中国科学 : 信息科学 2026 年 第 56 卷 第 5 期: 1195–1205
SCIENTIA SINICA Informationis
论文

极小 – 极大 (k,z)- 聚类问题的参数化近似

张震[1,2], 袁依格[2*], 武迪[3], 田志平[1], 冯启龙[2,4]

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. 长沙理工大学计算机学院, 长沙 410076
4. 中南大学计算机学院, 长沙 410083
* 通信作者. E-mail: immyyuan23@163.com

收稿日期: 2025–08–24; 修回日期: 2025–11–13; 接受日期: 2025–12–24; 网络出版日期: 2026–04–30

湘江实验室开放课题 (批准号: 25XJ03009)、国家自然科学基金 (批准号: 62202161, 62432016) 和湖南省科技创新计划 (批准号: 2025RC3207) 资助项目

摘要 给定度量空间中的一个点集和一个候选中心集以及正整数 k 和 z, 极小 – 极大 (k,z)- 聚类问题要求选取最多 k 个聚类中心, 在点集中移除最多 z 个异常点, 并将剩余点划分为 k 个簇, 从而最小化各簇聚类代价中的最大值, 其中, 每个簇的聚类代价为簇内点与对应中心间的距离之和. 极小 – 极大 (k,z)- 聚类问题在对负载均衡要求较高的应用中发挥重要作用, 但同时也面临较高的计算复杂性: 在多项式时间内, 目前已知的最佳近似算法仅能保证 $O(k)$- 近似比; 而在参数化设置下, 即使在同时以 k 和 z 为参数且中心集已给定的情况下, 该问题也不具备固定参数可解性. 鉴于此, 本文研究该问题的参数化近似, 提出了时间复杂度为 $2^{\tilde{O}(kz\varepsilon^{-1})}n^{O(1)}$ 的 $(3+\varepsilon)$- 近似算法. 这是关于该问题的首个常数近似结果. 当输入实例位于欧氏空间时, 应用相同方法可得出时间复杂度相近的 $(1+\varepsilon)$- 近似算法.

关键词 近似算法, 参数算法, 归约, 聚类, 采样

1 引言

基于中心的聚类 (center-based clustering) 是机器学习领域中的一类重要问题, 其目标是确定若干聚类中心, 并将每个数据点分配至一个距离较近的中心, 从而将数据集划分为若干具有高度内聚性的簇. 求解该类问题能有效刻画数据分布特征并识别具有代表性的数据原型, 因此, 相关算法已成为揭示数据内在结构、提取关键信息的重要工具.

基于中心的聚类问题多以所有点的聚类代价总和作为优化目标. 例如, k- 中位 (k-median) 问题要求最小化所有点与其对应中心之间的距离之和, 而 k- 平均 (k-means) 问题要求最小化距离的平方和. 由于其定义简洁且易于计算, 这类基于代价总和的目标函数在聚类问题的理论研究与实际应用中均得到了广泛关注. 然而, 根据此类目标函数定义的聚类问题在簇间聚类代价分布上缺乏均衡性约束: 即使在整体聚类代价较低的情况下, 个别簇的局部代价仍可能显著高于其他簇. 在资源调度、物流规划、任务分配等对负载均衡具有显著要求的应用中,

引用格式: 张震, 袁依格, 武迪, 等. 极小 – 极大 (k,z)- 聚类问题的参数化近似. 中国科学: 信息科学, 2026, 56: 1195–1205, doi: 10.1360/SSI-2025-0367
Zhang Z, Yuan Y G, Wu D, et al. Parameterized approximations for the minimax (k,z)-clustering problem. Sci Sin Inform, 2026, 56: 1195–1205, doi: 10.1360/SSI-2025-0367

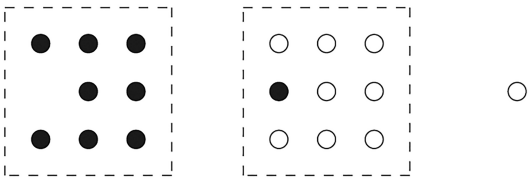

图 1　异常点对聚类结果的影响.

Figure 1　Effect of outliers on clustering.

这种簇间代价分布不均衡的现象可能导致系统性能的严重失衡. 例如, 在多服务器任务调度中, 若部分簇承担的计算或传输负载远高于其他簇, 容易出现局部资源过载、响应延迟增加等问题 [1,2]; 在物流与配送规划中, 若个别中心对应的运输代价显著偏高, 则会引发整体效率下降及资源利用不均等问题 [3,4]. 对于这些应用, 最小化聚类代价总和的方法并不适用, 更合理的优化目标应关注各簇聚类代价中的最大值, 以规避因簇间代价悬殊引起的风险. 鉴于此, 人们提出了极小 – 极大 k- 聚类 (minimax k-clustering, MkC) 问题 [5,6]. 给定度量空间中的点集 $\mathcal{P}$ 和候选中心集 $\mathcal{C}$ 以及正整数 k, MkC 问题要求在 $\mathcal{C}$ 中选取 k 个中心 $c_1,\cdots,c_k$ 并将 $\mathcal{P}$ 划分为 k 个簇 $\mathcal{P}_1,\cdots,\mathcal{P}_k$, 以最小化目标函数 $\max_{i\in[1,k]}\sum_{p\in\mathcal{P}_i}\delta(p,c_i)$, 其中, $\delta(p,c_i)$ 为 p 与 c_i 之间的距离.

MkC 问题根据点与对应中心之间的距离定义目标函数, 其求解过程对远离主体数据分布的异常点 (outliers) 具有固有敏感性. 在噪声干扰显著的数据环境下, 不加区分地将所有点纳入聚类过程往往会影响对主体数据分布特性的刻画, 如图 1 所示 (其中, 同一个簇内的点具有相同颜色). 针对这一挑战的常用策略是剔除一定数量的异常点, 以最小化剩余点在聚类问题中的目标函数值 [7~9]. 本文研究 MkC 问题在异常点剔除机制下的扩展形式, 即极小 – 极大 (k,z)- 聚类 (minimax (k,z)-clustering, MkzC) 问题, 其定义如下.

定义1 (MkzC 问题)　MkzC 问题的一个实例 $\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$ 包含以 δ 为距离函数、定义在集合 $\mathcal{X}$ 上的度量空间 $(\mathcal{X},\delta)$、点集 $\mathcal{P}\subseteq\mathcal{X}$、候选中心集 $\mathcal{C}\subseteq\mathcal{X}$ 以及整数 $k\in[1,|\mathcal{C}|]$ 和 $z\in[0,|\mathcal{P}|]$. 该实例的一个可行解 $(\mathcal{S},\mathcal{O},\tau)$ 包含规模不超过 k 的中心集 $\mathcal{S}\subseteq\mathcal{C}$、规模不超过 z 的异常点集 $\mathcal{O}\subseteq\mathcal{P}$ 以及映射 $\tau:\mathcal{P}\backslash\mathcal{O}\to\mathcal{S}$. 解 $(\mathcal{S},\mathcal{O},\tau)$ 的费用为 $\max_{c\in\mathcal{S}}\sum_{p\in\tau^{-1}(c)}\delta(p,c)$. M$kz$C 问题的目标是找到费用最低的可行解.

基于异常点剔除机制提升聚类分析的鲁棒性一直是机器学习和组合优化领域的热门研究方向. 人们已经利用该机制扩展了 k- 中位、k- 平均、拟阵中位 (matroid median) 等诸多以最小化聚类代价总和为目标的聚类问题, 并提出了一系列常数近似算法 [10~14]. 然而, 当目标函数从聚类代价总和转变为各簇聚类代价中的最大值时, 该类问题的求解难度明显提升. 这一点在以下方面得到了体现.

