epsilon^{-1}\max_{x,y\in\mathcal{X}}\Delta(x,y)\right)^d\}$ 的 τ-数据网.

数据网的定义为本文划分环形搜索空间并选取开设设施提供了明确的思路. 对于每个引导点和以其为中心的环形空间, 本文根据实例的公平性约束条件选取空间中包含最优开设设施的一个设施子集, 并根据空间直径确定距离参数、为该子集构造数据网. 由数据网的定义可知, 当考虑上述数据网对应的维诺图 (Voronoi diagram) 时, 最优解中的每个设施与所在维诺单元的中心点之间都有较短的距离. 通过将数据网中的每个设施作为候选开设设施, 本文可以保证候选开设设施集合有一个与最优解较为接近的子集.

在基于数据网构造近似解时, 本文面对的一个问题是现有数据网构造方法的时间复杂度及其构造的数据网规模都对空间维度有指数级依赖. 为了保证求解算法的时间复杂度在高维欧氏空间中是关于 k 和 ℓ 的固定参数时间, 本文基于约翰逊 – 林登施特劳斯变换和核心集构造方法将实例 $\mathcal{I}$ 映射为 $O(\log k+\log\log n)$-维空间中用户数量不超过 $(k\log n)^{O(1)}$ 的小规模实例. 结合这一实例消减方法与基于数据网选取开设设施的求解思路, 本文为多样性公平 k-中位问题提出了固定参数时间的 $(1+\varepsilon)$-近似算法.

3　实例消减

如前文所述, 本文通过构造设施集合的数据网压缩开设设施搜索范围. 为了降低数据网规模、保证算法的时间复杂度为固定参数时间, 本节将实例 $\mathcal{I}$ 映射为低维空间中的小规模实例. 在该过程中, 本节考虑用户带有权重的加权多样性公平 k-中位问题.

定义3 (加权多样性公平 k-中位问题)　给定正整数 ℓ, 度量空间中的 ℓ 个设施集合 $\mathcal{F}_1,\ldots,\mathcal{F}_\ell$ 和一个用户集合 $\mathcal{C}$、权重函数 $w:\mathcal{C}\to[1,+\infty)$、正整数 k 和 ℓ 个非负整数组成的向量 $\boldsymbol{r}=(r_1,\ldots,r_\ell)$, 多样性公平 k-中位问题的加权实例 $(\ell,\{\mathcal{F}_1,\ldots,\mathcal{F}_\ell\},\mathcal{C},w,k,\boldsymbol{r})$ 与非加权实例 $(\ell,\{\mathcal{F}_1,\ldots,\mathcal{F}_\ell\},\mathcal{C},k,\boldsymbol{r})$ 之间唯一的区别是加权实例以 $\sum_{c\in\mathcal{C}}w(c)\Delta(c,\mathcal{S})$ 作为解 $\mathcal{S}\subseteq\bigcup_{i=1}^{\ell}\mathcal{F}_i$ 的费用.

当以用户集合作为输入调用引理 3 中的算法时, 我们可以在不明显改变每对用户与设施之间距离的前提下, 将实例映射到维度与用户数量呈对数相关的低维空间中. 这一结论为本节提供了可行的实例消减思路. 本节首先利用引理 4 压缩用户集合规模, 然后以被压缩的用户集合作为输入调用引理 3 中的降维算法. 此外, 由于引理 4 构造的用户子集规模与空间维度相关, 本节在使用引理 4 之前先基于引理 2 在近线性时间内将实例 $\mathcal{I}$ 映射到维度与 d 无关的低维空间中. 算法 2 中给出了本

算法 2 多样性公平 k-中位问题实例消减方法.

输入: 常数 $\epsilon \in (0, \frac{1}{2})$ 以及满足 $\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C} \subset \mathbb{R}^d$ 的实例 $\mathcal{I} = (\ell, \{\mathcal{F}_1, \dots, \mathcal{F}_\ell\}, \mathcal{C}, k, \boldsymbol{r})$;

输出: 映射 $g : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ 以及满足 $\bigcup_{i=1}^{\ell} \tilde{\mathcal{F}}_i \cup \tilde{\mathcal{C}} \subset \mathbb{R}^{\tilde{d}}$ 和 $\tilde{\mathcal{F}}_i = \{g(f) : f \in \mathcal{F}_i\} \forall i \in [\ell]$ 的加权实例 $(\ell, \{\tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_\ell\}, \tilde{\mathcal{C}}, w, k, \boldsymbol{r})$;

1: 令 $g_1 : \mathbb{R}^d \to \mathbb{R}^{d'}$ 为基于引理 2 为常数 ϵ 和集合 $\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C}$ 构造的映射;
2: 令 $\mathcal{C}'$ 为基于引理 4 为常数 ϵ、集合 $\{g_1(c) : c \in \mathcal{C}\}$ 和正整数 k 构造的加权集合, 并令 $w' : \mathcal{C}' \to [1, +\infty)$ 为权重函数;
3: 令 $g_2 : \mathbb{R}^{d'} \to \mathbb{R}^{\tilde{d}}$ 为基于引理 3 为常数 ϵ 和集合 $\mathcal{C}'$ 构造的映射;
4: $\tilde{\mathcal{C}} \leftarrow \{g_2(c) : c \in \mathcal{C}'\}$;
5: **for** each $c \in \tilde{\mathcal{C}}$ **do**
6: $\quad w(c) \leftarrow w'(g_2^{-1}(c))$;
7: **end for**
8: 令 $g : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ 为复合映射 $g_2 \circ g_1$;
9: **for** each $i \in [\ell]$ **do**
10: $\quad \tilde{\mathcal{F}}_i \leftarrow \{g(f) : f \in \mathcal{F}_i\}$;
11: **end for**
12: **return** $g : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$, $(\ell, \{\tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_\ell\}, \tilde{\mathcal{C}}, w, k, \boldsymbol{r})$.

节提出的实例消减方法. 该算法首先基于引理 2 构造 $O(\log n)$-维欧氏空间, 然后在该空间中利用引理 4 中的核心集构造方法将实例的用户集合规模压缩为 $(k \log n)^{O(1)}$, 最后利用引理 3 将实例映射到 $O(\log k + \log\log n)$-维欧氏空间中. 以下引理给出了算法 2 的性能保证.

