着色 (k,ℓ)-中值问题的固定参数近似算法

陈晓红[1),2)] 张 震[1),2)] 徐雪松[1),2)] 陈 杰[1),2)] 袁汉春[3)] 石 峰[4)]

[1)](湖南工商大学数据智能与智慧社会国家重点实验室培育基地 长沙 410205)

[2)](湘江实验室 长沙 410205)

[3)](浙江师范大学计算机科学与技术学院 浙江 金华 321004)

[4)](中南大学计算机学院 长沙 410083)

摘 要 给定正整数 k 和非负整数 ℓ 以及度量空间中的一组设施和着色用户，着色 (k,ℓ)-中值问题旨在选取不超过 k 个开设设施、在用户集合中移除最多 ℓ 个异常点并为剩余的每个用户分配一个开设设施，使得颜色相同的用户对应不同设施，且用户与对应设施之间的距离之和最小。本文利用新的随机采样方法确定用来选取开设设施的引导点集合，并围绕引导点为问题实例构造小规模候选解集合。本文基于此为着色 (k,ℓ)-中值问题提出了时间复杂度为 $((k+\ell)\varepsilon^{-1})^{O(k)}n^{O(1)}$ 的 $(3+\varepsilon)$-近似算法，其中，n 为问题实例中设施与用户数量之和。这是关于着色 (k,ℓ)-中值问题的第一个具有性能保证的求解算法。该算法与此前人们在不考虑着色约束的情况下提出的固定参数近似算法有相同的时间复杂度和近似比。

关键词 固定参数算法；近似算法；k-中值；随机采样；异常点

中图法分类号 TP301 **DOI号** 10.11897/SP.J.1016.2026.00001

On FPT Approximations for the Chromatic (k,ℓ)-Median Problem

CHEN Xiao-Hong[1),2)] ZHANG Zhen[1),2)] XU Xue-Song[1),2)] CHEN Jie[1),2)] YUAN Han-Chun[3)] SHI Feng[4)]

[1)](National Key Laboratory Cultivation Base for Data Intelligence and Smart Society, Hunan University of Technology and Business, Changsha 410205)

[2)](Xiangjiang Laboratory, Changsha 410205)

[3)](School of Computer Science and Technology, Zhejiang Normal University, Jinhua, Zhejiang 321004)

[4)](School of Computer Science, Central South University, Changsha 410083)

Abstract Clustering is a fundamental task in data mining and machine learning, where the goal is to partition a set of data points into clusters with high intra-cluster similarity. Among various clustering formulations, the k-median problem has received significant attention due to its conceptual simplicity and wide applicability. Given a set of facilities and clients in a metric space, along with an integer k, the k-median problem aims at opening at most k facilities and assigning each client to an opened facility to minimize the total distance between clients and their assigned facilities. In many real-world applications, additional constraints need to be incorporated into the k-median problem to account for domain-specific structures. One such example is in the study of the spatial organization of chromosomes in cell nuclei, where it is common to analyze spatial patterns of probe points extracted from homologous chromosomes within multiple cells. These

收稿日期：2025-01-08；在线发布日期：2025-05-20。本课题得到国家自然科学基金卓越研究群体项目(72088101)、湘江实验室重大项目(24XJJCYJ01003)、国家自然科学基金青年科学基金项目(62202160,62202161)、国家重点研发计划项目(2022YFC3302302)、湖南省自然科学基金项目(2023JJ40240)、湖南省教育厅科学研究项目(23B0597)资助。**陈晓红**，博士，教授，中国工程院院士，主要研究领域为数据智能、决策理论。E-mail：csu_cxh@163.com。**张 震(通信作者)**，博士，副教授，中国计算机学会(CCF)会员，主要研究领域为组合优化、近似算法。E-mail：zz@hutb.edu.cn。**徐雪松**，博士，教授，中国计算机学会(CCF)会员，主要研究领域为复杂系统优化、计算机算法优化。**陈 杰**，博士，副教授，中国计算机学会(CCF)会员，主要研究领域为计算智能、隐私计算。**袁汉春**，博士，讲师，中国计算机学会(CCF)会员，主要研究领域为固定参数算法、核心化。**石 峰**，博士，副教授，中国计算机学会(CCF)会员，主要研究领域为图论、固定参数算法。

probe points can be modeled as clients in an instance of the k-median problem, but it is necessary to enforce that probe points from the same cell be assigned to different representative points. Motivated by this, the chromatic k-median problem has been introduced, where each client is associated with a color (e. g., representing the cell it comes from), and clients with the same color must be assigned to different facilities. This chromatic constraint has been used to support more accurate clustering in fields such as computational biology, privacy-preserving learning, and logistics planning. In this paper, we consider a more general and challenging variant of the problem, known as the chromatic (k,ℓ)-median problem, which additionally allows the removal of up to ℓ outliers from the client set. This extension enhances the robustness of the clustering process, particularly in real-world applications where input data may be affected by noise. However, the simultaneous presence of the chromatic constraint and the requirement of outlier detection introduces substantial algorithmic challenges, and no algorithms with provable approximation guarantees are currently known for this setting. We deal with the chromatic (k,ℓ)-median problem under the leader-based framework for algorithm design, which has been widely applied to various clustering problems. In this framework, the client closest to each facility opened in an optimal solution is referred to as a leader. Candidate facilities are then selected from annular regions centered at these leaders. Existing algorithms based on this approach typically rely on coreset constructions to reduce the size of the client set, so that the enumeration over all possible leaders can be performed within fixed-parameter tractable time. However, this strategy cannot be directly applied to the chromatic (k,ℓ)-median problem, as no effective coreset construction is currently known for this setting. As a remedy, we propose a relaxed notion of leaders, where it suffices to find clients whose distances to the corresponding optimal facilities are bounded by a given threshold. We present a sampling-based approach for identifying such weak leaders. Around each weak leader, we construct a carefully selected set of candidate facilities and derive a limited collection of candidate solutions. This enables us to propose a $(3+\varepsilon)$-approximation algorithm that runs in $((k+\ell)\varepsilon^{-1})^{O(k)}n^{O(1)}$ time for the chromatic (k,ℓ)-median problem. Notably, our result matches the previously known result obtained in the simpler case without the chromatic constraint, in terms of both running time and approximation ratio.