(1) 当以聚类代价总和作为目标函数时, 最优解中的异常点集由与最近中心距离最远的 z 个点构成. 然而, 在以各簇聚类代价中的最大值为目标函数的 MkzC 问题中, 异常点不再具备这一明确的分布特性. 这明显增加了异常点选择的复杂性.

(2) 给定 k 个中心和 z 个异常点, 我们可以在剔除异常点后通过将剩余的每个点分配到与其距离最近的中心来最小化聚类代价总和; 而在 MkzC 问题中, 从数值划分 (number partitioning) 以及最小化最大完成时间 (makespan minimization) 等问题出发构造的归约表明, 即使是根据给定的中心和异常点确定点集的最优分配方式也是 NP- 难的 [5,6].

目前, 关于 MkzC 问题的最佳近似结果仍源于通过最小化聚类代价总和来间接控制各簇聚类代价最大值的思路. 由于聚类代价总和的最小值不超过任意划分方式中各簇聚类代价最大值的 k 倍, 基于该思路构造的解具有 $O(k)$- 近似比 [5,6]. 如何挖掘 MkzC 问题本身的组合结构特性以设计近似比优于 $O(k)$ 的近似算法还是未知的.

2　本文贡献

在聚类问题的实际应用场景中, 可选择的中心数量 k 以及可剔除的异常点数量 z 往往远小于输入数据的总体规模 n. 鉴于此, 已有大量研究致力于在 k 和 z 取值较小的设定下改进相关聚类问题的近似结果 [14~19].

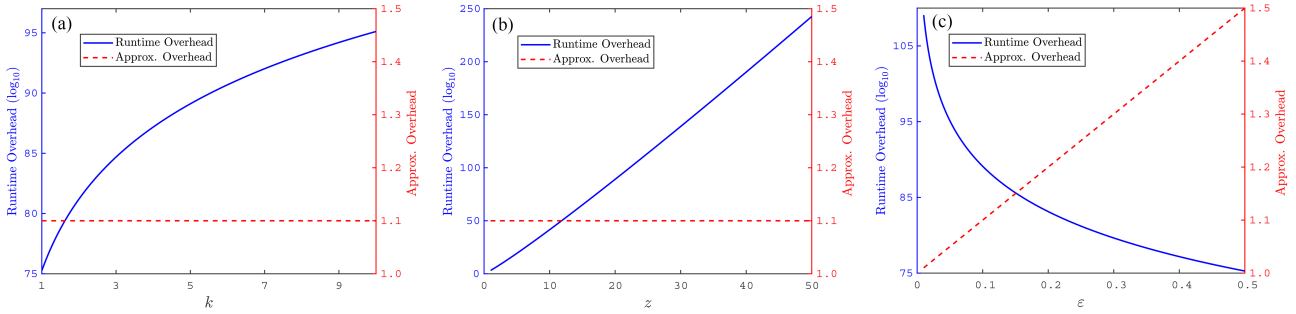

图 2　(网络版彩图) 参数取值对时间复杂度与近似比乘法放大因子的影响. (a) 不同 k 值 ($z=20$, $\varepsilon=0.1$); (b) 不同 z 值 ($k=5$, $\varepsilon=0.1$); (c) 不同 ε 值 ($k=5$, $z=20$).

Figure 2　(Color online) Effects of parameters on the multiplicative overheads in the running time and the approximation ratio. (a) Varying k ($z=20$, $\varepsilon=0.1$); (b) varying z ($k=5$, $\varepsilon=0.1$); (c) varying ε ($k=5$, $z=20$).

本文利用相同设定求解 MkzC 问题. 虽然可以通过枚举所有可能的中心集以及点集划分方式在 $n^{O(k)}$ 时间内找到 MkzC 问题实例的最优解, 但本文旨在避免对实例组合空间的全局搜索, 为 MkzC 问题提出固定参数时间 (即 $f(k,z)\cdot n^{O(1)}$ 时间, 其中, $f()$ 为任意可计算函数) 的求解算法.

Bandyapadhyay 等 [6] 从数值划分问题出发构造的归约表明, 即便最优解中的中心集已明确且实例中的点分布在线性度量空间中, MkC 问题在参数 k 下仍不具备固定参数可解性, 除非 P=NP. 由于 MkC 问题是 MkzC 问题在 $z=0$ 时的特例, 这同时意味着我们无法在 $f(k,z)\cdot n^{O(1)}$ 时间内精确求解 MkzC 问题. 然而, 该结论并未排除在固定参数时间内改进 MkzC 问题近似比的可能性. 例如, 在同样不具备固定参数可解性、要求剔除 z 个异常点的 k- 中位问题中 (其目标函数为移除异常点后剩余点与对应中心的距离之和), Agrawal 等 [14] 提出了时间复杂度为 $f(k,z,\varepsilon)\cdot n^{O(1)}$、近似比为 $1+2e^{-1}+\varepsilon$ 的参数化近似算法; 这一近似结果明显优于当前多项式时间内最佳的 $(6.994+\varepsilon)$- 近似比 [13]. 本文对 MkzC 问题开展类似研究, 在 $f(k,z,\varepsilon)\cdot n^{O(1)}$ 时间内将现有的 $O(k)$- 近似比改进为常数, 并进一步在欧氏空间中提出了具有相似时间复杂度、近似比为 $1+\varepsilon$ 的参数化近似算法.

本文为 MkzC 问题提出的参数化近似算法通过随机采样寻找实例最优解中异常点的邻近点, 并通过剔除这些点将实例归约为等价的 MkC 问题实例. 以下定理给出了该归约过程的理论保证.

定理1　令 ε 为 $(0,1)$ 内的任意常数, $T(n)$ 为 MkC 问题的任意 α- 近似算法在规模为 n 的实例中所需的运行时间, 则 (1) 给定一般度量空间中的 MkzC 问题实例 $\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$, 存在时间复杂度为 $\rho(T(|\mathcal{P}\cup\mathcal{C}|-z)+O(|\mathcal{P}|z))+O(|\mathcal{P}|kz\varepsilon^{-1})+(|\mathcal{P}|\cdot|\mathcal{C}|)^{O(1)}$、近似比为 $\alpha(1+\varepsilon)$ 的随机近似算法; (2) 给定欧氏空间中的 MkzC 问题实例 $\big((\mathbb{R}^d,||\cdot||_2),\mathcal{P},\mathbb{R}^d,k,z\big)$, 存在时间复杂度为 $\rho(T(|\mathcal{P}|-z)+O(|\mathcal{P}|dz))+O(|\mathcal{P}|dkz\varepsilon^{-1})+|\mathcal{P}|^{O(1)}d$、近似比为 $\alpha(1+\varepsilon)$ 的随机近似算法. 其中, ρ 在一般度量空间和欧氏空间中分别不超过 $\big(e(10.452kz\varepsilon^{-1}+k+2z)\big)^z$ 和 $\big(e(9.624kz\varepsilon^{-1}+k+2z)\big)^z$.

定理 1 表明, 对于 MkC 问题的任意多项式时间或参数化近似算法, 在近似比和时间复杂度分别放大 $1+\varepsilon$ 和 $(kz\varepsilon^{-1})^{O(z)}$ 倍的条件下, 可将其转换为求解 MkzC 问题的参数化近似算法. 图 2 给出了 k, z 和 ε 的取值在度量空间中对上述乘法放大因子的影响. 基于定理 1 与 Bandyapadhyay 等 [6] 为 MkC 问题提出的时间复杂度为 $2^{\tilde{O}(k\varepsilon^{-1})}n^{O(1)}$ 的 $(3+\varepsilon)$- 近似算法, 可以得出近似比为 $3+\varepsilon$、时间复杂度为 $2^{\tilde{O}(kz\varepsilon^{-1})}n^{O(1)}$ 的 MkzC 问题求解算法, 如推论 1 所述.