引理7 给定满足 $\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C} \subset \mathbb{R}^d$ 和 $|\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C}| = n$ 的多样性公平 k-中位问题实例 $\mathcal{I} = (\ell, \{\mathcal{F}_1, \dots, \mathcal{F}_\ell\}, \mathcal{C}, k, \boldsymbol{r})$ 和常数 $\epsilon \in (0, \frac{1}{2})$, 算法 2 可以在 $O(nd \log n + (k\epsilon^{-1}n)^{O(1)})$ 时间内构造满足 $\tilde{d} = \epsilon^{-O(1)}(\log k + \log\log n)$ 的映射 $g : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ 以及满足 $\bigcup_{i=1}^{\ell} \tilde{\mathcal{F}}_i \cup \tilde{\mathcal{C}} \subset \mathbb{R}^{\tilde{d}}$, $\sum_{c \in \tilde{\mathcal{C}}} w(c) = |\mathcal{C}|$ 和 $|\tilde{\mathcal{C}}| \leqslant (k\epsilon^{-1}\log n)^{O(1)}$ 的加权实例 $(\ell, \{\tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_\ell\}, \tilde{\mathcal{C}}, w, k, \boldsymbol{r})$, 使得任意规模不超过 k 的集合 $\mathcal{S} \subseteq \bigcup_{i=1}^{\ell} \mathcal{F}_i$ 都满足 $\sum_{c \in \tilde{\mathcal{C}}} w(c)\Delta(c, \{g(f) : f \in \mathcal{S}\}) \in [1-\epsilon, 1+5\epsilon) \sum_{c \in \mathcal{C}} \Delta(c, \mathcal{S})$.

证明 令 $g_1 : \mathbb{R}^d \to \mathbb{R}^{d'}$ 表示算法 2 在第 1 步中构造的映射, 令 $\mathcal{C}'$ 和 $w' : \mathcal{C}' \to [1, +\infty)$ 分别表示算法 2 在第 2 步构造的加权集合和权重函数, 令 $g_2 : \mathbb{R}^{d'} \to \mathbb{R}^{\tilde{d}}$ 表示算法 2 在第 3 步构造的映射, 并令 $g : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ 和 $(\ell, \{\tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_\ell\}, \tilde{\mathcal{C}}, w, k, \boldsymbol{r})$ 分别表示算法 2 返回的映射和加权实例. 可以得出

$$\sum_{c \in \tilde{\mathcal{C}}} w(c) = \sum_{c \in \tilde{\mathcal{C}}} w'(g_2^{-1}(c)) = \sum_{c \in \mathcal{C}'} w'(c) = |\{g_1(c) : c \in \mathcal{C}\}| = |\mathcal{C}|, \tag{5}$$

其中, 第 1 步基于算法 2 在第 6 步中的操作得出, 第 2 步基于 $\tilde{\mathcal{C}} = \{g_2(c) : c \in \mathcal{C}'\}$ 这一事实得出, 第 3 步基于引理 4 得出, 第 4 步基于引理 2 构造的映射为单映射这一假设得出. 此外, 由引理 2 和 4 可知,

$$d' = O(\epsilon^{-2} \log n), \tag{6}$$

且

$$|\tilde{\mathcal{C}}| = |\mathcal{C}'| \leqslant d'(k\epsilon^{-1}\log|\mathcal{C}|)^{O(1)} \leqslant (k\epsilon^{-1}\log n)^{O(1)}. \tag{7}$$

结合等式 (6) 和 (7) 以及引理 2∼4 可知, 算法 2 的时间复杂度不超过

$$O(nd\log n + |\mathcal{C}|d'k + |\mathcal{C}'|d'\log|\mathcal{C}'|) \leqslant O(nd\log n + (k\epsilon^{-1}n)^{O(1)}), \tag{8}$$

且

$$\tilde{d} = O(\epsilon^{-2}\log|\mathcal{C}'|) \leqslant O(\epsilon^{-2}\log(k\epsilon^{-1}\log n)) = \epsilon^{-O(1)}(\log k + \log\log n). \tag{9}$$

给定一个满足 $|\mathcal{S}| \leqslant k$ 的集合 $\mathcal{S} \subseteq \bigcup_{i=1}^{\ell} \mathcal{F}_i$, 可以得出

$$\begin{aligned}\sum_{c\in\tilde{\mathcal{C}}} w(c)\Delta(c,\{g(f): f\in\mathcal{S}\}) &\in [1,1+\epsilon]\sum_{c\in\mathcal{C}'} w(c)\Delta(c,\{g_1(f): f\in\mathcal{S}\})\\ &\subseteq [1-\epsilon,(1+\epsilon)^2]\sum_{c\in\mathcal{C}}\Delta(g_1(c),\{g_1(f): f\in\mathcal{S}\})\\ &\subseteq [1-\epsilon,(1+\epsilon)^3]\sum_{c\in\mathcal{C}}\Delta(c,\mathcal{S})\\ &\subseteq [1-\epsilon,1+5\epsilon)\sum_{c\in\mathcal{C}}\Delta(c,\mathcal{S}),\end{aligned} \tag{10}$$

其中, 第 1 步基于引理 3 得到, 第 2 步基于引理 4 得到, 第 3 步基于引理 2 得到, 第 4 步基于 $\epsilon \in (0, \frac{1}{2})$ 这一事实得出. 由式 (5)∼(10) 可知, 引理 7 正确.

4　小规模加权实例求解算法

本节通过构造设施集合的数据网求解多样性公平 k-中位问题在低维欧氏空间中的小规模加权实例. 下面给出本节中使用的一些定义. 令 $g:\mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ 和 $\tilde{\mathcal{I}} = (\ell, \{\tilde{\mathcal{F}}_1,\ldots,\tilde{\mathcal{F}}_\ell\}, \tilde{\mathcal{C}}, w, k, \boldsymbol{r})$ 分别表示基于算法 2 构造的映射和加权实例, 其中, $\bigcup_{i=1}^{\ell}\tilde{\mathcal{F}}_i \cup \tilde{\mathcal{C}} \subset \mathbb{R}^{\tilde{d}}$, 且 $\tilde{\mathcal{F}}_i = \{g(f): f\in\mathcal{F}_i\}\,\forall\, i\in[\ell]$. 令 $\tilde{\mathcal{F}} = \bigcup_{i=1}^{\ell}\tilde{\mathcal{F}}_i$. 由于 $\tilde{\mathcal{I}}$ 和 $\mathcal{I}$ 有相同的设施类别分布, 这两个实例的可行解存在性是一致的. 在 $\mathcal{I}$ 和 $\tilde{\mathcal{I}}$ 有可行解的情况下, 令 $\tilde{\mathcal{S}}^* = \{f_1^*,\ldots,f_k^*\} \subseteq \tilde{\mathcal{F}}$ 表示 $\tilde{\mathcal{I}}$ 的一个最优解, 并令 $\mathrm{opt} = \sum_{c\in\tilde{\mathcal{C}}} w(c)\Delta(c,\tilde{S}^*)$ 表示 $\tilde{\mathcal{S}}^*$ 的费用. 给定整数 $i\in[k]$, 定义 $\mathcal{L}_i^* = \{j\in[\ell]: f_i^*\in\tilde{\mathcal{F}}_j\}$, 并令 $\mathcal{H}_i = \bigcap_{j\in\mathcal{L}_i^*}\tilde{\mathcal{F}}_j$. 令 $\Delta_{\max} = \max_{c\in\tilde{\mathcal{C}}}\Delta(c,\tilde{\mathcal{S}}^*)$. 由 $w(c)\geqslant 1\,\forall\, c\in\tilde{\mathcal{C}}$ 这一事实可知,

$$\Delta_{\max} \leqslant \sum_{c\in\tilde{\mathcal{C}}} w(c)\Delta(c,\tilde{\mathcal{S}}^*) = \mathrm{opt}. \tag{11}$$