Keywords fixed-parameter algorithms; approximation algorithms; k-median; random sampling; outliers

1 引 言

聚类是数据挖掘领域中的核心问题之一。该问题要求将一组用户划分为若干子集，以在不同子集间最大化用户的差异性，同时在同一子集内最大化用户的相似性。在诸多聚类模型中，k-中值（k-Median）问题因其简洁的表达形式和广泛的应用而备受关注。给定度量空间中的一组设施和用户以及正整数 k，k-中值问题旨在开设最多 k 个设施并为每个用户分配一个开设设施，使得用户与对应设施的距离之和最小。k-中值问题是一个 NP-难问题[1]，因此，人们针对其近似算法开展了广泛研究。目前，关于 k-中值问题的最好近似结果是 Gowda 等人[2]基于 Li 和 Svensson[3]提出的亚可行近似解修正方法得到的 $(2.613+\varepsilon)$-近似比（其中，ε 为 $(0,1)$ 内的任意常数）。在相较于一般化的度量空间更为特殊的欧几里得空间中，Cohen-Addad 等人[4]通过挖掘空间性质为 k-中值问题提出了 $(2.406+\varepsilon)$-近似算法。

k-中值问题要求确保用户与其对应的开设设施之间具有较高的相似性。基于这一特性，人们在 k-中值问题的很多应用场景中将开设设施作为用户集合的代表性数据点。例如，在染色体拓扑结构的研究中，可将从一组细胞的目标同源染色体中提取的探针点集作为用户集合，并利用开设设施的位置表征这些探针点在细胞群中的共同空间分布模式。然而，由于同一细胞中提取的探针点彼此对应

不同的同源体，为了更准确地刻画同源体探针点的空间分布模式，需要在聚类过程中引入以下约束：每个开设设施在同一细胞中至多连接一个探针点。针对上述需求，Ding 和 Xu[5]在 k -中值问题的基础上引入着色约束、提出了着色 k -中值（Chromatic k -Median）问题。在该问题中，每个用户被赋予一个颜色（例如，在分析染色体拓扑结构时，同一细胞内的探针点具有相同颜色），且颜色相同的用户必须与不同设施相连。通过引入着色约束，人们在生物信息学[5-6]、隐私计算[7]、交通规划[8]等领域实现了更具针对性的聚类分析。

着色 k -中值问题的求解算法基于用户与开设设施之间的距离划分用户集合。在划分过程中，与开设设施距离较远的少量异常点（Outliers）能明显影响解的结构。在需要处理噪声数据的问题实例中，移除这些异常点往往能产生更合理的聚类结果，如图 1 所示（其中，用户和设施分别为圆形和方形）。鉴于此，本文研究着色 (k,ℓ) -中值（Chromatic (k,ℓ) -Median）问题。给定非负整数 ℓ，该问题允许实例的解在用户集合中移除不超过 ℓ 个异常点以降低其对应的目标函数值。第 3 节中给出了着色 (k,ℓ) -中值问题的形式化定义。

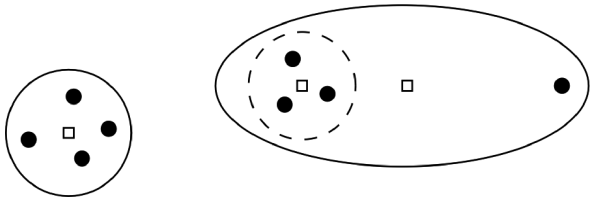


图 1　异常点对聚类结果的影响

在实际应用中处理带异常点的聚类问题（Clustering with Outliers）时，开设设施的数量 k 和异常点的数量 ℓ 通常远小于点集规模。因此，在 k 和 ℓ 取值较小的假设下设计相关问题的求解算法一直是一个热门研究领域[9-13]。本文基于相同假设求解着色 (k,ℓ) -中值问题。不难发现，通过枚举设施集合的子集以及用户的连接方式，我们可以在 $n^{O(k)}$ 时间内得到着色 (k,ℓ) -中值问题实例的最优解，其中，n 为设施与用户数量之和。然而，本文旨在避免对解空间的全面搜索，在固定参数时间（即 $g(k,\ell)n^{O(1)}$ 时间，其中，$g()$ 为任意正值函数）内求解着色 (k,ℓ) -中值问题。

Cohen-Addad 等人[14]基于最大覆盖（Max-Coverage）问题的归约结果表明，存在一个正值函数 $g()$，使得任何能为 k -中值问题实现优于 $1+2e^{-1}$ 的近似比的算法都有不低于 $n^{g(k)}$ 的时间复杂度。鉴于 k -中值问题是着色 (k,ℓ) -中值问题的特殊情况（其中，$\ell=0$ 且用户颜色各不相同），这一结果说明我们无法在 k 和 ℓ 为固定参数的情况下为着色 (k,ℓ) -中值问题设计固定参数时间的精确算法。然而，该归约结果并未否定在固定参数时间内优化 k -中值相关问题近似结果的可能性。例如：Chen 等人[9]以 k 和 ℓ 作为固定参数，为考虑异常点的 (k,ℓ) -中值（(k,ℓ) -Median）问题提出了时间复杂度为 $((k+\ell)\varepsilon^{-1})^{O(k)}n^{O(1)}$ 的 $(3+\varepsilon)$ -近似算法；该结果明显优于此前 Gupta 等人[15]在多项式时间内得到的 $(6.994+\varepsilon)$ -近似结果。本文根据着色约束对现有固定参数时间算法设计框架进行调整，为着色 (k,ℓ) -中值问题提出的算法与 Chen 等人[9]在不考虑着色约束的情况下提出的固定参数近似算法具有相同的时间复杂度和近似比（如第 7 节定理 1 所述）。这是关于着色 (k,ℓ) -中值问题的第一个具有近似保证的固定参数时间求解算法。

本文主要贡献概括如下。

(1)本文结合着色约束为着色 (k,ℓ) -中值问题设计了用于挖掘最优开设设施邻近用户的随机采样算法。本文利用这一算法估计最优解中开设设施的位置，并构造了有效的小规模解搜索空间。

(2)在上述技术的基础上，本文以 k 和 ℓ 作为固定参数，为着色 (k,ℓ) -中值问题提出了第一个具有近似保证的求解算法。这对于更准确地刻画着色 (k,ℓ) -中值问题的求解难度、完善其求解理论具有重要意义。

2　相关工作

本文所研究的着色 (k,ℓ) -中值问题是 (k,ℓ) -中值（(k,ℓ) -Median）问题的推广形式，其中，前者在用户颜色各不相同时等价于 (k,ℓ) -中值问题。鉴于鲁棒聚类分析领域的实际需求，人们广泛研究了 (k,ℓ) -中值问题的求解算法。关于该问题的第一个具有近似保证的算法是 Charikar 等人[16]利用原始-对偶和拉格朗日松弛方法提出的 $(4(1+\varepsilon^{-1}),1+\varepsilon)$ -双标准近似算法。该算法在允许适当违反异常点数量上限、移除至多 $\ell(1+\varepsilon)$ 个异常点的情况下，能保证所得解的费用至多为目标函数最小值的 $4(1+\varepsilon^{-1})$ 倍。Chen[17]基于拉格朗日松弛方法提出了第一个能严格满足实例约束条件的常数近似

算法。此后，Krishnaswamy 等人[18]和 Gupta 等人[15]利用迭代舍入技术分别将 (k,ℓ)-中值问题的近似结果改进为 $7.081+\varepsilon$ 和 $6.994+\varepsilon$。当 k 和 ℓ 为固定参数时，Feldman 和 Schulman[10]基于加权连接费用刻画 (k,ℓ)-中值问题，提出了时间复杂度为 $n\log n(k+\ell)^{O(k+\ell)}$ 的常数近似算法。此后，Chen 等人[9]基于最优解中开设设施的邻近用户构造 (k,ℓ)-中值问题的近似解，提出了时间复杂度为 $((k+\ell)\varepsilon^{-1})^{O(k)}n^{O(1)}$ 的 $(3+\varepsilon)$-近似算法。类似地，在欧几里得空间中，人们以 k 和 ℓ 作为固定参数求解 (k,ℓ)-中值问题的连续型实例（其中，设施可被开设在空间中的任意位置）、基于空间性质提出了近似比为 $1+\varepsilon$ 的固定参数时间算法[10-11]。