推论1　给定常数 $\varepsilon\in(0,1)$ 和一般度量空间中的 MkzC 问题实例 $\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$, 存在时间复杂度为 $2^{\tilde{O}(kz\varepsilon^{-1})}n^{O(1)}$、近似比为 $3+\varepsilon$ 的随机近似算法, 其中, $n=|\mathcal{P}\cup\mathcal{C}|$.

对于 d- 维欧氏空间中的 MkC 问题实例, Bandyapadhyay 等 [6] 提出了时间复杂度为 $2^{\tilde{O}(k\varepsilon^{-O(1)})}n^{O(1)}d$ 的 $(1+\varepsilon)$- 近似算法. 通过结合该算法与定理 1, 我们可以在欧氏空间中得出 MkzC 问题的参数化 $(1+\varepsilon)$- 近似算法, 如推论 2 所述.

推论2　给定常数 $\varepsilon\in(0,1)$ 和欧氏空间中的 MkzC 问题实例 $\big((\mathbb{R}^d,||\cdot||_2),\mathcal{P},\mathbb{R}^d,k,z\big)$, 存在时间复杂度为

$2^{\tilde{O}(kz\varepsilon^{-O(1)})}n^{O(1)}d$、近似比为 $1+\varepsilon$ 的随机近似算法, 其中, $n=|\mathcal{P}|$.

3　相关工作

参数计算理论是处理难解问题的重要工具, 其应用价值源于问题实例的某些关键参数在实际场景中的取值通常远小于整体数据规模的现象. 参数计算理论根据实例的 “小参数” 特性构建参数化问题模型, 并致力于设计时间复杂度为 $f(\kappa)\cdot n^{O(1)}$ 的参数化算法, 其中, κ 为被选定的参数, $f()$ 为任意可计算函数. 若某一问题可在该形式的时间复杂度内被精确求解, 则称该问题在相应参数下具有固定参数可解性. 如前文所述, 已有复杂性结果 [6] 表明 M$k$$z$C 问题在参数 k 和 z 下不具备固定参数可解性. 本文据此设计 M$k$$z$C 问题的参数化近似算法.

当 $z=0$ 时, M$k$$z$C 问题特化为 M$k$C 问题. 尽管在此特例下无需处理异常点, 但与一般情形相似, 目前在多项式时间内关于 MkC 问题的最佳近似结果依然是通过最小化聚类代价总和得出的 $O(k)$- 近似比 [5,6]. 在允许实例的解适度违反聚类中心数量约束的情况下, Even 等 [20] 和 Arkin 等 [21] 为 MkC 问题提出了选取 $O(k)$ 个中心的 (双标准) 常数近似算法. 对于 MkC 问题在一维欧氏空间中的实例, Ahmadian 等 [5] 提出了多项式时间的 $(1+\varepsilon)$- 近似算法. 此外, 如第 2 节所述, Bandyapadhyay 等 [6] 以 k 为参数, 分别在一般度量空间和高维欧氏空间中提出了固定参数时间的 $(3+\varepsilon)$- 近似算法和 $(1+\varepsilon)$- 近似算法.

本文通过结合异常点识别方法与 MkC 问题的近似算法求解 M$k$$z$C 问题, 如定理 1 所述. 这一思路与 Agrawal 等 [14] 以及 Jaiswal 和 Kumar [15] 针对剔除 z 个异常点的 k- 中位问题所提出的求解框架较为相似. Agrawal 等 [14] 和 Jaiswal 和 Kumar [15] 分别利用严格限制采样范围的均匀采样方法与根据点的位置设定其采样概率的非均匀采样方法剔除异常点, 并在剩余点中求解 k- 中位问题, 从而提出了以 k 和 z 为参数的参数化近似算法. 本文同样致力于在固定参数时间内确定异常点集, 但 M$k$$z$C 问题在解的组合结构上呈现更高的复杂性: 正如第 1 节中所指出的, 与最小化所有簇聚类代价之和的情形不同, M$k$$z$C 问题最优解中的异常点缺乏可供算法直接利用的分布规律; 即使在中心集已知的前提下, 能否以非穷举的方式在有限时间内确定这些异常点也仍是未知的. 为此, 本文针对 M$k$$z$C 问题提出了新的二阶段异常点识别方法. 该方法在第一阶段松弛了对最优异常点准确定位的要求, 转而寻找与其位置相近的 “锚点”; 第二阶段则利用这些锚点引导异常点的选取过程, 从而构造近似最优的异常点集. 本文利用这一方法实现了从 M$k$$z$C 问题向 M$k$C 问题的有效归约, 并基于此为前者设计了参数化近似算法.

4　M$k$$z$C 问题的求解算法

本节证明定理 1 的正确性. 首先给出本节所用的基本定义和引理. 令 ϵ 为 $(0,1)$ 内的一个常数. 令 $\mathcal{I}=\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$ 表示 M$k$$z$C 问题的一个实例. 给定点 $x\in\mathcal{X}$ 和集合 $\mathcal{Y}\subseteq\mathcal{X}$, 令 $\delta(x,\mathcal{Y})=\min_{y\in\mathcal{Y}}\delta(x,y)$ 为 x 与 $\mathcal{Y}$ 中最近邻的距离. 给定整数 $i\geqslant 1$, 令 $[i]=[1,i]\cap\mathbb{Z}$ 为从 1 到 i 的整数集合. 令 $(\mathcal{S}^*,\mathcal{O}^*,\tau^*)$ 为实例 $\mathcal{I}$ 的一个最优解, 其中, $\mathcal{S}^*=\{c_1^*,\cdots,c_k^*\}$. 给定整数 $i\in[k]$, 令 $\mathcal{P}_i^*=(\tau^*)^{-1}(c_i^*)$ 为中心 c_i^* 对应的簇. 令 $\mathrm{opt}=\max_{i\in[k]}\sum_{p\in\mathcal{P}_i^*}\delta(p,c_i^*)$ 为实例 $\mathcal{I}$ 最优解的费用. 给定正整数 k' 和子集 $\mathcal{P}'\subseteq\mathcal{P}$, 令 $\mathrm{opt}_{k'}^{\mathrm{sum}}(\mathcal{P}')=\min_{\mathcal{S}\subseteq\mathcal{C}\wedge|\mathcal{S}|=k'}\sum_{p\in\mathcal{P}'}\delta(p,\mathcal{S})$ 表示利用 $\mathcal{C}$ 中的 k' 个中心对 $\mathcal{P}'$ 中的点进行聚类时所能达到的最小聚类代价总和.

k- 中位问题要求最小化所有点与其对应中心之间的距离之和, 其定义如下. 在基于锚点的求解过程中, 本节通过构造相应的 k- 中位问题实例最小化所选锚点与对应异常点之间的距离之和.

定义2 (k- 中位问题)　k- 中位问题的一个实例 $\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k\big)$ 包含以 δ 为距离函数、定义在集合 $\mathcal{X}$ 上的度量空间 $(\mathcal{X},\delta)$、点集 $\mathcal{P}\subseteq\mathcal{X}$、候选中心集 $\mathcal{C}\subseteq\mathcal{X}$ 以及整数 $k\in[|\mathcal{C}|]$. 该实例的一个可行解是一个规模不超过 k 的中心集 $\mathcal{S}\subseteq\mathcal{C}$, 其费用为 $\sum_{p\in\mathcal{P}}\delta(p,\mathcal{S})$. k- 中位问题的目标是找到费用最低的可行解.

以下引理是关于 k- 中位问题的常数近似保证.