如第 2 节中所述, 本节以 $\tilde{\mathcal{S}}^*$ 中设施附近的用户为中心构造开设设施的环形搜索空间. 给定整数 $i\in[k]$ 和实数 $\alpha>0$, 令 c_i 表示集合 $\tilde{\mathcal{C}}$ 中与 f_i^* 距离最近的用户, 并定义 $\mathcal{B}_i(\alpha) = \{f\in\tilde{\mathcal{F}}: \Delta(c_i,f)\leqslant\alpha\}$. 给定整数 $i\in[k]$ 和 $j\in[\lceil\epsilon^{-2}\log n\rceil]$, 令 $\mathcal{D}(i,0) = \mathcal{B}_i(\epsilon\Delta_{\max}n^{-1})$, 并令 $\mathcal{D}(i,j) = \mathcal{B}_i(\epsilon(1+\epsilon)^j\Delta_{\max}n^{-1})\backslash\mathcal{B}_i(\epsilon(1+\epsilon)^{j-1}\Delta_{\max}n^{-1})$. $\Delta_{\max}$ 的定义说明

$$\epsilon(1+\epsilon)^{\lceil\epsilon^{-2}\log n\rceil}\Delta_{\max}n^{-1} > \Delta_{\max} \geqslant \max_{i\in[k]}\Delta(c_i,f_i^*). \tag{12}$$

由不等式 (12) 以及 $\mathcal{D}(i,j)$ 和 $\mathcal{H}_i$ 的定义可知, 给定任意整数 $i\in[k]$, 都存在满足 $f_i^*\in\mathcal{D}(i,j)\cap\mathcal{H}_i$ 的整数 $j\in[0,\lceil\epsilon^{-2}\log n\rceil]$. 令 $\mathcal{D}_i$ 表示上述包含 f_i^* 的集合 $\mathcal{D}(i,j)\cap\mathcal{H}_i$. 本节将开设设施的搜索范围限制在集合 $\mathcal{D}_1,\ldots,\mathcal{D}_k$ 的成员中. 以下引理说明, 通过将实例 $\tilde{\mathcal{I}}$ 求解算法的时间复杂度乘以 $2^{\ell k}(k\epsilon^{-1})^{O(k)}n^{O(1)}$, 可以假设 $\mathcal{D}_1,\ldots,\mathcal{D}_k$ 的成员是已知的.

引理8　给定加权实例 $(\ell,\{\tilde{\mathcal{F}}_1,\ldots,\tilde{\mathcal{F}}_\ell\},\tilde{\mathcal{C}},w,k,\boldsymbol{r})$ 和常数 $\epsilon\in(0,\frac{1}{2})$, 集合 $\{\mathcal{D}_1,\ldots,\mathcal{D}_k\}$ 的取值不超过 $2^{\ell k}(k\epsilon^{-1})^{O(k)}n^{O(1)}$ 种.

证明　我们首先分析集合 $\{\mathcal{D}(1,0)\cap\mathcal{H}_1,\ldots,\mathcal{D}(k,\lceil\epsilon^{-2}\log n\rceil)\cap\mathcal{H}_k\}$ 的取值. 由 $\mathcal{D}(i,j)$ 的定义可知, $\{\mathcal{D}(1,0)\cap\mathcal{H}_1,\ldots,\mathcal{D}(k,\lceil\epsilon^{-2}\log n\rceil)\cap\mathcal{H}_k\}$ 的取值取决于实数 $\Delta_{\max}$ 以及集合 $\{c_1,\ldots,c_k\}$ 和 $\{\mathcal{H}_1,\ldots,\mathcal{H}_k\}$ 的取值. 由 $\Delta_{\max}$, c_i 和 $\mathcal{H}_i$ 的定义可知: $\Delta_{\max}$ 的取值不超过 $|\tilde{\mathcal{F}}||\tilde{\mathcal{C}}|$ 种; $\{c_1,\ldots,c_k\}$ 的取值不超过 $|\tilde{\mathcal{C}}|^k$ 种; $\{\mathcal{H}_1,\ldots,\mathcal{H}_k\}$ 的取值不超过 $2^{\ell k}$ 种. 有上述论证可知, 集合 $\{\mathcal{D}(1,0)\cap\mathcal{H}_1,\ldots,\mathcal{D}(k,\lceil\epsilon^{-2}\log n\rceil)\cap\mathcal{H}_k\}$ 最多有 $2^{\ell k}|\tilde{\mathcal{C}}|^{k+1}|\tilde{\mathcal{F}}|$ 种取值.

由 $\mathcal{D}_i$ 的定义可知, 每个整数 $i\in[k]$ 都满足 $\mathcal{D}_i\in\{\mathcal{D}(i,0)\cap\mathcal{H}_i,\ldots,\mathcal{D}(i,\lceil\epsilon^{-2}\log n\rceil)\cap\mathcal{H}_i\}$. 结合这一事实与上文中关于 $\{\mathcal{D}(1,0)\cap\mathcal{H}_1,\ldots,\mathcal{D}(k,\lceil\epsilon^{-2}\log n\rceil)\cap\mathcal{H}_k\}$ 取值的分析可知, $\{\mathcal{D}_1,\ldots,\mathcal{D}_k\}$ 的取值

算法 3 加权实例求解算法.

输入: 常数 $\epsilon \in (0, \frac{1}{2})$ 和加权实例 $\tilde{\mathcal{I}} = (\ell, \{\tilde{\mathcal{F}}_1, \ldots, \tilde{\mathcal{F}}_\ell\}, \tilde{\mathcal{C}}, w, k, \boldsymbol{r})$;
输出: 实例 $\tilde{\mathcal{I}}$ 的近似解 $\mathcal{S}^\dagger$;