在实际应用需求的推动下，人们通过为用户赋加颜色提出了着色聚类问题、公平聚类（Fair Clustering）[19-21]问题等带有额外约束的聚类问题。在为用户的连接方式引入着色约束后，(k,ℓ)-中值问题的求解难度显著提升。例如，在给定开设设施的情况下，我们可以通过将每个用户连接到与其距离最近的开设设施上并将与对应设施距离最远的一部分用户作为异常点来最小化聚类代价，而在着色 (k,ℓ)-中值问题中，用相同方式构造的解无法保证能满足实例的着色约束。这使得着色 (k,ℓ)-中值问题实例最优解的组合结构更加难以利用。目前，还不存在关于着色 (k,ℓ)-中值问题的带有近似保证的求解算法。即使在 $\ell=0$ 的情况下（此时着色 (k,ℓ)-中值问题等价于着色 k-中值问题），是否能为该问题设计多项式时间的常数近似算法也仍然是未知的。

近年来，人们开始基于参数计算理论处理聚类问题中的着色约束。Feng 等人[22]和 Goyal 等人[23]利用 D-采样方法[24]为着色 k-中值问题实例构造候选开设设施集合，并基于此提出了以 k 为固定参数的 $(3+\varepsilon)$-近似算法。当点集位于欧几里得空间且 k 为固定参数时，人们采用枚举方法为最优解中的每个簇构造代表性子集，并将这些子集的中位点作为开设设施，进而为着色 k-中值问题的连续型实例提出了固定参数时间的 $(1+\varepsilon)$-近似算法[5,25-26]。由于着色 k-中值问题要求在目标函数中计算所有用户的连接费用，现有算法在着色约束下无法有效处理用户集合中的异常点。本文利用新的随机采样算法在考虑异常点的情况下构造满足着色约束的小规模解搜索空间，并基于此为着色 (k,ℓ)-中值问题提出了固定参数时间的近似算法。

3 基本定义及引理

本节给出文中将使用的一些定义和引理。以下是着色 (k,ℓ)-中值问题的定义。

定义 1.（着色 (k,ℓ)-中值问题）

输入：着色 (k,ℓ)-中值问题的一个实例 $((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$。该实例包含定义在集合 $\mathcal{X}$ 上且以 d 为距离函数的度量空间 $(\mathcal{X},d)$、用户集合 $\mathcal{C}\subseteq\mathcal{X}$、设施集合 $\mathcal{F}\subseteq\mathcal{X}$、正整数 k 以及非负整数 ℓ，其中，每个用户 $c\in\mathcal{C}$ 都有一个颜色 $\eta(c)$。

输出：实例 $((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$ 的最小费用可行解。该实例的一个可行解 $(\mathcal{D},\mathcal{O},\tau)$ 由满足 $|\mathcal{D}|\leqslant k$ 的开设设施集合 $\mathcal{D}\subseteq\mathcal{F}$、满足 $|\mathcal{O}|\leqslant m$ 的异常点集合 $\mathcal{O}\subseteq\mathcal{C}$ 以及对于满足 $|\tau^{-1}(f)|\geqslant 2$ 的每个开设设施 $f\in\mathcal{D}$ 和每对用户 $c_1,c_2\in\tau^{-1}(f)$ 都有 $\eta(c_1)\neq\eta(c_2)$ 的映射 $\tau:\mathcal{C}\setminus\mathcal{O}\to\mathcal{D}$ 组成。可行解 $(\mathcal{D},\mathcal{O},\tau)$ 的费用为 $\sum_{c\in\mathcal{C}\setminus\mathcal{O}}d(c,\tau(c))$。

给定正整数 i，定义 $[i]=\{1,2,\cdots,i\}$。令 ϵ 表示 $(0,1)$ 内的一个常数。令 $\mathcal{I}=((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$ 表示着色 (k,ℓ)-中值问题的一个实例，其中，$|\mathcal{C}\cup\mathcal{F}|=n$。令 $(\mathcal{D}^*,\mathcal{O}^*,\tau^*)$ 表示 $\mathcal{I}$ 的一个最优解，其中，$\mathcal{D}^*=\{f_1^*,f_2^*,\cdots,f_k^*\}$。给定整数 $i\in[k]$，令 $\mathcal{C}_i^*=\{c\in\mathcal{C}\setminus\mathcal{O}^*:\tau^*(c)=f_i^*\}$ 表示 f_i^* 对应的簇，并令 $opt_i=\sum_{c\in\mathcal{C}_i^*}d(c,f_i^*)$ 表示该簇的聚类代价。令 $opt=\sum_{i=1}^{k}opt_i$ 表示实例 $\mathcal{I}$ 最优解的费用。不失一般性，可以假设 $|\mathcal{C}_1^*|\geqslant|\mathcal{C}_2^*|\geqslant\cdots\geqslant|\mathcal{C}_k^*|$。给定点 $u\in\mathcal{X}$ 和集合 $\mathcal{V}\subseteq\mathcal{X}$，令 $d^{\mathrm{sum}}(\mathcal{V},u)=\sum_{v\in\mathcal{V}}d(v,u)$ 表示 $\mathcal{V}$ 中的点与 u 的距离之和，并令 $d^{\min}(u,\mathcal{V})=\min_{v\in\mathcal{V}}d(u,v)$ 表示 $\mathcal{V}$ 中的点与 u 之间的最小距离。

以下引理为分析 $\mathcal{D}^*$ 中每个设施的邻近用户在对应簇中所占的比重提供了途径。

引理 1. 给定实数 $\gamma>1$、点 $u\in\mathcal{X}$ 和集合 $\mathcal{V}\subseteq\mathcal{X}$，不等式 $|\{v\in\mathcal{V}:d(v,u)\leqslant\gamma d^{\mathrm{sum}}(\mathcal{V},u)\,|\,\mathcal{V}|^{-1}\}|>(1-\gamma^{-1})|\mathcal{V}|$ 成立。

证明. 令 $\mathcal{V}(\gamma)=\{v\in\mathcal{V}:d(v,u)\leqslant\gamma d^{\mathrm{sum}}(\mathcal{V},u)\,|\,\mathcal{V}|^{-1}\}$ 表示以点 u 为中心、半径为 $\gamma d^{\mathrm{sum}}(\mathcal{V},u)\,|\mathcal{V}|^{-1}$ 的球形区域所覆盖的点集。这一定义说明每个点 $v\in\mathcal{V}\setminus\mathcal{V}(\gamma)$ 都满足 $d(v,u)>\gamma d^{\mathrm{sum}}(\mathcal{V},u)\,|\mathcal{V}|^{-1}$。由此可知，

$$d(\mathcal{V}\setminus\mathcal{V}(\gamma),u)>\frac{1}{|\mathcal{V}|}\gamma d^{\mathrm{sum}}(\mathcal{V},u)|\mathcal{V}\setminus\mathcal{V}(\gamma)| \quad (1)$$