引理1 (参见文献 [22∼24])　对于 k- 中位问题在一般度量空间中的实例 $\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k\big)$ 以及在欧氏空间中的实例 $\big((\mathbb{R}^d,||\cdot||_2),\mathcal{P},\mathbb{R}^d,k\big)$, 分别存在时间复杂度为 $(|\mathcal{P}|\cdot|\mathcal{C}|)^{O(1)}$ 的 2.613- 近似算法和时间复杂度为 $|\mathcal{P}|^{O(1)}d$

的 2.406- 近似算法.

以下引理给出了独立的二元随机变量之和显著低于其期望值的概率上界.

引理2 (切尔诺夫界 [25])　给定实数 $p,\lambda\in(0,1)$ 以及 t 个取值为 0 或 1 的二元变量 $v_1,\cdots,v_t$, 其中, 每个整数 $i\in[t]$ 都满足 $\Pr[v_i=1]=p$, 不等式 $\sum_{i=1}^{t}v_i<(1-\lambda)pt$ 成立的概率低于 $e^{-\lambda^2pt/2}$.

下面给出 MkzC 问题的求解算法. 具体而言, 第 4.1 小节基于随机采样方法选取用于定位异常点的锚点; 第 4.2 小节围绕这些锚点构造近似最优异常点集; 第 4.3 小节通过移除所选异常点将实例 $\mathcal{I}$ 归约为 MkC 问题的实例, 并利用该归约完成求解.

4.1　锚点选取算法

算法 1 描述了本节所使用的锚点选取算法 Anchor-Sel. 给定常数 $\epsilon\in(0,1)$ 以及 MkzC 问题实例 $\mathcal{I}=\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$, 算法 Anchor-Sel 首先基于引理 1 求解 $(k+z)$- 中位问题实例 $\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k+z\big)$ 以选取 $k+z$ 个中心. 由于该中心集具有 β- 近似比, $\mathcal{P}$ 中所有点与其最近中心之间的距离之和不超过 $\beta\cdot\mathrm{opt}_{k+z}^{\mathrm{sum}}$. 鉴于此, 算法 Anchor-Sel 将这 $k+z$ 个中心作为用于定位异常点的初始锚点集. 为了进一步提升锚点集对最优异常点集 $\mathcal{O}^*$ 的覆盖能力, 确保 $\mathcal{O}^*$ 中的每个异常点都与某个锚点足够接近, 该算法迭代地在 $\mathcal{P}$ 中额外选取 $\lceil 2kz\beta\epsilon^{-1}\rceil$ 个锚点, 其中, 每个点被选取的概率正比于其与当前最近锚点间的距离.

算法 1 Anchor-Sel$(\epsilon,\mathcal{I})$.

输入: 常数 $\epsilon\in(0,1)$ 以及 MkzC 问题的实例 $\mathcal{I}=\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$;
输出: 锚点集 $\mathcal{A}\subseteq\mathcal{P}\cup\mathcal{C}$;
1: 基于引理 1 为 $(k+z)$- 中位问题实例 $\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k+z\big)$ 构造近似解 $\mathcal{A}$, 令 β 为其近似比;
2: **for** $i\in[\lceil 2kz\beta\epsilon^{-1}\rceil]$ **do**
3:　以概率 $\delta(p,\mathcal{A})/\sum_{q\in\mathcal{P}}\delta(q,\mathcal{A})$ 在 $\mathcal{P}$ 中选取一个点 p;
4:　$\mathcal{A}\leftarrow\mathcal{A}\cup\{p\}$;
5: **end for**
6: **return** $\mathcal{A}$.

令 $\mathcal{A}$ 为 Anchor-Sel$(\epsilon,\mathcal{I})$ 返回的锚点集. 以下引理给出了该算法的性能保证.

引理3　不等式 $\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A})\leqslant\epsilon\cdot\mathrm{opt}$ 成立的概率不低于 $1-e^{-1/4}$.

证明　令 $\mathcal{A}_0$ 表示算法 Anchor-Sel 在第 1 步通过求解 $(k+z)$- 中位问题实例得出的中心集. 给定整数 $i\in[\lceil 2kz\beta\epsilon^{-1}\rceil]$, 令 $\mathcal{A}_i$ 表示算法 Anchor-Sel 在前 i 次迭代中选取的锚点与 $\mathcal{A}_0$ 的并集. 为了证明引理 3, 分别分析以下两种情况:

(1) 存在某个整数 $i\in[0,\lceil 2kz\beta\epsilon^{-1}\rceil-1]$, 使得

$$\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A}_i)\leqslant\frac{\epsilon}{k\beta}\sum_{p\in\mathcal{P}}\delta(p,\mathcal{A}_i);$$

(2) 每个整数 $i\in[0,\lceil 2kz\beta\epsilon^{-1}\rceil-1]$ 都满足

$$\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A}_i)>\frac{\epsilon}{k\beta}\sum_{p\in\mathcal{P}}\delta(p,\mathcal{A}_i).$$

在情况 1 中, 存在某个整数 $i\in[0,\lceil 2kz\beta\epsilon^{-1}\rceil-1]$, 使得

$$\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A})\leqslant\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A}_i)\leqslant\frac{\epsilon}{k\beta}\sum_{p\in\mathcal{P}}\delta(p,\mathcal{A}_i)\leqslant\frac{\epsilon}{k\beta}\sum_{p\in\mathcal{P}}\delta(p,\mathcal{A}_0),\tag{1}$$

其中, 第 1 步和第 3 步源于包含关系 $\mathcal{A}_0\subseteq\mathcal{A}_i\subseteq\mathcal{A}$. 此外,

$$\sum_{p\in\mathcal{P}}\delta(p,\mathcal{A}_0)\leqslant\beta\cdot\mathrm{opt}_{k+z}^{\mathrm{sum}}(\mathcal{P})\leqslant\beta\cdot\mathrm{opt}_{k}^{\mathrm{sum}}(\mathcal{P}\backslash\mathcal{O}^*)\leqslant\beta\sum_{i=1}^{k}\sum_{p\in\mathcal{P}_i^*}\delta(p,c_i^*)\leqslant k\beta\cdot\mathrm{opt},\tag{2}$$

其中, 第 1 步基于 $\mathcal{A}_0$ 是实例 $\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k+z\big)$ 的 β- 近似解这一事实得出, 第 2 步将 $\mathcal{O}^*$ 中的 z 个异常点视为 z 个单元素簇, 第 3 步根据 $\mathrm{opt}_k^{\mathrm{sum}}(\mathcal{P}\backslash\mathcal{O}^*)$ 的定义得出, 第 4 步由等式 $\mathrm{opt}=\max_{i\in[k]}\sum_{p\in\mathcal{P}_i^*}\delta(p,c_i^*)$ 得出. 由不等式 (1) 和 (2) 可知, 引理 3 中声明的不等式在情况 1 中成立.

下面分析情况 2. 给定整数 $i\in[0,\lceil 2kz\beta\epsilon^{-1}\rceil-1]$, 由算法 `Anchor-Sel` 在第 $i+1$ 次迭代中为每个点 $p\in\mathcal{P}$ 设定的正比于 $\delta(p,\mathcal{A}_i)$ 的采样概率可知,

$$\Pr[|\mathcal{A}_{i+1}\cap\mathcal{O}^*|>|\mathcal{A}_i\cap\mathcal{O}^*|]=\frac{\sum_{o\in\mathcal{O}^*\backslash\mathcal{A}_i}d(o,\mathcal{A}_i)}{\sum_{p\in\mathcal{P}\backslash\mathcal{A}_i}d(p,\mathcal{A}_i)}=\frac{\sum_{o\in\mathcal{O}^*}d(o,\mathcal{A}_i)}{\sum_{p\in\mathcal{P}}d(p,\mathcal{A}_i)}>\frac{\epsilon}{k\beta}. \tag{3}$$

令 $v_0,\cdots,v_{\lceil 2kz\beta\epsilon^{-1}\rceil-1}$ 为一组取值为 0 或 1 的独立随机变量, 其中, 每个变量 v_i 都满足 $\Pr[v_i=1]=\epsilon(k\beta)^{-1}$. 可以得出

$$\Pr\left[\mathcal{O}^*\nsubseteq\mathcal{A}\right]=\Pr[|\mathcal{A}\cap\mathcal{O}^*|<|\mathcal{O}^*|]<\Pr\left[\sum_{i=0}^{\lceil 2kz\beta\epsilon^{-1}\rceil-1}v_i<z\right]<e^{-1/4}, \tag{4}$$

其中, 第 2 步源于不等式 (3) 和变量 v_i 取值为 1 的概率, 第 3 步根据引理 2 得出. 不等式 (4) 说明, $\mathcal{O}^*\subseteq\mathcal{A}$ 且 $\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A})=0$ 成立的概率不低于 $1-e^{-1/4}$. 由此可知, 引理 3 中声明的不等式在情况 2 中成立.