$\mathbb{S} \leftarrow \emptyset$;
令 $\mathbb{D}$ 为基于引理 8 为集合 $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$ 确定的取值范围;
for each $\{\mathcal{D}'_1, \ldots, \mathcal{D}'_k\} \in \mathbb{D}$ **do**
  for $j \leftarrow 1$ to k^k **do**
    $\mathcal{U} \leftarrow \emptyset$;
    for each $f \in \bigcup_{i=1}^{k} \mathcal{D}'_i$ **do**
      令 $\gamma(f)$ 为在 $[k]$ 中随机选取的一个整数;
    end for
    for $j \leftarrow 1$ to k **do**
      $\mathcal{D}'_i \leftarrow \{f \in \mathcal{D}'_i : \gamma(f) = i\}$;
      if $|\mathcal{D}'_i| > 1$ **then**
        基于引理 6 构造 $\mathcal{D}'_i$ 的 $\max_{x,y \in \mathcal{D}'_i} \epsilon\Delta(x, y)$-数据网 $\mathcal{N}_i$;
        $\mathcal{U} \leftarrow \mathcal{U} \cup \mathcal{N}_i$;
      else
        $\mathcal{U} \leftarrow \mathcal{U} \cup \mathcal{D}'_i$;
      end if
    end for
    for each $\mathcal{S} \subseteq \mathcal{U}$ **do**
      if $|\mathcal{S} \cap \tilde{\mathcal{F}}_i| \geqslant r_i \, \forall i \in [\ell]$ 且 $|\mathcal{S}| \leqslant k$ **then**
        $\mathbb{S} \leftarrow \mathbb{S} \cup \{\mathcal{S}\}$;
      end if
    end for
  end for
end for
return $\mathcal{S}^\dagger \leftarrow \arg\min_{\mathcal{S} \in \mathbb{S}} \sum_{c \in \tilde{\mathcal{C}}} w(c)\Delta(c, \mathcal{S})$.

不超过 $2^{\ell k}|\tilde{\mathcal{C}}|^{k+1}|\tilde{\mathcal{F}}|(\epsilon^{-2}\log n + 1)^k$ 种. 可以得出

$$\begin{aligned} 2^{\ell k}|\tilde{\mathcal{C}}|^{k+1}|\tilde{\mathcal{F}}|(\epsilon^{-2}\log n + 1)^k &\leqslant 2^{\ell k}|\tilde{\mathcal{F}}|(k\epsilon^{-1}\log n)^{O(k)} \\ &\leqslant 2^{\ell k}|\tilde{\mathcal{F}}|(k\epsilon^{-1})^{O(k)} n^{O(1)} \\ &= 2^{\ell k}(k\epsilon^{-1})^{O(k)} n^{O(1)}, \end{aligned} \tag{13}$$

其中, 第 1 步基于 $|\tilde{\mathcal{C}}| = (k\epsilon^{-1}\log n)^{O(1)}$ 这一事实 (引理 7) 得出, 第 2 步基于引理 1 得出, 第 3 步根据 $|\tilde{\mathcal{F}}| = |\mathcal{F}| < n$ 这一事实 (由 $\tilde{\mathcal{F}}$ 的定义得出) 得出. 不等式 (13) 说明引理 8 成立.

算法 3 是本节为加权实例 $\tilde{\mathcal{I}}$ 提出的求解算法. 该算法在引理 8 给出的取值范围中寻找集合 $\mathcal{D}_1, \ldots, \mathcal{D}_k$. 给定满足 $f_i \in \mathcal{D}_i \, \forall i \in [k]$ 且不包含重复元素的集合 $\{f_1, \ldots, f_k\} \subseteq \tilde{\mathcal{F}}$, 由 $\mathcal{D}_i \subseteq \mathcal{H}_i \, \forall i \in [k]$ 这一事实以及 $\mathcal{H}_i$ 的定义可知, 每个整数 $i \in [\ell]$ 都满足

$$|\{f_1, \ldots, f_k\} \cap \tilde{\mathcal{F}}_i| \geqslant |\tilde{\mathcal{S}}^* \cap \tilde{\mathcal{F}}_i| \geqslant r_i. \tag{14}$$

不等式 (14) 说明 $\{f_1, \ldots, f_k\}$ 是实例 $\tilde{\mathcal{I}}$ 的可行解. 该不等式为算法 3 构造满足实例公平性约束条件的解提供了明确的设施搜索范围. 对于每个整数 $i \in [k]$, 算法 3 在集合 $\mathcal{D}_i$ 中选取与 f_i^* 较为接近的设施作为候选开设设施. 然而, 当集合 $\mathcal{D}_1, \ldots, \mathcal{D}_k$ 之间存在交集时, 满足 $f_i \in \mathcal{D}_i \, \forall i \in [k]$ 的解 $\{f_1, \ldots, f_k\}$ 可能会因为包含重复元素而违反不等式 (14) 以及实例的公平性约束条件. 为了避免上述问题, 算法 3 在第 7 步和第 10 步中基于彩色编码 (color coding) 技术去除集合 $\mathcal{D}_1, \ldots, \mathcal{D}_k$ 之间的交集. 给定整数 $i \in [k]$, 算法 3 在 7 步为每个设施 $f \in \mathcal{D}_i$ 在 $[k]$ 内随机选取一个标签 $\gamma(f)$. 可以得出, 等式 $\gamma(f_i^*) = i \, \forall i \in [k]$ 成立的概率不低于 k^{-k}. 由于算法 3 在第 4 步的循环中重复执行 k^k 次, 该等式在

至少一次循环中成立的概率可以被提升为 $1-(1-k^{-k})^{k^k} > 1-\mathrm{e}^{-1}$. 在第 10 步中, 为了保证集合 $\mathcal{D}_1,\dots,\mathcal{D}_k$ 之间不存在交集, 算法 3 从 $\mathcal{D}_i$ 中移除标签不为 i 的设施. 在等式 $\gamma(f_i^*)=i\,\forall\, i\in[k]$ 成立的情况下, 算法 3 的第 10 步操作不会影响 $f_i^*\in\mathcal{D}_i\,\forall\, i\in[k]$ 这一事实. 在去除集合 $\mathcal{D}_1,\dots,\mathcal{D}_k$ 之间的交集以后, 算法 3 为规模大于 1 的每个集合构造数据网, 并将数据网中的设施添加到候选开设设施集合 $\mathcal{U}$ 中. 最后, 算法 3 基于候选开设设施集合构造候选解集合 $\mathbb{S}$, 并返回 $\mathbb{S}$ 中费用最低的解. 令 $\mathcal{S}^\dagger$ 表示算法 3 返回的解.

以下引理给出了算法 3 的时间复杂度.

引理9　算法 3 的时间复杂度为 $2^{\ell k+(k\epsilon^{-1})^{O(1)}}n^{O(1)}$.