此外，由 $\mathcal{V}\setminus\mathcal{V}(\gamma)\subseteq\mathcal{V}$ 这一事实可知 $d^{\mathrm{sum}}(\mathcal{V}\setminus\mathcal{V}(\gamma),u)\leqslant d^{\mathrm{sum}}(\mathcal{V},u)$。结合该不等式与不等式(1)可以得出 $|\mathcal{V}|>\gamma|\mathcal{V}\setminus\mathcal{V}(\gamma)|$。因此，

$$|\mathcal{V}(\gamma)|=|\mathcal{V}|-|\mathcal{V}\setminus\mathcal{V}(\gamma)|>\left(1-\frac{1}{\gamma}\right)|\mathcal{V}|。$$

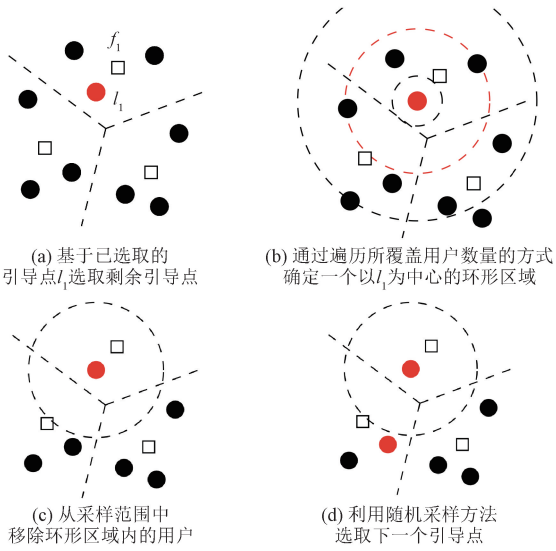
该不等式说明引理 1 成立。　　证毕。

4　算法概述

本文算法基于 Cohen-Addad 等人[14]提出的固定参数近似算法设计框架。该框架将与最优解中的每个开设设施距离最近的用户作为“引导点”(Leader)、在以引导点为中心的环形区域中选取候选开设设施。该思路已被应用于一系列聚类问题求解算法的设计中[27-29]，其中包括 Chen 等人[9]以 k 和 ℓ 作为固定参数、针对 (k,ℓ)-中值问题提出的 $(3+\varepsilon)$-近似算法。Chen 等人[9]基于 Agrawal 等人[11]提出的核心集构造方法压缩用户集合规模，使得枚举用户集合以寻找引导点所需时间可以被限制在固定参数时间。本文所处理的情况更加复杂：是否存在关于着色 (k,ℓ)-中值问题的有效核心集构造方法尚未明确。为此，本文基于 Cohen-Addad 等人[14]给出的框架提出了新的算法设计思路。与现有方法不同的是，本文弱化了对引导点的定义，所提出的算法仅需找到与目标设施之间的距离不超过给定上限的用户（如第 5.2 节中的不变式 $\varphi(i)$ 所述），而非与其距离最近的用户。这一弱化的要求为本文在不构造核心集的情况下于有限时间内寻找引导点提供了可能性。下面概述本文的求解思路。

4.1　引导点选取方法

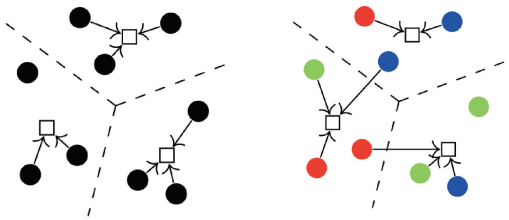
本文在第 5 节中基于采样方法为开设设施选取引导点。由引理 1 可知，在实例最优解的每个簇中，与对应开设设施的距离小于给定阈值的用户所占比例具有与该阈值正相关的下界。本文通过引理 1 分析每个簇中可以作为开设设施引导点的用户所占的比例，并采用随机采样方法从用户集合中选取第一个引导点。图 2 中给出了本文基于已选取的引导点为其他开设设施寻找引导点的基本思路。给定设施 f_1 的引导点 l_1，本文证明了存在一个以 l_1 为中心的球形区域，使得与 l_1 距离较远的某个开设设施所对应的引导点在该区域外的比重较高（不等式(13)）。根据这一结论，本文以 l_1 为中心、通过遍历所覆盖用户数量的方式寻找满足上述要求的球形区域，并在该区域外选取剩余开设设施的引导点。在上述采样过程中所面临的一个挑战是，实例的着色约束使得我们无法直接根据用户与开设设施之间的距离估计最优解中的异常点集合与簇的分布，这明显提高了为每个簇确定合理采样范围的难度。出现这一问题的原因在于实例的最优解中存在为了满足着色约束而与对应开设设施距离较远的用户，如图 3 所示。本文利用已经选取的、与这些用户距离较近的引导点估计这些用户的位置，在已有的引导点集合与采样结果的并集中选择目标设施的引导点。



(a) 基于已选取的引导点l_1选取剩余引导点　(b) 通过遍历所覆盖用户数量的方式确定一个以l_1为中心的环形区域

(c) 从采样范围中移除环形区域内的用户　(d) 利用随机采样方法选取下一个引导点

图 2　算法思路



(a) 不考虑着色约束时，移除异常点后用户的连接方式　(b) 着色约束下，移除异常点后用户的连接方式

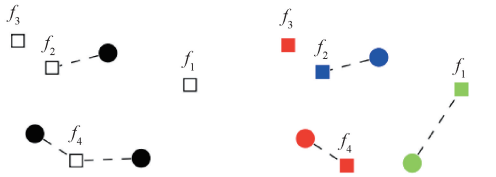
图 3　着色约束对用户连接方式的影响

4.2　基于引导点集合的求解算法

在基于引导点集合为 (k,ℓ)-中值问题构造近似解时，Chen 等人[9]开设与每个引导点距离最近的设施、将每个用户连接到与其距离最近的开设设施并将与对应设施距离最远的 ℓ 个用户作为异常点。然而，在着色约束下，我们无法保证能利用相

同方式构造实例的可行解。首先，在选取与每个引导点距离最近的设施作为开设设施时，可能会出现同一个设施被多次选取的情况，如图 4-(a)所示。此时，我们需要合并对应相同设施的簇，这可能导致解违反实例的着色约束。其次，在着色约束下，不能仅根据用户与设施之间的距离来确定用户的连接方式，还需要保证所选连接方式的可行性。

针对第一个问题，本文在第 6.1 节中利用彩色编码（Color-Coding）技术选取开设设施。具体来说，本文为每个设施分配一个标签，并通过选取标签互不相同的设施避免开设设施集合中出现重复元素，如图 4-(b)所示（其中：颜色相同的设施具有相同的标签；与每个引导点距离最近且带有指定标签的设施被选作开设设施）。针对第二个问题，本文在第 6.2 节中利用最小费用循环（Minimum-Cost Circulation）问题刻画异常点集合的选取以及用户连接映射的构造。在给定开设设施集合的情况下，本文基于最小费用循环问题的多项式时间求解算法[30]构造满足实例着色约束的近似解。结合上述引导点选取方法以及基于引导点集合的求解算法，本文为着色 (k,ℓ)-中值问题提出了固定参数时间的 $(3+\varepsilon)$-近似算法。