4.2　基于锚点的异常点选取算法

本节利用 `Anchor-Sel`$(\epsilon,\mathcal{I})$ 返回的锚点集 $\mathcal{A}$ 构造候选异常点集族, 如算法 2 (`Outlier-Sel`) 所述. 给定常数 $\epsilon\in(0,1)$、实例 $\mathcal{I}=\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$ 以及锚点集 $\mathcal{A}$, 算法 `Outlier-Sel` 遍历所有满足 $\sum_{a\in\mathcal{A}}\gamma(a)=z$ 的映射 $\gamma:\mathcal{A}\to\mathbb{Z}_{\geqslant 0}$ 以估计每个锚点在最优异常点集 $\mathcal{O}^*$ 中所对应的异常点数量. 在每次枚举中, 算法 `Outlier-Sel` 为每个锚点选取相应数量的邻近点作为候选异常点. 在此过程中, 为避免 $\mathcal{P}$ 中的某些点因靠近多个锚点而被重复选作异常点, 算法 `Outlier-Sel` 构造随机映射 $\ell:\mathcal{P}\to\mathcal{A}$, 并仅在每个锚点在该映射下的原像中选取其对应的异常点.

算法 2 `Outlier-Sel`$(\epsilon,\mathcal{I},\mathcal{A})$.

输入: 常数 $\epsilon\in(0,1)$, `MkzC` 问题的实例 $\mathcal{I}=\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$, 以及锚点集 $\mathcal{A}\subseteq\mathcal{P}\cup\mathcal{C}$;
输出: 候选异常点集族 $\mathbb{O}$;

```
𝕆 ← ∅;
for γ : 𝒜 → ℤ_{⩾0} s.t. ∑_{a∈𝒜} γ(a) = z do
    令 ℓ : 𝒫 → {a ∈ 𝒜 | γ(a) ≠ 0} 为一个均匀随机映射;
    if 不等式 |ℓ^{-1}(a)| ⩾ γ(a) 对于满足 γ(a) ≠ 0 的每个锚点 a ∈ 𝒜 都成立 then
        𝒪 ← ∅;
        for a ∈ 𝒜 s.t. γ(a) ≠ 0 do
            令 NN(a) 为 a 在 ℓ^{-1}(a) 中的 γ(a) 个最近邻组成的集合;
            𝒪 ← 𝒪 ∪ NN(a);
        end for
        𝕆 ← 𝕆 ∪ {𝒪};
    end if
end for
return 𝕆.
```

令 $\mathbb{O}$ 为 `Outlier-Sel`$(\epsilon,\mathcal{I},\mathcal{A})$ 返回的候选异常点集族. 以下引理给出了 $\mathbb{O}$ 包含与 $\mathcal{O}^*$ 足够接近的候选集的概率下界.

引理4　以下事件成立的概率不低于 z^{-z}: 存在候选集 $\mathcal{O}\in\mathbb{O}$ 及双射 $\varphi:\mathcal{O}^*\backslash\mathcal{O}\to\mathcal{O}\backslash\mathcal{O}^*$, 使得 $\sum_{o\in\mathcal{O}^*\backslash\mathcal{O}}\delta(o,\varphi(o))\leqslant 2\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A})$.

证明　给定锚点 $a\in\mathcal{A}$, 令 $\mathcal{O}^*(a)=\{o\in\mathcal{O}^*\mid\arg\min_{a'\in\mathcal{A}}\delta(o,a')=a\}$ 表示该锚点在 $\mathcal{O}^*$ 中对应的异常点子集, 并令 $\gamma^*(a)=|\mathcal{O}^*(a)|$. 令 $\mathcal{A}^*=\{a\in\mathcal{A}\mid\gamma^*(a)\neq 0\}$ 表示在 $\mathcal{O}^*$ 中对应的异常点数量不为 0 的锚点子集.

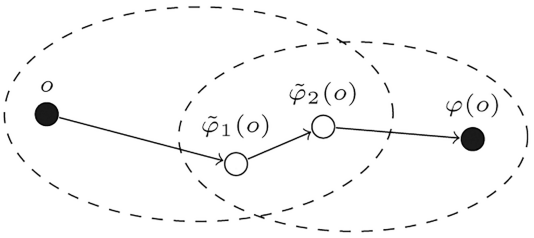

图 3　映射 $\varphi:\mathcal{O}^*\backslash\mathcal{O}\to\mathcal{O}\backslash\mathcal{O}^*$ 的构造过程.

Figure 3　Construction of the mapping $\varphi:\mathcal{O}^*\backslash\mathcal{O}\to\mathcal{O}\backslash\mathcal{O}^*$.

给定映射 $\gamma^*:\mathcal{A}\to\mathbb{Z}_{\geqslant 0}$, 算法 `Outlier-Sel` 在第 3 步为每个点 $p\in\mathcal{P}$ 随机分配一个锚点 $\ell(p)\in\mathcal{A}^*$. 由不等式 $|\mathcal{A}^*|\leqslant z$ 可以得出, 包含关系

$$\mathcal{O}^*(a)\subseteq\ell^{-1}(a)\,\forall\, a\in\mathcal{A}^* \tag{5}$$

成立的概率不低于 z^{-z}. 给定锚点 $a\in\mathcal{A}^*$, 令 $\mathrm{NN}(a)\subseteq\ell^{-1}(a)$ 表示算法 `Outlier-Sel` 在包含关系 (5) 成立的情况下在第 7 步为锚点 $a\in\mathcal{A}^*$ 选取的 $\gamma^*(a)$- 最近邻集合; 等式 $|\mathrm{NN}(a)|=\gamma^*(a)=|\mathcal{O}^*(a)|$ 说明可构造双射 $\tilde{\varphi}:\mathcal{O}^*(a)\to\mathrm{NN}(a)$. 可以得出, 映射 $\tilde{\varphi}:\mathcal{O}^*\to\bigcup_{a\in\mathcal{A}^*}\mathrm{NN}(a)$ 满足

$$\begin{aligned}\sum_{o\in\mathcal{O}^*}\delta(o,\tilde{\varphi}(o))&=\sum_{a\in\mathcal{A}^*}\sum_{o\in\mathcal{O}^*(a)}\delta(o,\tilde{\varphi}(o))\\&\leqslant\sum_{a\in\mathcal{A}^*}\sum_{o\in\mathcal{O}^*(a)}(\delta(o,a)+\delta(\tilde{\varphi}(o),a))\\&=\sum_{a\in\mathcal{A}^*}\sum_{o\in\mathcal{O}^*(a)}\delta(o,a)+\sum_{a\in\mathcal{A}^*}\sum_{o\in\mathrm{NN}(a)}\delta(o,a)\\&\leqslant 2\sum_{a\in\mathcal{A}^*}\sum_{o\in\mathcal{O}^*(a)}\delta(o,a)\\&=2\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A})\end{aligned} \tag{6}$$

的概率不低于 z^{-z} (这一概率下界等同于包含关系 (5) 成立的概率下界), 其中, 第 2 步由三角不等式得出, 第 3 步基于 $\tilde{\varphi}:\mathcal{O}^*(a)\to\mathrm{NN}(a)$ 的双射性得出, 第 4 步源于假设 $\mathcal{O}^*(a)\subseteq\ell^{-1}(a)$ (基于包含关系 (5) 提出) 以及 $\mathrm{NN}(a)$ 是锚点 a 在 $\ell^{-1}(a)$ 中的最近邻集合这一事实得出.