证明　令 $\mathbb{D}$ 表示算法 3 在第 2 步中基于引理 8 为集合 $\{\mathcal{D}_1,\dots,\mathcal{D}_k\}$ 确定的取值范围. 给定 $\tilde{\mathcal{F}}$ 的 k 个子集 $\mathcal{D}_1',\dots,\mathcal{D}_k'$, 令 $\mathcal{U}'$ 表示算法 3 在第 5∼17 步中构造的候选开设设施集合. 令 $\mathcal{J}=\{i\in[k]:|\mathcal{D}_i'|>1\}$. 给定整数 $i\in\mathcal{J}$, 算法 3 根据 $\mathcal{D}_i'$ 中设施之间的最大距离构造 $\max_{x,y\in\mathcal{D}_i'}\epsilon\Delta(x,y)$-数据网 $\mathcal{N}_i'$, 并将 $\mathcal{N}_i'$ 中的设施添加到 $\mathcal{U}'$ 中. 由引理 6 可知, 该过程所需时间不超过

$$\sum_{i\in\mathcal{J}}2^{O(\tilde{d})}|\mathcal{D}_i'|^{O(1)}\leqslant 2^{O(\tilde{d})}k|\tilde{\mathcal{F}}|^{O(1)}<2^{O(\tilde{d})}kn^{O(1)}. \tag{15}$$

此外, 由 $\mathcal{N}_i'$ 的定义和引理 6 可知, 每个整数 $i\in\mathcal{J}$ 都满足

$$|\mathcal{N}_i'|\leqslant\left(\frac{\max_{x,y\in\mathcal{D}_i'}\Delta(x,y)}{\epsilon\max_{x,y\in\mathcal{D}_i'}\Delta(x,y)}\right)^{\tilde{d}}=\epsilon^{-\tilde{d}}. \tag{16}$$

结合不等式 (16) 与 $\mathcal{U}'$ 的构造过程可知

$$|\mathcal{U}'|=\sum_{i\in\mathcal{J}}|\mathcal{N}_i'|+k-|\mathcal{J}|\leqslant k\epsilon^{-\tilde{d}}. \tag{17}$$

算法 3 在第 18∼22 步的循环中枚举 $\mathcal{U}'$ 的规模不超过 k 的子集, 并将其中的可行解添加到候选解集合 $\mathbb{S}$ 中. 由算法 3 将第 5∼22 步中的操作重复执行 $|\mathbb{D}|k^k$ 次这一事实可知

$$|\mathbb{S}|\leqslant|\mathbb{D}|k^k|\mathcal{U}'|^k\leqslant|\mathbb{D}|k^{O(k)}\epsilon^{-\tilde{d}k}, \tag{18}$$

其中, 第 2 步基于不等式 (17) 得出. 此外, 不等式 (15) 和 (17) 说明构造 $\mathbb{S}$ 所需时间不超过

$$|\mathbb{D}|k^k(2^{O(\tilde{d})}kn^{O(1)}+|\mathcal{U}'|^k)\leqslant|\mathbb{D}|k^{O(k)}(2^{O(\tilde{d})}n^{O(1)}+\epsilon^{-\tilde{d}k}). \tag{19}$$

构造候选解集合 $\mathbb{S}$ 以后, 算法 3 在 $\mathbb{S}$ 中花费 $|\mathbb{S}||\tilde{\mathcal{C}}|\tilde{d}k\leqslant|\mathbb{D}|k^{O(k)}\epsilon^{-\tilde{d}k}|\tilde{\mathcal{C}}|d$ (不等式 (18)) 时间寻找费用最小的解. 结合这一事实与不等式 (19) 可知, 算法 3 的时间复杂度为

$$\begin{aligned}|\mathbb{D}|k^{O(k)}(2^{O(\tilde{d})}n^{O(1)}+\epsilon^{-\tilde{d}k}|\tilde{\mathcal{C}}|d)&\leqslant 2^{\ell k}(k\epsilon^{-1})^{O(k)}n^{O(1)}\epsilon^{-O(\tilde{d}k)}\\&\leqslant 2^{\ell k}(k\epsilon^{-1})^{O(k)}n^{O(1)}(k\log n)^{(k\epsilon^{-1})^{O(1)}}\\&\leqslant 2^{\ell k}(k\epsilon^{-1})^{(k\epsilon^{-1})^{O(1)}}n^{O(1)}\\&=2^{\ell k+(k\epsilon^{-1})^{O(1)}}n^{O(1)},\end{aligned} \tag{20}$$

其中, 第 1 步基于不等式 $|\mathbb{D}|\leqslant 2^{\ell k}(k\epsilon^{-1})^{O(k)}n^{O(1)}$ (引理 8) 和 $|\tilde{\mathcal{C}}|\leqslant(\epsilon^{-1}k\log n)^{O(1)}$ (引理 7) 得出, 第 2 步基于等式 $\tilde{d}=\epsilon^{-O(1)}(\log k+\log\log n)$ (引理 7) 得出, 第 3 步基于引理 1 得出. 不等式 (20) 说明引理 9 成立.

以下引理说明, $\mathcal{S}^\dagger$ 有较高的常数概率是实例 $\tilde{\mathcal{I}}$ 的 $(1+5\epsilon)$-近似解.

引理10　不等式 $\sum_{c\in\tilde{\mathcal{C}}}w(c)\Delta(c,\mathcal{S}^\dagger)<(1+5\epsilon)\mathrm{opt}$ 成立的概率不低于 $1-\mathrm{e}^{-1}$.

证明　如前文所述, 以下事件成立的概率不低于 $1-\mathrm{e}^{-1}$: 算法 3 能基于彩色编码技术构造 k 个互不相交的设施集合 $\mathcal{D}_1^*,\dots,\mathcal{D}_k^*$, 使得每个整数 $i\in[k]$ 都满足 $\mathcal{D}_i^*\subseteq\mathcal{D}_i$ 和 $f_i^*\in\mathcal{D}_i^*$. 令 $\mathcal{U}$ 表示算法 3 在第 9∼17 步中基于 $\mathcal{D}_1^*,\dots,\mathcal{D}_k^*$ 构造的候选开设设施集合. 给定整数 $i\in[k]$, 本文分别考虑以下 3 种情况: (1) $|\mathcal{D}_i^*|=1$; (2) $|\mathcal{D}_i^*|>1$ 且 $\mathcal{D}_i^*\subseteq\mathcal{D}(i,0)$; (3) $|\mathcal{D}_i^*|>1$ 且存在满足 $\mathcal{D}_i^*\subseteq\mathcal{D}(i,j)$ 的整数 $j\in[\lceil\epsilon^{-2}\log n\rceil]$.

在情况 (1) 中, 算法 3 在第 15 步将 $\mathcal{D}_i^*$ 中的设施添加至 $\mathcal{U}$ 中. 由 $f_i^*\in\mathcal{D}_i^*$ 这一事实可知, 等式

$$\{f_i^*\}=\mathcal{D}_i^*\subseteq\mathcal{U} \tag{21}$$

在情况 (1) 中成立.