(a) 开设设施集合为$\{f_2,f_4\}$，其中，设施f_4被选取两次　(b) 开设设施集合为$\{f_1,f_2,f_4\}$

图 4　基于彩色编码的设施选取方法

5　引导点挖掘

本节通过随机采样寻找与 $\mathcal{D}^*$ 中的设施较为接近的引导点。在 5.1 节中，我们给出针对引导点的具体采样过程。在 5.2 节中，我们分析基于该采样过程获取满足要求的引导点集合的概率。

5.1　随机采样算法

本节为寻找引导点提出的采样过程在算法 Sampling（算法 1）中给出。该算法旨在递归地更新一个候选引导点集合族。给定正整数 k（引导点集合规模上限）、多重集合 $\mathcal{L}$（一个候选引导点集合）、集合 $\mathcal{C}^\dagger$（算法采样范围）以及变量参数 $\mathbb{L}$（待更新的候选引导点集合族），Sampling$(k,\mathcal{L},\mathcal{C}^\dagger,\mathbb{L},d)$ 按照以下方式递归地在 $\mathbb{L}$ 中添加候选引导点集合：(1) 从 $\mathcal{C}^\dagger$ 中随机选取一个用户 c，并以 $\mathcal{L}\cup+\{c\}$ 作为输入集合调用其本身；(2) 移除 $\mathcal{C}^\dagger$ 中与 $\mathcal{L}$ 的成员距离较近的一部分用户，并以消减后的用户集合 $\mathcal{C}^{\neq}\subset\mathcal{C}^\dagger$ 作为采样范围再次调用其本身。

算法 1. Sampling$(k,\mathcal{L},\mathcal{C}^\dagger,\mathbb{L},d)$

输入：正整数 k，集合 $\mathcal{L}$、$\mathcal{C}^\dagger$ 和 $\mathbb{L}$ 以及距离函数 d；

1. IF $|\mathcal{L}|=k$ THEN
2. 　$\mathbb{L}\leftarrow\mathbb{L}\cup\{\mathcal{L}\}$；
3. ELSE
4. 　均匀随机地在 $\mathcal{C}^\dagger$ 中选取一个用户 c；
5. 　Sampling$(k,\mathcal{L}\uplus\{c\},\mathcal{C}^\dagger,\mathbb{L},d)$；
6. 　IF $\mathcal{L}\neq\emptyset$ 且 $|\mathcal{C}^\dagger|>1$ THEN
7. 　　$\mathcal{C}^{\pm}\leftarrow\underset{\mathcal{C}'\subseteq\mathcal{C}^\dagger\wedge|\mathcal{C}'|=|\mathcal{C}^\dagger|/2}{\arg\max}\sum_{c\in\mathcal{C}'}d^{\min}(c,\mathcal{L})$；
8. 　　Sampling$(k,\mathcal{L},\mathcal{C}^{\pm},\mathbb{L},d)$.

以下引理给出了算法 Sampling 的时间复杂度及其构造的候选引导点集合规模。

引理 2. 给定集合 $\mathbb{L}$ 和着色 (k,ℓ)-中值问题的实例 $((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$，Sampling$(k,\emptyset,\mathcal{C},\mathbb{L},d)$ 的时间复杂度为 $k^{O(k)}|\mathcal{C}|$，且其向 $\mathbb{L}$ 中添加的集合数量不超过 $2^k|\mathcal{C}|$。

证明. 给定正整数 s 和非负整数 t 以及满足 $|\mathcal{C}^\dagger|=s$ 和 $|\mathcal{L}|=k-t$ 的集合 $\mathcal{C}^\dagger$ 和 $\mathcal{L}$，令 $T(s,t)$ 表示 Sampling$(k,\mathcal{L},\mathcal{C}^\dagger,\mathbb{L},d)$ 的运行时间，并令 $S(s,t)$ 表示其向 $\mathbb{L}$ 中添加的集合数量。Sampling$(k,\mathcal{L},\mathcal{C}^\dagger,\mathbb{L},d)$ 在第 5 步递归调用 Sampling$(k,\mathcal{L}\uplus\{c\},\mathcal{C}^\dagger,\mathbb{L},d)$，在第 7 步花费 $O(s(k-t))$ 时间从 $\mathcal{C}^\dagger$ 中移除一半的用户以构造集合 $\mathcal{C}^{\pm}$，并在第 8 步递归调用 Sampling$(k,\mathcal{L},\mathcal{C}^{\pm},\mathbb{L},d)$。由此可知，给定整数 $s\geqslant 2$ 和 $t\geqslant 1$，等式

$$S(s,t)=S(s,t-1)+S\left(\frac{s}{2},t\right) \tag{2}$$

成立，且存在满足

$$T(s,t)=T(s,t-1)+T\left(\frac{s}{2},t\right)+\gamma s(k-t) \tag{3}$$

的常数 γ。

我们声明每个整数 $s\geqslant 1$ 和 $t\geqslant 0$ 都满足不等式 $T(s,t)\leqslant 2^t\gamma(t+1)^t ks$ 和 $S(s,t)\leqslant 2^t s$。由于 $T(s,0)$、$T(1,t)$、$S(s,0)$ 和 $S(1,t)$ 都是常数，这一声明在 $s=1$ 或 $t=0$ 的情况下成立。给定整数 $s\geqslant 2$ 和 $t\geqslant 1$，本节假设上述声明对于任意整数 $s'\in[1,s)$ 和 $t'\in[0,t)$ 都成立。结合这一假设与

不等式(2)和不等式(3)可知

$$\begin{aligned}T(s,t)&\leqslant 2^{t-1}\gamma t^{t-1}ks+2^t\gamma(t+1)^t k\ \frac{s}{2}\\&\quad+\gamma s(k-t)\\&\leqslant \gamma ks(2^{t-1}t^{t-1}+2^{t-1}(t+1)^t+1)\\&\leqslant \gamma ks(2^{t-1}(t^{t-1}+1)+2^{t-1}(t+1)^t)\\&\leqslant \gamma ks(2^{t-1}(t+1)^t+2^{t-1}(t+1)^t)\\&=2^t\gamma(t+1)^t ks,\end{aligned}$$

且

$$S(s,t)\leqslant 2^{t-1}s+2^t\ \frac{s}{2}=2^t s。$$

由此可知,我们以归纳假设的方式证明了每个整数 $s\geqslant 1$ 和 $t\geqslant 0$ 都满足 $T(s,t)\leqslant 2^t\gamma(t+1)^t ks$ 和 $S(s,t)\leqslant 2^t s$。令 $s=|\mathcal{C}|$、$t=k$,则上述论证表明 Sampling $(k,\emptyset,\mathcal{C},\mathbb{L},d)$ 的时间复杂度为 $O(2^k(k+1)^k k\,|\mathcal{C}|)\leqslant k^{O(k)}|\mathcal{C}|$,且其向 $\mathbb{L}$ 中添加的集合数量不超过 $2^k|\mathcal{C}|$。　　证毕。