令 $\mathcal{O}=\bigcup_{a\in\mathcal{A}^*}\mathrm{NN}(a)$ 表示算法 `Outlier-Sel` 在包含关系 (5) 成立的情况下围绕 $\mathcal{A}^*$ 中的锚点选取的异常点集. 由于每个锚点 $a\in\mathcal{A}^*$ 都满足 $\mathrm{NN}(a)\subseteq\ell^{-1}(a)$ (由算法 `Outlier-Sel` 的第 7 步得出), 等式 $\bigcup_{a\in\mathcal{A}^*}\mathrm{NN}(a)=\biguplus_{a\in\mathcal{A}^*}\mathrm{NN}(a)$ 成立. 结合该等式与

$$|\mathcal{O}|=\sum_{a\in\mathcal{A}}\gamma^*(a)=z=|\mathcal{O}^*|$$

这一事实可知, $\tilde{\varphi}:\mathcal{O}^*\to\mathcal{O}$ 是一个双射. 下面基于该双射构造引理 4 中声明的双射 $\varphi:\mathcal{O}^*\backslash\mathcal{O}\to\mathcal{O}\backslash\mathcal{O}^*$. 我们通过迭代调用 $\tilde{\varphi}$ 将每个异常点 $o\in\mathcal{O}^*\backslash\mathcal{O}$ 映射到 $\mathcal{O}\backslash\mathcal{O}^*$ 中, 如图 3 所示 (其中, 白色点位于 $\mathcal{O}^*\cap\mathcal{O}$, 箭头表示 $\tilde{\varphi}:\mathcal{O}^*\to\mathcal{O}$ 的映射方向). 具体而言, 给定异常点 $o\in\mathcal{O}^*\backslash\mathcal{O}$, 我们按照以下方式定义一个序列 $\tilde{\varphi}_0(o),\tilde{\varphi}_1(o),\cdots$:

(1) 令 $\tilde{\varphi}_0(o)=o$;

(2) 给定正整数 i, 若 $\tilde{\varphi}_{i-1}(o)\in\mathcal{O}^*$, 则令 $\tilde{\varphi}_i(o)=\tilde{\varphi}(\tilde{\varphi}_{i-1}(o))$, 否则 (即 $\tilde{\varphi}_{i-1}(o)\in\mathcal{O}\backslash\mathcal{O}^*$) 令序列在 $i-1$ 处终止, 并记 $t(o)=i-1$.

由等式 $|\mathcal{O}^*\backslash\mathcal{O}|=|\mathcal{O}\backslash\mathcal{O}^*|$ 和 $\tilde{\varphi}:\mathcal{O}^*\to\mathcal{O}$ 的双射性可知, 对于每个异常点 $o\in\mathcal{O}^*\backslash\mathcal{O}$, 上述构造过程都在有限步内终止, 且存在满足 $\tilde{\varphi}_{t(o)}(o)\in\mathcal{O}\backslash\mathcal{O}^*$ 的整数 $t(o)\geqslant 0$. 据此, 我们定义 $\varphi(o)=\tilde{\varphi}_{t(o)}(o)$. 由于 φ 是通过迭代

调用双射 $\tilde{\varphi}$ 构造的, φ 同样具备双射性. 此外,

$$\sum_{o\in\mathcal{O}^*\backslash\mathcal{O}}\delta(o,\varphi(o))\leqslant\sum_{o\in\mathcal{O}^*\backslash\mathcal{O}}\sum_{i=1}^{t(o)}\delta(\tilde{\varphi}_{i-1}(o),\tilde{\varphi}_i(o))\leqslant\sum_{o\in\mathcal{O}^*}\delta(o,\tilde{\varphi}(o))\leqslant 2\sum_{o\in\mathcal{O}^*}\delta(o,\mathcal{A}),$$

其中, 第 1 步由三角不等式得出, 第 2 步由 $\tilde{\varphi}_i(o)$ 的定义得出, 第 3 步由不等式 (6) 得出. 因此, 引理 4 成立.

4.3　基于归约的求解算法

算法 3 为本节针对 MkzC 问题提出的求解算法. 该算法首调用算法 Anchor-Sel 构造锚点集 $\mathcal{A}$, 然后结合锚点集与算法 Outlier-Sel 构造候选异常点集. 为了保证引理 4 中声明的候选集被成功构造的概率可被提升为常数, 算法 3 将算法 Outlier-Sel 重复执行 z^z 次以生成候选异常点集族. 对于该族中的每个候选异常点集, 算法 3 将其从 $\mathcal{P}$ 中移除, 并调用 MkC 问题的求解算法 Clust 得出实例 $\mathcal{I}$ 的候选解.

算法 3 基于归约的求解算法.

输入: 常数 $\epsilon\in(0,1)$, MkzC 问题的实例 $\mathcal{I}=\big((\mathcal{X},\delta),\mathcal{P},\mathcal{C},k,z\big)$, 以及只在 $z=0$ 时适用的求解算法 Clust;
输出: 实例 $\mathcal{I}$ 的解 $(\mathcal{S}^\dagger,\mathcal{O}^\dagger,\tau^\dagger)$;

$\mathcal{A}\leftarrow$ Anchor-Sel$(\epsilon,\mathcal{I})$;
$\mathcal{D}\leftarrow\emptyset$;
for $i\in[z^z]$ **do**
　$\mathbb{O}\leftarrow$ Outlier-Sel$(\epsilon,\mathcal{I},\mathcal{A})$;
　for $\mathcal{O}\in\mathbb{O}$ **do**
　　令 $(\mathcal{S},\emptyset,\tau)$ 为算法 Clust 为实例 $\big((\mathcal{X},\delta),\mathcal{P}\backslash\mathcal{O},\mathcal{C},k,0\big)$ 输出的解;
　　$\mathcal{D}\leftarrow\mathcal{D}\cup\{(\mathcal{S},\mathcal{O},\tau)\}$;
　end for
end for
return $(\mathcal{S}^\dagger,\mathcal{O}^\dagger,\tau^\dagger)\leftarrow\underset{(\mathcal{S},\mathcal{O},\tau)\in\mathcal{D}}{\arg\min}\max_{c\in\mathcal{C}}\sum_{p\in\tau^{-1}(c)}d(p,c)$.

下面通过分析算法 3 证明定理 1 的正确性.