下面分析 $|\mathcal{D}_i^*|>1$ 的情况. 在该情况中, 算法 3 在第 13 步将 $\mathcal{D}_i^*$ 的 $\max_{x,y\in\mathcal{D}_i^*}\epsilon\Delta(x,y)$-数据网 $\mathcal{N}_i$ 中的设施添加到 $\mathcal{U}$ 中. $\mathcal{N}_i\subseteq\mathcal{D}_i^*$ 这一事实说明 $\mathcal{D}_i^*\cap\mathcal{U}\neq\emptyset$. 在情况 (2) 中, 令 f_i 为 $\mathcal{D}_i^*\cap\mathcal{U}$ 中的一个设施. 可以得出

$$\Delta(f_i,f_i^*)\leqslant\Delta(c_i,f_i)+\Delta(c_i,f_i^*)\leqslant 2\max_{f\in\mathcal{D}_i^*}\Delta(c_i,f)\leqslant\frac{2\epsilon}{n}\Delta_{\max}\leqslant\frac{2\epsilon}{n}\mathrm{opt}, \tag{22}$$

其中, 第 1 步基于三角不等式[1)]得到, 第 2 步基于 $\{f_i,f_i^*\}\subseteq\mathcal{D}_i^*$ 这一事实得出, 第 3 步根据 $\mathcal{D}_i^*\subseteq\mathcal{D}(i,0)$ 这一假设以及 $\mathcal{D}(i,0)$ 的定义得出, 第 4 步基于不等式 (11) 得出.

在情况 (3) 中, 令 f_i 为 $\mathcal{N}_i$ 中与 f_i^* 距离最近的设施. 可以得出

$$\Delta(f_i,f_i^*)\leqslant\epsilon\max_{x,y\in\mathcal{D}_i^*}\Delta(x,y)\leqslant 2\epsilon\max_{f\in\mathcal{D}_i^*}\Delta(c_i,f)\leqslant 2\epsilon(1+\epsilon)\Delta(c_i,f_i^*), \tag{23}$$

其中, 第 1 步根据 $\mathcal{N}_i$ 是 $\mathcal{D}_i^*$ 的 $\max_{x,y\in\mathcal{D}_i^*}\epsilon\Delta(x,y)$-数据网这一事实以及数据网的定义得出, 第 2 步根据三角不等式得出, 第 3 步基于存在满足 $\mathcal{D}_i^*\subseteq\mathcal{D}(i,j)$ 的整数 $j\in[\lceil\epsilon^{-2}\log n\rceil]$ 这一假设以及 $\mathcal{D}(i,j)$ 的定义得出.

由等式 (21)、不等式 (22) 和 (23) 可知, 候选开设设施集合 $\mathcal{U}$ 有一个对于任意整数 $i\in[k]$ 都满足以下不等式的子集 $\mathcal{S}=\{f_1,\dots,f_k\}$:

$$\Delta(f_i,f_i^*)\leqslant\frac{2\epsilon}{n}\mathrm{opt}+2\epsilon(1+\epsilon)\Delta(c_i,f_i^*). \tag{24}$$

给定整数 $i\in[k]$, 定义 $\tilde{\mathcal{C}}_i^*=\{c\in\tilde{\mathcal{C}}:\arg\min_{f\in\tilde{\mathcal{S}}^*}\Delta(c,f)=f_i^*\}$. 可以得出

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{c\in\tilde{\mathcal{C}}_i^*}w(c)\Delta(f_i,f_i^*)&\leqslant\sum_{i=1}^{k}\sum_{c\in\tilde{\mathcal{C}}_i^*}w(c)\left(\frac{2\epsilon}{n}\mathrm{opt}+2\epsilon(1+\epsilon)\Delta(c_i,f_i^*)\right)\\
&<2\epsilon\cdot\mathrm{opt}+2\epsilon(1+\epsilon)\sum_{i=1}^{k}\sum_{c\in\tilde{\mathcal{C}}_i^*}w(c)\Delta(c_i,f_i^*)\\
&\leqslant 2\epsilon\cdot\mathrm{opt}+2\epsilon(1+\epsilon)\sum_{i=1}^{k}\sum_{c\in\tilde{\mathcal{C}}_i^*}w(c)\Delta(c,f_i^*)\\
&=2\epsilon\cdot\mathrm{opt}+2\epsilon(1+\epsilon)\sum_{c\in\tilde{\mathcal{C}}}w(c)\Delta(c,\tilde{\mathcal{S}}^*)\\
&<5\epsilon\cdot\mathrm{opt},
\end{aligned} \tag{25}$$

其中, 第 1 步基于不等式 (24) 得出, 第 2 步基于 $\sum_{c\in\tilde{\mathcal{C}}}w(c)=|\mathcal{C}|<n$ 这一事实 (引理 7) 得出, 第 3 步基于 c_i 是 $\tilde{\mathcal{C}}$ 中与 f_i^* 距离最近的用户这一事实得出, 第 4 步根据 $\tilde{\mathcal{C}}_i^*$ 的定义得出, 第 5 步根据 $\epsilon\in(0,\frac{1}{2})$

1) 欧氏空间中的任意 3 个点 x, y 和 z 都满足 $\Delta(x,z)\leqslant\Delta(x,y)+\Delta(y,z)$.

算法 4 多样性公平 k-中位问题的求解算法.

输入: 常数 $\epsilon \in (0, \frac{1}{2})$ 以及满足 $\bigcup_{i=1}^{\ell} \mathcal{F}_i \cup \mathcal{C} \subset \mathbb{R}^d$ 的实例 $\mathcal{I} = (\ell, \{\mathcal{F}_1, \dots, \mathcal{F}_\ell\}, \mathcal{C}, k, \boldsymbol{r})$;
输出: 布尔值 `False` 或实例 $\mathcal{I}$ 的近似解 $\mathcal{S}^{\ddagger}$;
1: 令 `bool` 为算法 1 为实例 $\mathcal{I}$ 给出的判定结果;
2: **if** `bool` = `False` **then**
3:　　**return** `False`;
4: **else**
5:　　令 $g : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ 和 $\tilde{\mathcal{I}} = (\ell, \{\tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_\ell\}, \tilde{\mathcal{C}}, w, k, \boldsymbol{r})$ 为基于算法 2 为 ϵ 和 $\mathcal{I}$ 构造的映射和加权实例;
6:　　令 $\mathcal{S}^{\dagger}$ 为基于算法 3 为 ϵ 和 $\tilde{\mathcal{I}}$ 构造的集合;
7:　　**return** $\mathcal{S}^{\ddagger} \leftarrow \{g^{-1}(f) : f \in \mathcal{S}^{\dagger}\}$;
8: **end if**

这一事实得出. 由此可知,

$$\begin{aligned}
\sum_{c \in \tilde{\mathcal{C}}} w(c)\Delta(c, \mathcal{S}) &= \sum_{i=1}^{k} \sum_{c \in \tilde{\mathcal{C}}_i^*} w(c)\Delta(c, \mathcal{S}) \\
&\leqslant \sum_{i=1}^{k} \sum_{c \in \tilde{\mathcal{C}}_i^*} w(c)\big(\Delta(c, f_i^*) + \Delta(f_i^*, \mathcal{S})\big) \\
&\leqslant \sum_{i=1}^{k} \sum_{c \in \tilde{\mathcal{C}}_i^*} w(c)\big(\Delta(c, f_i^*) + \Delta(f_i^*, f_i)\big) \\
&= \sum_{c \in \tilde{\mathcal{C}}} w(c)\Delta(c, \tilde{\mathcal{S}}^*) + \sum_{i=1}^{k} \sum_{c \in \tilde{\mathcal{C}}_i^*} w(c)\Delta(f_i^*, f_i) \\
&< (1 + 5\epsilon)\mathrm{opt}, \qquad (26)
\end{aligned}$$

其中, 第 2 步基于三角不等式得到, 第 4 步根据 $\tilde{\mathcal{C}}_i^*$ 的定义得出, 第 5 步根据不等式 (25) 得出.