本节用一个树 $\mathcal{T}$ 表示 Sampling $(k,\emptyset,\mathcal{C},\emptyset,d)$ 的运行过程。$\mathcal{T}$ 中的每个节点 $(\mathcal{L},\mathcal{C}^\dagger)$ 对应算法以集合 $\mathcal{L}$ 和 $\mathcal{C}^\dagger$ 作为输入的一次执行过程。一个节点的子节点对应其执行过程中的递归调用操作,如图 5 所示(其中:白色用户表示算法选取的引导点;节点 $(\{c_2\},\{c_1,c_2,\cdots,c_6\})$ 对应的算法执行过程在第 4 步选取用户 c_4,在第 7 步将用户集合消减为 $\{c_4,c_5,c_6\}$)。$\mathcal{T}$ 的叶子节点以规模为 k、被添加到 $\mathbb{L}$ 中的候选引导点集合作为输入。

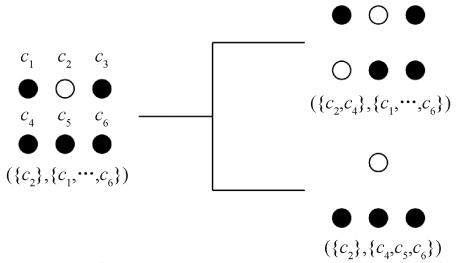


图 5　节点 $(\{c_2\},\{c_1,c_2,\cdots,c_6\})$ 的子节点

5.2　算法分析

令 $\mathbb{L}$ 表示 Sampling $(k,\emptyset,\mathcal{C},\emptyset,d)$ 构造的集合。给定实数 $\gamma>1$ 和整数 $i\in[k]$,令 $\mathcal{H}_i^\gamma=\{c\in\mathcal{C}_i^*:d(c,f_i^*)\leqslant opt_i|\mathcal{C}_i^*|^{-1}\gamma\}$ 表示位于以 f_i^* 为球心、以 $opt_i|\mathcal{C}_i^*|^{-1}\gamma$ 为半径的球体内的用户子集。引理 1 说明在 $\mathcal{C}_i^*$ 中随机选取的用户有较高的概率属于 $\mathcal{H}_i^{1+\epsilon}$。由此可知,如果 $\mathcal{C}_i^*$ 是采样范围的子集且在其中占有较大比重,则算法 Sampling 可以通过随机采样找到 $\mathcal{H}_i^{1+\epsilon}$ 中的一个用户并将其作为 f_i^* 的引导点。为了构造满足要求的引导点集合,算法 Sampling 递归地调整采样范围以确保这一前提条件的成立。

本节通过考虑以下不变式证明 $\mathbb{L}$ 有较高的概率包含与 $\mathcal{D}^*$ 较为接近的集合。

$\varphi(i)$:给定整数 $i\in[k]$,$\mathcal{T}$ 中存在满足以下性质的节点 $(\mathcal{L}_i,\mathcal{C}^\dagger)$ 的概率不低于 $(20(k+\ell)\epsilon^{-2})^{-i}$:(1)$|\mathcal{L}_i|=i$;(2) $\{c\in\mathcal{C}:d^{\min}(c,\mathcal{L}_i)>\epsilon(k\,|\mathcal{C}_i^*|)^{-1}opt\}\subseteq\mathcal{C}^\dagger$;(3) $\sum_{j=1}^{i}|\mathcal{C}_j^*|d^{\min}(f_j^*,\mathcal{L}_i)\leqslant(1+\epsilon)\sum_{j=1}^{i}opt_j+\epsilon\, ik^{-1}opt$。

本节基于归纳假设方法证明不变式 $\varphi(i)$ 在 $i\in[k]$ 时的正确性。可以得出,

$$\begin{aligned}\frac{\mathcal{H}_1^{1+\epsilon}}{|\mathcal{C}|}&\geqslant\frac{|\mathcal{H}_1^{1+\epsilon}|}{k\,|\mathcal{C}_1^*|+\ell}\\&>\frac{\epsilon}{(1+\epsilon)(k+\ell)}\\&>\frac{\epsilon^2}{20(k+\ell)}\end{aligned}\tag{4}$$

其中,第 1 步基于 $|\mathcal{C}_1^*|\geqslant|\mathcal{C}_2^*|\geqslant\cdots\geqslant|\mathcal{C}_k^*|$ 这一假设以及不等式 $|\mathcal{O}^*|\leqslant\ell$ 得出,第 2 步基于 $\mathcal{H}_1^{1+\epsilon}$ 的定义和引理 1 得出。结合不等式 (4) 和 $\mathcal{L}_1$ 中的用户是算法 Sampling 在 $\mathcal{C}$ 中通过均匀随机采样的方式选取这一事实可知,$\mathcal{L}_1\subseteq\mathcal{H}_1^{1+\epsilon}$ 成立的概率不低于 $(20(k+\ell)\epsilon^{-2})^{-1}$。如果 $\mathcal{L}_1\subseteq\mathcal{H}_1^{1+\epsilon}$ 成立,则 $\mathcal{H}_1^{1+\epsilon}$ 的定义说明 $|\mathcal{C}_1^*|d^{\min}(f_1^*,\mathcal{L}_1)\leqslant(1+\epsilon)opt_1$。因此,节点 $(\mathcal{L}_1,\mathcal{C})$ 满足 $\varphi(1)$ 中声明的性质。

下面考虑 $i>1$ 的情况。本节在 $\varphi(i-1)$ 成立的假设下证明 $\varphi(i)$。由 $\varphi(i-1)$ 可知,满足

$$\begin{aligned}&\sum_{j=1}^{i-1}|\mathcal{C}_j^*|d^{\min}(f_j^*,\mathcal{L}_{i-1})\\&\leqslant(1+\epsilon)\sum_{j=1}^{i-1}opt_j+\frac{1}{k}\epsilon(i-1)opt\end{aligned}\tag{5}$$

和

$$\{c\in\mathcal{C}:d^{\min}(c,\mathcal{L}_{i-1})>\epsilon(k\,|\mathcal{C}_{i-1}^*|)^{-1}opt\}\subseteq\mathcal{C}^\dagger\tag{6}$$

的节点 $(\mathcal{L}_{i-1},\mathcal{C}^\dagger)$ 存在的概率不低于 $(20(k+\ell)\epsilon^{-2})^{-(i-1)}$。本节以存在这一节点 $(\mathcal{L}_{i-1},\mathcal{C}^\dagger)$ 为前提条件分析 $\varphi(i)$ 的正确性。令 $\mathcal{G}_{i-1}=\{c\in\mathcal{C}:d^{\min}(c,\mathcal{L}_{i-1})\leqslant\epsilon(k\,|\mathcal{C}_i^*|)^{-1}opt\}$。本节分别分析 $\mathcal{H}_i^{1+\epsilon}\cap\mathcal{G}_{i-1}\neq\emptyset$ 和 $\mathcal{H}_i^{1+\epsilon}\cap\mathcal{G}_{i-1}=\emptyset$ 两种情况。

在 $\mathcal{H}_i^{1+\epsilon}\cap\mathcal{G}_{i-1}\neq\emptyset$ 的情况下,本节结合三角不等式以及 $\mathcal{H}_i^{1+\epsilon}$ 和 $\mathcal{G}_{i-1}$ 的定义分析 f_i^* 与 $\mathcal{L}_{i-1}$ 中用户之间的距离(如图 6 所示),并基于此证明 $\varphi(i)$ 的正确性。