证明　(定理 1) 我们首先分析算法 3 的近似比. 令 α 为算法 Clust 的近似比. 令 $(\mathcal{S}^\dagger,\mathcal{O}^\dagger,\tau^\dagger)$ 表示算法 3 为实例 $\mathcal{I}$ 返回的解. 令 $\mathcal{O}$ 和 $\varphi:\mathcal{O}^*\backslash\mathcal{O}\to\mathcal{O}\backslash\mathcal{O}^*$ 分别表示引理 4 中声明的异常点集和双射. 由于算法 3 将算法 Outlier-Sel 重复执行 z^z 次以生成候选异常点集族, 上述异常点集被成功构造的概率可被提升为

$$1-(1-z^{-z})^{z^z}>1-e^{-1}.$$

在该事件成立的条件下, 令 $(\mathcal{S},\mathcal{O},\tau)$ 表示算法 3 在实例 $\big((\mathcal{X},\delta),\mathcal{P}\backslash\mathcal{O},\mathcal{C},k,0\big)$ 上调用算法 Clust 得出的候选解. 给定整数 $i\in[k]$, 令 $\mathcal{P}_i=(\mathcal{P}_i^*\backslash\mathcal{O})\cup\{\varphi^{-1}(p)|p\in\mathcal{P}_i^*\cap\mathcal{O}\}$. 由 $\mathcal{P}_i$ 的定义可知,

$$\bigcup_{i=1}^{k}\mathcal{P}_i=\left(\bigcup_{i=1}^{k}\mathcal{P}_i^*\backslash\mathcal{O}\right)\cup\{\varphi^{-1}(p)\mid p\in\mathcal{O}\backslash\mathcal{O}^*\}=\left(\bigcup_{i=1}^{k}\mathcal{P}_i^*\backslash\mathcal{O}\right)\cup\mathcal{O}^*\backslash\mathcal{O}=\mathcal{P}\backslash\mathcal{O}.\tag{7}$$

因此,

$$\begin{aligned}\max_{c\in\mathcal{S}^\dagger}\sum_{p\in(\tau^\dagger)^{-1}(c)}\delta(p,c)&\leqslant\max_{c\in\mathcal{S}}\sum_{p\in\tau^{-1}(c)}\delta(p,c)\\&\leqslant\alpha\cdot\max_{i\in[k]}\sum_{p\in\mathcal{P}_i}\delta(p,c_i^*)\\&=\alpha\cdot\max_{i\in[k]}\left(\sum_{p\in\mathcal{P}_i^*\backslash\mathcal{O}}\delta(p,c_i^*)+\sum_{p\in\mathcal{P}_i^*\cap\mathcal{O}}\delta(\varphi^{-1}(p),c_i^*)\right)\end{aligned}$$

$$\leqslant \alpha \cdot \max_{i\in[k]} \left(\sum_{p\in\mathcal{P}_i^*} \delta(p, c_i^*) + \sum_{p\in\mathcal{P}_i^*\cap\mathcal{O}} \delta(\varphi^{-1}(p), p) \right)$$
$$= \alpha \left(\mathrm{opt} + \max_{i\in[k]} \sum_{p\in\mathcal{P}_i^*\cap\mathcal{O}} \delta(\varphi^{-1}(p), p) \right), \tag{8}$$

其中, 第 1 步利用算法 3 返回费用最低的候选解这一事实, 第 2 步根据等式 (7) 和 `Clust` 的近似比得出, 第 3 步根据 $\mathcal{P}_i$ 的定义得出, 第 4 步由三角不等式得出. 下面分析 $\max_{i\in[k]} \sum_{p\in\mathcal{P}_i^*\cap\mathcal{O}} \delta(\varphi^{-1}(p), p)$ 的上界. 可以得出, 不等式

$$\begin{aligned}
\max_{i\in[k]} \sum_{p\in\mathcal{P}_i^*\cap\mathcal{O}} \delta(\varphi^{-1}(p), p) &\leqslant \sum_{i=1}^{k} \sum_{p\in\mathcal{P}_i^*\cap\mathcal{O}} \delta(\varphi^{-1}(p), p) \\
&= \sum_{p\in\mathcal{O}\setminus\mathcal{O}^*} \delta(\varphi^{-1}(p), p) \\
&= \sum_{p\in\mathcal{O}^*\setminus\mathcal{O}} \delta(p, \varphi(p)) \\
&\leqslant 2 \sum_{o\in\mathcal{O}^*} \delta(o, \mathcal{A}) \\
&\leqslant 2\epsilon \cdot \mathrm{opt}
\end{aligned} \tag{9}$$

成立的概率不低于 $1-e^{-1/4}$ (该概率下界等同于引理 3 中声明的不等式成立的概率下界), 其中, 第 4 步由引理 4 得出, 第 5 步根据引理 3 得出. 结合不等式 (8) 和 (9) 以及引理 4 中声明的异常点集在重复执行算法 `Outlier-Sel` 的过程中被成功构造的概率不低于 $1-e^{-1}$ 这一事实可知, 不等式

$$\max_{c\in\mathcal{S}^\dagger} \sum_{p\in(\tau^\dagger)^{-1}(c)} \delta(p, c) \leqslant \alpha(1+2\epsilon)\mathrm{opt}$$

成立的概率不低于 $(1-e^{-1/4})(1-e^{-1})$. 该不等式说明算法 3 的近似比为 $\alpha(1+2\epsilon)$.

令 T 为算法 `Clust` 的时间复杂度. 算法 3 在调用 `Anchor-Sel` 构造锚点集 $\mathcal{A}$ 时需要利用引理 1 中的多项式时间算法求解一个 $(k+z)$- 中位问题, 并根据每个点与已选最近锚点间的距离随机选取 $O(kz\epsilon^{-1})$ 个点. 该过程在一般度量空间及 d- 维欧氏空间中所需时间分别为 $(|\mathcal{P}|\cdot|\mathcal{C}|)^{O(1)}+O(|\mathcal{P}|kz\epsilon^{-1})$ 和 $|\mathcal{P}|^{O(1)}d+O(|\mathcal{P}|dkz\epsilon^{-1})$. 令 ρ 为算法 3 所构造的候选异常点集的数量. 在构造每个候选异常点集及相应的候选解时, 算法 3 通过 `Outlier-Sel` 在 $\mathcal{A}$ 中最多 z 个锚点的邻域中选取候选异常点, 并调用算法 `Clust` 生成候选解. 该过程在一般度量空间和欧氏空间中所需时间分别为 $T+O(|\mathcal{P}|z)$ 和 $T+O(|\mathcal{P}|dz)$. 综上所述, 算法 3 在一般度量空间和 d- 维欧氏空间中的时间复杂度分别为 $\rho(T+O(|\mathcal{P}|z))+O(|\mathcal{P}|kz\epsilon^{-1})+(|\mathcal{P}|\cdot|\mathcal{C}|)^{O(1)}$ 和 $\rho(T+O(|\mathcal{P}|dz))+O(|\mathcal{P}|dkz\epsilon^{-1})+|\mathcal{P}|^{O(1)}d$.

下面分析 ρ 的取值. 由于算法 3 将 `Outlier-Sel` 重复执行 z^z 次以构造候选异常点集, ρ 为 z^z 与满足 $\sum_{i=1}^{|\mathcal{A}|}\gamma_i = z$ 的非负整数 $|\mathcal{A}|$- 元组 $(\gamma_1, \cdots, \gamma_{|\mathcal{A}|})$ 的数量的乘积. 利用隔板法分析该类 $|\mathcal{A}|$- 元组的数量可得

$$\rho = z^z \cdot \binom{|\mathcal{A}|+z-1}{z} \leqslant z^z \cdot \frac{(|\mathcal{A}|+z-1)^z}{z!} \leqslant \left(e(|\mathcal{A}|+z-1)\right)^z, \tag{10}$$

其中, 第 3 步由斯特林下界得出. 由算法 1 的构造过程可知, 锚点集 $\mathcal{A}$ 的规模为 $[\lceil 2kz\beta\epsilon^{-1}\rceil]+k+z$, 其中, β 为引理 1 中算法的近似比, 在一般度量空间和欧氏空间中分别取 2.613 和 2.406. 结合这一事实与不等式 (10) 可知, ρ 在一般度量空间和欧氏空间中的上界分别为 $\left(e(5.226kz\epsilon^{-1}+k+2z)\right)^z$ 和 $\left(e(4.812kz\epsilon^{-1}+k+2z)\right)^z$.

令 $\varepsilon = 2\epsilon$, 则上述论证说明定理 1 成立.