由 $f_i \in \mathcal{D}_i^* \, \forall i \in [k]$ 这一事实和 $\mathcal{D}_1^*, \dots, \mathcal{D}_k^*$ 的定义可知, 集合 $\mathcal{S} = \{f_1, \dots, f_k\}$ 中不包含重复元素, 且 $f_i \in \mathcal{D}_i \, \forall i \in [k]$. 结合这一事实与不等式 (14) 可知, $\mathcal{S}$ 是实例 $\tilde{\mathcal{I}}$ 的可行解, 且算法 3 会在第 18∼22 步中将 $\mathcal{S}$ 添加到候选解集合 $\mathbb{S}$ 中. 由此可知,

$$\sum_{c \in \tilde{\mathcal{C}}} w(c)\Delta(c, \mathcal{S}^{\dagger}) = \min_{\mathcal{S}' \in \mathbb{S}} \sum_{c \in \tilde{\mathcal{C}}} w(c)\Delta(c, \mathcal{S}') \leqslant \sum_{c \in \tilde{\mathcal{C}}} w(c)\Delta(c, \mathcal{S}) < (1 + 5\epsilon)\mathrm{opt},$$

其中, 第 3 步基于不等式 (26) 得到.

5　多样性公平 k-中位问题的求解算法

本节基于第 3 节中给出的实例消减方法和第 4 节中给出的小规模加权实例求解算法为实例 $\mathcal{I}$ 构造近似解, 如算法 4 所述. 给定常数 ϵ 和实例 $\mathcal{I}$, 算法 4 首先基于算法 1 判定实例 $\mathcal{I}$ 的可行解存在性. 在实例 $\mathcal{I}$ 有可行解的情况下, 算法 4 利用算法 2 构造映射 $g : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ 以及满足 $\bigcup_{i=1}^{\ell} \tilde{\mathcal{F}}_i \cup \tilde{\mathcal{C}} \subset \mathbb{R}^{\tilde{d}}$ 和 $\tilde{\mathcal{F}}_i = \{g(f) : f \in \mathcal{F}_i\} \, \forall i \in [\ell]$ 的加权实例 $\tilde{\mathcal{I}} = (\ell, \{\tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_\ell\}, \tilde{\mathcal{C}}, w, k, \boldsymbol{r})$, 并基于算法 3 求解加权实例 $\tilde{\mathcal{I}}$. 该算法将 $\tilde{\mathcal{I}}$ 的解在映射 g 下的原像集合作为实例 $\mathcal{I}$ 的近似解. 下面通过分析算法 4 的性能证明定理 1 的正确性.

证明　(定理 1) 由引理 5 可知, 算法 4 调用算法 1 以判定实例 $\mathcal{I}$ 可行解存在性所需时间不超过 $O(2^{\ell k} nk)$. 在实例 $\mathcal{I}$ 有可行解的情况下, 令 $\mathcal{S}^{\ddagger}$ 表示算法 4 返回的设施集合, 并令 $\mathcal{S}^{\dagger}$ 表示算法 4 调用算法 3 为加权实例 $\tilde{\mathcal{I}}$ 构造的近似解. 可以得出, 每个整数 $i \in [\ell]$ 都满足

$$|\mathcal{S}^{\ddagger} \cap \mathcal{F}_i| = |\{g(f) : f \in \mathcal{S}^{\ddagger} \cap \mathcal{F}_i\}| = |\mathcal{S}^{\dagger} \cap \tilde{\mathcal{F}}_i| \geqslant r_i, \qquad (27)$$

其中, 第 1 步根据引理 2 和 3 构造的映射都是单映射这一假设得出, 第 2 步基于 $\mathcal{S}^\dagger = \{g(f) : f \in \mathcal{S}^\ddagger\}$ 和 $\tilde{\mathcal{F}}_i = \{g(f) : f \in \mathcal{F}_i\}$ 这一事实得出, 第 3 步根据 $\mathcal{S}^\dagger$ 是 $\tilde{\mathcal{I}}$ 的可行解这一事实得出. 不等式 (27) 说明 $\mathcal{S}^\ddagger$ 是实例 $\mathcal{I}$ 的可行解. 下面分析 $\mathcal{S}^\ddagger$ 的近似比. 令 $\mathcal{S}^*$ 和 $\tilde{\mathcal{S}}^*$ 分别表示实例 $\mathcal{I}$ 和加权实例 $\tilde{\mathcal{I}}$ 的最优解. 在引理 10 中声明的不等式成立的情况下, 可以得出

$$\begin{aligned}\sum_{c\in\mathcal{C}}\Delta(c,\mathcal{S}^\ddagger) &\leqslant \frac{1}{1-\epsilon}\sum_{c\in\tilde{\mathcal{C}}} w(c)\Delta(c,\mathcal{S}^\dagger)\\ &< \frac{1+5\epsilon}{1-\epsilon}\sum_{c\in\tilde{\mathcal{C}}} w(c)\Delta(c,\tilde{\mathcal{S}}^*)\\ &\leqslant \frac{1+5\epsilon}{1-\epsilon}\sum_{c\in\tilde{\mathcal{C}}} w(c)\Delta(c,\{g(f) : f\in\mathcal{S}^*\})\\ &< \frac{(1+5\epsilon)^2}{1-\epsilon}\sum_{c\in\mathcal{C}}\Delta(c,\mathcal{S}^*)\\ &< (1+47\epsilon)\sum_{c\in\mathcal{C}}\Delta(c,\mathcal{S}^*),\end{aligned} \tag{28}$$

其中, 第 1 和 4 步基于引理 7 得出, 第 2 步基于引理 10 得出, 第 5 步基于 $\epsilon \in (0, \frac{1}{2})$ 这一事实得出. 由不等式 (28) 可知, 当引理 10 中声明的不等式成立时 (其概率不低于 $1-\mathrm{e}^{-1}$), 算法 4 返回的解近似比为 $(1+47\epsilon)$. 通过重复运行算法 4 并返回费用最低的解, 我们可以将得到 $(1+47\epsilon)$-近似解的概率提升为任意小于 1 的常数. 此外, 引理 7 和 9 说明算法 4 的时间复杂度为 $O(nd\log n)+2^{\ell k+(k\epsilon^{-1})^{O(1)}}n^{O(1)}$. 令 $\epsilon = \frac{\varepsilon}{47}$, 则上述论证表明, 算法 4 是时间复杂度为 $O(nd\log n)+2^{\ell k+(k\varepsilon^{-1})^{O(1)}}n^{O(1)}$ 的 $(1+\varepsilon)$-近似算法.