引理 3. 如果 $\mathcal{H}_i^{1+\epsilon}\cap\mathcal{G}_{i-1}\neq\emptyset$,则 $(\mathcal{L}_{i-1},\mathcal{C}^\dagger)$ 的子节点 $(\mathcal{L}_{i-1}\uplus\{c_i\},\mathcal{C}^\dagger)$ 满足

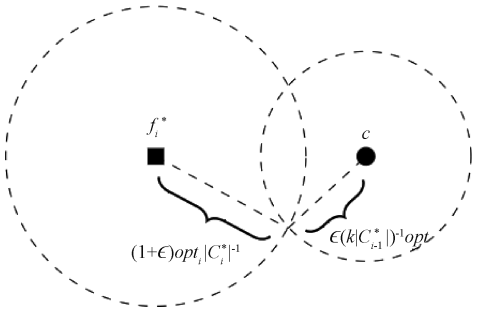


图 6 $\mathcal{H}_i^{1+\epsilon}\cap\mathcal{G}_{i-1}\neq\emptyset$ 的情况下 f_i^* 与 $\mathcal{L}_{i-1}$ 中的某个用户 c 之间的距离

$$\sum_{j=1}^{i}|\mathcal{C}_j^*|\,d^{\min}(f_j^*,\mathcal{L}_{i-1}\uplus\{c_i\})\leqslant(1+\epsilon)\sum_{j=1}^{i}opt_j+\epsilon\,ik^{-1}opt$$

和

$$\{c\in\mathcal{C}:d^{\min}(c,\mathcal{L}_{i-1}\uplus\{c_i\}>\epsilon(k|\mathcal{C}_i^*|)^{-1}opt\}\subseteq\mathcal{C}^{\dagger}$$

证明. 我们首先基于 $\mathcal{H}_i^{1+\epsilon}$ 和 $\mathcal{G}_{i-1}$ 的定义证明 f_i^* 与 $\mathcal{L}_{i-1}$ 中的一个用户较为接近。令 $c^{\dagger}$ 表示 $\mathcal{H}_i^{1+\epsilon}\cap\mathcal{G}_{i-1}$ 中的一个用户，并令 $l(c^{\dagger})$ 表示 $\mathcal{L}_{i-1}$ 中与 $c^{\dagger}$ 距离最近的用户。可以得出，

$$\begin{aligned}&d^{\min}(f_i^*,\mathcal{L}_{i-1})\\&\leqslant d(f_i^*,l(c^{\dagger}))\\&\leqslant d(f_i^*,(c^{\dagger})+d(c^{\dagger},l(c^{\dagger})))\\&\leqslant\max_{c\in\mathcal{H}_i^{1+\epsilon}}d(f_i^*,c)+d^{\min}(c^{\dagger},\mathcal{L}_{i-1})\\&\leqslant(1+\epsilon)\frac{opt_i}{|\mathcal{C}_i^*|}+d^{\min}(c^{\dagger},\mathcal{L}_{i-1})\\&\leqslant(1+\epsilon)\frac{opt_i}{|\mathcal{C}_i^*|}+\epsilon\frac{opt_i}{k|\mathcal{C}_i^*|}\end{aligned}\tag{7}$$

其中，第 2 步由三角不等式得出，第 3 步利用式 $c^{\dagger}\in\mathcal{H}_i^{1+\epsilon}$ 以及 $l(c^{\dagger})$ 的定义得出，第 4 步基于 $\mathcal{H}_i^{1+\epsilon}$ 的定义得出，第 5 步基于 $c^{\dagger}\in\mathcal{G}_{i-1}$ 这一事实以及 $\mathcal{G}_{i-1}$ 的定义得出。不等式 (7) 和不等式 (5) 说明，

$$\begin{aligned}&\sum_{j=1}^{i}|\mathcal{C}_j^*|\,d^{\min}(f_j^*,\mathcal{L}_{i-1}\uplus\{c_i\})\\&\leqslant\sum_{j=1}^{i}|\mathcal{C}_j^*|\,d^{\min}(f_j^*,\mathcal{L}_{i-1})\\&=\sum_{j=1}^{i-1}|\mathcal{C}_j^*|\,d^{\min}(f_j^*,\mathcal{L}_{i-1})+|\mathcal{C}_i^*|\,d^{\min}(f_j^*,\mathcal{L}_{i-1})\\&\leqslant1+\epsilon\sum_{j=1}^{i-1}opt_j+\frac{1}{k}\epsilon(i-1)opt+|\mathcal{C}_j^*|\,d^{\min}(f_j^*,\mathcal{L}_{i-1})\\&\leqslant(1+\epsilon)\sum_{j=1}^{i}opt_j+\frac{\epsilon_i}{k}opt\end{aligned}\tag{8}$$

由 $|\mathcal{C}_1^*|\geqslant|\mathcal{C}_2^*|\geqslant\cdots\geqslant|\mathcal{C}_k^*|$ 这一假设以及每对用户 $c,c_i\in\mathcal{C}$ 都满足 $d^{\min}(c,\mathcal{L}_{i-1}\uplus\{c_i\})\leqslant d^{\min}(c,\mathcal{L}_{i-1})$ 这一事实可知，

$$\begin{aligned}&\{c\in\mathcal{C}:d^{\min}(c,\mathcal{L}_{i-1}\uplus\{c_i\})>\epsilon(k|\mathcal{C}_i^*|)^{-1}opt\}\\&\subseteq\{c\in\mathcal{C}:d^{\min}(c,\mathcal{L}_{i-1})>\epsilon(k|\mathcal{C}_{i-1}^*|)^{-1}opt\}\\&\subseteq\mathcal{C}^{\dagger}\end{aligned}\tag{9}$$

其中，第 2 步基于式(6)得出。结合不等式(8)和式(9)可知，引理 3 成立。 证毕。

针对 $\mathcal{H}_i^{1+\epsilon}\cap\mathcal{G}_{i-1}=\emptyset$ 的情况，本节一方面分析 $|\mathcal{H}_i^{1+\epsilon}|$ 与 $|\mathcal{C}\setminus\mathcal{G}_{i-1}|$ 的比值，另一方面分析 $(\mathcal{L}_{i-1},\mathcal{C}^{\dagger})$ 的后代节点所对应的采样范围与集合 $\mathcal{C}\setminus\mathcal{G}_{i-1}$ 之间的关系，其目标是证明 $\mathcal{H}_i^{1+\epsilon}$ 中的用户在 Sampling$(k,\emptyset,\mathcal{C},\emptyset,d)$ 构造的采样范围中占有较大比重，并基于此分析与 f_i^* 距离较近的引导点被选取的概率。

引理 4. 如果 $\mathcal{H}_i^{1+\epsilon}\cap\mathcal{G}_{i-1}=\emptyset$，则 $(\mathcal{L}_{i-1},\mathcal{C}^{\dagger})$ 的一个后代节点（包含其本身）$(\mathcal{L}_{i-1},\mathcal{C}^{\pm})$ 满足 $\mathcal{C}\setminus\mathcal{G}_{i-1}\subseteq\mathcal{C}^{\pm}$ 和 $|\mathcal{H}_i^{1+\epsilon}|\geqslant\epsilon^2(20(k+\ell))^{-1}|\mathcal{C}^{\pm}|$。