5　总结

本文以中心及异常点数量上限作为参数, 分别在一般度量空间和欧氏空间中为极小 – 极大 (k,z)- 聚类问题提出了近似比为 $3+\varepsilon$ 和 $1+\varepsilon$ 的参数化近似算法. 这是关于该问题的首个常数近似保证. 本文算法通过识别近似最优的异常点集将极小 – 极大 (k,z)- 聚类问题实例归约为相应的极小 – 极大 k- 聚类问题实例, 并在归约后的实例上调用已有的参数化近似算法 [6] 以获取中心集及点集分配方式. 为保证所识别的异常点与最优解中的异常点足够接近, 算法对每个锚点的邻域中可能存在的异常点数量进行枚举, 并据此选取相同数量的锚点最近邻作为候选异常点. 该枚举策略能在归约过程中确保在算法近似比中引入的乘法放大因子不超过 $1+\varepsilon$, 但在时间复杂度中引入了对异常点数量上限的指数级依赖. 一个值得进一步探索的方向是, 能否通过放宽对算法近似比的要求降低其时间复杂度. 例如, 当允许归约过程在近似比中引入较大的常数级乘法放大因子时, 是否能够通过松弛锚点与最优异常点间的距离约束压缩锚点集及其对应的候选异常点集, 从而提出时间复杂度更低的常数近似算法, 还有待深入研究. 另一方面, 当不再要求算法具有可被严格证明的近似保证时, 可以考虑采用非枚举式的异常点选择机制: 在由锚点确定的候选集中, 根据孤立森林 [26]、k- 近邻 [27]、局部离群因子 [28] 等异常度量筛选异常点. 这一方向的研究有望为极小 – 极大 (k,z)- 聚类问题提出更具实用性的求解框架. 需要指出的是, 由于基于归约的极小 – 极大 (k,z)- 聚类问题求解思路以极小 – 极大 k- 聚类问题的算法为基础, 上述关于时间复杂度的优化方向也依赖于对后者的进一步研究.

参考文献

1 Zhao J, Yang K, Wei X, et al. A heuristic clustering-based task deployment approach for load balancing using Bayes theorem in cloud environment. IEEE Trans Parallel Distrib Syst, 2016, 27: 305–316
2 Li T, Ying S, Zhao Y, et al. Batch jobs load balancing scheduling in cloud computing using distributional reinforcement learning. IEEE Trans Parallel Distrib Syst, 2024, 35: 169–185
3 Cao B, Glover F. Creating balanced and connected clusters to improve service delivery routes in logistics planning. J Syst Sci Syst Eng, 2010, 19: 453–480
4 Ruan J H, Wang X P, Chan F T S, et al. Optimizing the intermodal transportation of emergency medical supplies using balanced fuzzy clustering. Int J Production Res, 2016, 54: 4368–4386
5 Ahmadian S, Behsaz B, Friggstad Z, et al. Approximation algorithms for minimum-load k-facility location. ACM Trans Algorithms, 2018, 14: 1–29
6 Bandyapadhyay S, Fomin F V, Golovach P A, et al. FPT approximation for fair minimum-load clustering. In: Proceedings of the 17th International Symposium on Parameterized and Exact Computation, 2022
7 Sánchez Vinces B V, Schubert E, Zimek A, et al. A comparative evaluation of clustering-based outlier detection. Data Min Knowl Disc, 2025, 39: 13
8 Amagata D. Fair k-center clustering with outliers. In: Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, 2024. 10–18
9 Huang J, Liu W, Ding H. Bi-criteria sublinear time algorithms for clustering with outliers in high dimensions. Theor Comput Sci, 2025, 1057: 115538
10 Charikar M, Khuller S, Mount D M, et al. Algorithms for facility location problems with outliers. In: Proceedings of the 12th Annual Symposium on Discrete Algorithms, 2001. 642–651
11 Chen K. A constant factor approximation algorithm for k-median clustering with outliers. In: Proceedings of the 19th Annual Symposium on Discrete Algorithms, 2008. 826–835
12 Krishnaswamy R, Li S, Sandeep S. Constant approximation for k-median and k-means with outliers via iterative rounding. In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, 2018. 646–659
13 Gupta A, Moseley B, Zhou R. Structural iterative rounding for generalized k-median problems. Math Program, 2025, 212: 581–634
14 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
15 Jaiswal R, Kumar A. Clustering what matters in constrained settings. Algorithmica, 2025, 87: 1178–1198
16 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
17 Zhang Z, Huang J Y, Feng Q L. Faster approximation schemes for (constrained) k-means with outliers. In: Proceedings of the

49th International Symposium on Mathematical Foundations of Computer Science, 2024. 1–17
18 Dabas R, Gupta N, Inamdar T. FPT approximation for capacitated clustering with outliers. Theor Comput Sci, 2025, 1027: 115026
19 Zhang Z, Chen X, Liu L M, et al. A (1+varepsilon)-approximation algorithm for diversity-aware k-median. Sci Sin Inform, 2025, 55: 32–45 [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]
20 Even G, Garg N, Konemann J, et al. Covering graphs using trees and stars. In: Proceedings of the 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems and the 7th International Workshop on Randomization and Approximation Techniques in Computer Science, 2003. 24–35
21 Arkin E M, Hassin R, Levin A. Approximations for minimum and min-max vehicle routing problems. J Algorithms, 2006, 59: 1–18
22 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
23 Cohen-Addad V, Grandoni F, Lee E, et al. Breaching the 2 LMP approximation barrier for facility location with applications to k-median. In: Proceedings of the 34th ACM-SIAM Symposium on Discrete Algorithms, 2023. 940–986
24 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
25 Chernoff H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann Math Statist, 1952, 23: 493–507
26 Liu F T, Ting K M, Zhou Z H. Isolation-based anomaly detection. ACM Trans Knowl Discov Data, 2012, 6: 1–39
27 Gu X Y, Akoglu L, Rinaldo A. Statistical analysis of nearest-neighbor methods for anomaly detection. In: Proceedings of the 32nd Annual Conference on Neural Information Processing Systems, 2019. 10921–10931
28 Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying density-based local outliers. In: Proceedings of the 26th ACM SIGMOD International Conference on Management of Data, 2000. 93–104

Parameterized approximations for the minimax (k, z)-clustering problem

Zhen ZHANG[1,2], Yige YUAN[2*], Di WU[3], Zhiping TIAN[1] & Qilong FENG[2,4]

1. *School of Frontier Crossover Studies, Hunan University of Technology and Business, Changsha 410205, China*
2. *Xiangjiang Laboratory, Changsha 410205, China*
3. *School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410076, China*
4. *School of Computer Science, Central South University, Changsha 410083, China*
* Corresponding author. E-mail: immyyuan23@163.com

Abstract Given a set of points and a set of candidate centers in a metric space, along with two positive integers k and z, the minimax (k, z)-clustering problem is designed to select at most k centers, remove up to z outliers from the point set, and partition the remaining points into k clusters. The objective is to minimize the maximum clustering cost among all clusters, where the cost of a cluster is defined as the sum of distances from its points to the corresponding center. The minimax (k, z)-clustering problem plays an important role in applications with load-balancing requirements. However, it is computationally challenging: the best known polynomial-time approximation algorithm achieves only an approximation ratio of $O(k)$, and it is fixed-parameter intractable even when both k and z are treated as fixed parameters and the centers selected by an optimal solution are given. In light of this, we focus on parameterized approximation algorithms for the problem. We give a $(3+\varepsilon)$-approximation algorithm with a running time of $2^{\tilde{O}(kz\varepsilon^{-1})}n^{O(1)}$, which constitutes the first constant-factor approximation for the problem. When the input instance lies in a Euclidean space, applying the same approach yields a $(1+\varepsilon)$-approximation algorithm with a comparable running time.

Keywords approximation algorithms, parameterized algorithms, reduction, clustering, sampling