6　总结

本文为多样性公平 k-中位问题提出了时间复杂度为 $O(nd\log n)+2^{\ell k+(k\varepsilon^{-1})^{O(1)}}n^{O(1)}$ 的 $(1+\varepsilon)$-近似算法, 在欧氏空间中改进了 Thejaswi 等 [8] 在相近的固定参数时间内给出的 $(1+2\mathrm{e}^{-1}+\varepsilon)$-近似结果. 本文给出的实例消减技术和基于数据网的开设设施选取方法为求解欧氏空间中的设施选址型问题实例提供了新的求解思路. 基于这一技术处理其他相关问题是一个值得进一步研究的方向.

参考文献

1 Gowda K N, Pensyl T W, Srinivasan A, et al. Improved bi-point rounding algorithms and a golden barrier for k-median. In: Proceedings of the 34th ACM-SIAM Symposium on Discrete Algorithms, 2023. 987–1011

2 Cohen-Addad V, Esfandiari H, Mirrokni V S, et al. Improved approximations for Euclidean k-means and k-median, via nested quasi-independent sets. In: Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, 2022. 1621–1628

3 Cohen-Addad V, Klein P N, Mathieu C. Local search yields approximation schemes for k-means and k-median in euclidean and minor-free metrics. SIAM J Comput, 2019, 48: 644–667

4 Friggstad Z, Rezapour M, Salavatipour M R. Local search yields a PTAS for k-means in doubling metrics. SIAM J Comput, 2019, 48: 452–480

5 Cohen-Addad V, Feldmann A E, Saulpic D. Near-linear time approximation schemes for clustering in doubling metrics. J ACM, 2021, 68: 1–34

6 Kay M, Matuszek C, Munson S A. Unequal representation and gender stereotypes in image search results for occupations. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015. 3819–3828

7 Thejaswi S, Ordozgoiti B, Gionis A. Diversity-aware k-median: clustering with fair center representation. In: Proceedings of the 32nd European Conference on Machine Learning and the 25th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2021. 765–780

8 Thejaswi S, Gadekar A, Ordozgoiti B, et al. Clustering with fair-center representation: parameterized approximation algorithms and heuristics. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022. 1749–1759

9 Zhang Z, Yang J F, Liu L M, et al. Towards a theoretical understanding of why local search works for clustering with fair-center representation. In: Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence, 2024. 16953–16960

10 Hotegni S S, Mahabadi S, Vakilian A. Approximation algorithms for fair range clustering. In: Proceedings of the 40th International Conference on Machine Learning, 2023. 13270–13284

11 Cohen-Addad V, Gupta A, Kumar A, et al. Tight FPT approximations for k-median and k-means. In: Proceedings of the 46th International Colloquium on Automata, Languages, and Programming, 2019

12 Manurangsi P, Raghavendra P. A birthday repetition theorem and complexity of approximating dense CSPs. In: Proceedings of the 44th International Colloquium on Automata, Languages, and Programming, 2017

13 Chen K. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. SIAM J Comput, 2009, 39: 923–947

14 Kumar A, Sabharwal Y, Sen S. Linear-time approximation schemes for clustering problems in any dimensions. J ACM, 2010, 57: 1–32

15 Jaiswal R, Kumar A, Sen S. A simple D 2-sampling based PTAS for k-means and other clustering problems. Algorithmica, 2014, 70: 22–46

16 Jaiswal R, Kumar M, Yadav P. Improved analysis of D^2-sampling based PTAS for k-means and other clustering problems. Inf Process Lett, 2015, 115: 100–103

17 Johnson W B, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space. Contemp Math, 1984, 26: 189–206

18 Narayanan S, Nelson J. Optimal terminal dimensionality reduction in Euclidean space. In: Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, 2019. 1064–1069

19 Bandyapadhyay S, Fomin F V, Simonov K. On coresets for fair clustering in metric and Euclidean spaces and their applications. In: Proceedings of the 48th International Colloquium on Automata, Languages, and Programming, 2021

20 Chen X R, Han L, Xu D C, et al. k-median/means with outliers revisited: a simple FPT approximation. In: Proceedings of the 29th International Conference on Computing and Combinatorics, 2023. 295–302

21 Kong X, Zhang Z, Feng Q. On parameterized approximation algorithms for balanced clustering. J Comb Optim, 2023, 45: 49

22 Agrawal A, Inamdar T, Saurabh S, et al. Clustering what matters: optimal approximation for clustering with outliers. J Artif Intell Res, 2023, 78: 143–166

23 Gupta A, Krauthgamer R, Lee J R. Bounded geometries, fractals, and low-distortion embeddings. In: Proceedings of the 44th IEEE Annual Symposium on Foundations of Computer Science, 2003. 534–543

24 Har-Peled S, Mendel M. Fast construction of nets in low-dimensional metrics and their applications. SIAM J Comput, 2006, 35: 1148–1184

A $(1+\varepsilon)$-approximation algorithm for diversity-aware k-median

Zhen ZHANG[1,2,3], Xiaohong CHEN[1,2*], Limei LIU[1,2*], Jian REN[1,2], Lin JIANG[1,2] & Qilong FENG[4,2]

1. *School of Advanced Interdisciplinary Studies, Hunan University of Technology and Business, Changsha 410205, China*;
2. *Xiangjiang Laboratory, Changsha 410205, China*;
3. *Business School, Central South University, Changsha 410083, China*;
4. *School of Computer Science, Central South University, Changsha 410083, China*
* Corresponding author. E-mail: csu_cxh@163.com, seagullm@163.com

Abstract The diversity-aware k-median problem holds significant importance across various fields related to clustering, particularly in scenarios where clustering centers need to be fairly selected, such as in data summarization. Given a set of clients, ℓ sets of facilities, and a positive integer k, the problem aims to open a constrained number of facilities from each facility set, ensuring that the number of opened facilities is no more than k, and each client has a high similarity with the nearest opened facility. In this paper, we reduce the considered instance of the problem to a small-size one located in a low-dimensional space, and partition the space based on the points from the reduced instance to estimate the locations of the facilities opened in an optimal solution. This yields a $(1+\varepsilon)$-approximation algorithm running in $O(nd\log n)+2^{\ell k+(k\varepsilon^{-1})^{O(1)}}n^{O(1)}$ time in d-dimensional Euclidean space, where n is the number of clients and facilities. This improves upon the previously known $(1+2\mathrm{e}^{-1}+\varepsilon)$-approximation ratio obtained within a similar fixed-parameter tractable running time, despite the latter being applicable in a more general metric space.

Keywords parameterized algorithm, approximation algorithm, k-median, facility location, sampling