证明. 我们首先分析 $\mathcal{H}_i^{1+\epsilon}$ 在 $\mathcal{C}\setminus\mathcal{G}_{i-1}$ 中的比重。给定整数 $j\in[i-1]$，令 l_j 表示 $\mathcal{L}_{i-1}$ 中与 f_j^* 距离最近的用户。由 $\mathcal{G}_{i-1}$ 的定义可以得出，每个用户 $c\in\mathcal{C}\setminus\mathcal{G}_{i-1}$ 都满足

$$d(c,l_j)\geqslant d^{\min}(c,\mathcal{L}_{i-1})>\epsilon\frac{opt}{k|\mathcal{C}_i^*|}\tag{10}$$

由此可知，每个整数 $j\in[i-1]$ 都满足

$$\begin{aligned}&|\mathcal{C}_j^*\setminus\mathcal{G}_{i-1}|\\&<\frac{k|\mathcal{C}_i^*|\,d^{\mathrm{sum}}(\mathcal{C}_j^*\setminus\mathcal{G}_{i-1},l_j)}{\epsilon\cdot opt}\\&\leqslant\frac{k|\mathcal{C}_i^*|\,d^{\mathrm{sum}}(\mathcal{C}_j^*,l_j)}{\epsilon\cdot opt}\\&\leqslant\frac{k|\mathcal{C}_i^*|}{\epsilon\cdot opt}(opt_j+|\mathcal{C}_j^*|\,d(f_j^*,l_j))\\&=\frac{k|\mathcal{C}_i^*|}{\epsilon\cdot opt}(opt_j+|\mathcal{C}_j^*|\,d^{\min}(f_j^*,\mathcal{L}_{i-1}))\end{aligned}\tag{11}$$

其中，第 1 步由不等式 (10) 得出，第 3 步由三角不等式得出，第 4 步由 l_j 的定义得出。将不等式 (11) 的两端在 $j\in[i-1]$ 的范围内求和可以得出，

$$\begin{aligned}&\sum_{j=1}^{i-1}|\mathcal{C}_j^*\setminus\mathcal{G}_{i-1}|\\&<\sum_{j=1}^{i-1}\left(\frac{k|\mathcal{C}_i^*|}{\epsilon\cdot opt}(opt_j+|\mathcal{C}_j^*|\,d^{\min}(f_j^*,\mathcal{L}_{i-1}))\right)\\&\leqslant\frac{k|\mathcal{C}_i^*|}{\epsilon\cdot opt}\left((2+\epsilon)\sum_{j=1}^{i-1}opt_j+\frac{1}{k}\epsilon(i-1)opt\right)\\&<\frac{k|\mathcal{C}_i^*|}{\epsilon\cdot opt}(2+2\epsilon)opt\end{aligned}$$

$$< 4k \mid \mathcal{C}_i^* \mid \epsilon^{-1} \tag{12}$$

其中，第 2 步基于不等式 (5) 得出。因此，

$$\begin{aligned} \mid \mathcal{C} \setminus \mathcal{G}_{i-1} \mid &= \mid \mathcal{O}^* \mid + \sum_{j=1}^{i-1} \mathcal{C}_j^* \setminus \mathcal{G}_{i-1} + \sum_{j=i}^{k} \mid \mathcal{C}_j^* \setminus \mathcal{G}_{i-1} \mid \\ &< \ell + 4k \mid \mathcal{C}_i^* \mid \epsilon^{-1} + \sum_{j=i}^{k} \mid \mathcal{C}_i^* \setminus \mathcal{G}_{i-1} \mid \\ &< \ell + (4\epsilon^{-1}+1)k \mid \mathcal{C}_i^* \mid \end{aligned} \tag{13}$$

其中，第 2 步根据不等式 (12) 得出，第 3 步根据 $\mid \mathcal{C}_1^* \mid \geqslant \mid \mathcal{C}_2^* \mid \geqslant \cdots \geqslant \mid \mathcal{C}_k^* \mid$ 这一假设得出。由此可知，

$$\begin{aligned} \frac{\mid \mathcal{C} \setminus \mathcal{G}_{i-1} \mid}{\mid \mathcal{H}_i^{1+\epsilon} \mid} &= \frac{\mid \mathcal{C}_i^* \mid}{\mid \mathcal{H}_i^{1+\epsilon} \mid} \cdot \frac{\mid \mathcal{C} \setminus \mathcal{G}_{i-1} \mid}{\mid \mathcal{C}_i^* \mid} \\ &< \frac{1+\epsilon}{\epsilon \mid \mathcal{C}_i^* \mid} (\ell + (4\epsilon^{-1}+1)k \mid \mathcal{C}_i^* \mid) \\ &< 10\epsilon^{-2}(k+\ell) \end{aligned} \tag{14}$$

其中，第 2 步根据 $\mathcal{H}_i^{1+\epsilon}$ 的定义和引理 1 以及不等式 (13) 得出。

下面分析 $(\mathcal{L}_{i-1}, \mathcal{C}^\dagger)$ 的后代对应的采样范围。结合 $\mid \mathcal{C}_1^* \mid \geqslant \mid \mathcal{C}_2^* \mid \geqslant \cdots \geqslant \mid \mathcal{C}_k^* \mid$ 这一假设和不等式 (6) 可知，

$$\begin{aligned} &\mathcal{C} \setminus \mathcal{G}_{i-1} \\ &= \{c \in \mathcal{C} : d^{\min}(c, \mathcal{L}_{i-1}) > \epsilon(k \mid \mathcal{C}_i^* \mid)^{-1} opt\} \\ &\subseteq \{c \in \mathcal{C} : d^{\min}(c, \mathcal{L}_{i-1}) > \epsilon(k \mid \mathcal{C}_{i-1}^* \mid)^{-1} opt\} \\ &\subseteq \mathcal{C}^\dagger \end{aligned} \tag{15}$$

给定整数 $s \in [0, \lceil \log \mid \mathcal{C}^\dagger \mid \rceil]$，令 $\mathcal{C}_s^\dagger$ 表示 $\mathcal{C}^\dagger$ 中 $d(c, \mathcal{L}_{i-1})$ 取值最高的 $\lceil 2^{-s} \mid \mathcal{C}^\dagger \mid \rceil$ 个用户 c 组成的集合。等式 (15) 说明 $[0, \lceil \log \mid \mathcal{C}^\dagger \mid \rceil]$ 中存在一个满足 $2 \mid \mathcal{C} \setminus \mathcal{G}_{i-1} \mid \geqslant \mid \mathcal{C}_{\tilde{s}}^\dagger \mid$ 和 $\mathcal{C} \setminus \mathcal{G}_{i-1} \subseteq \mathcal{C}_{\tilde{s}}^\dagger$ 的整数 $\tilde{s}$。由不等式 (14) 可知该整数还满足 $20\epsilon^{-2}(k+\ell) \mid \mathcal{H}_i^{1+\epsilon} \mid \geqslant \mid \mathcal{C}_{\tilde{s}}^\dagger \mid$。此外，$\mathcal{C}_{\tilde{s}}^\dagger$ 的定义和算法 Sampling 在第 7 步和第 8 步进行的递归操作说明 $(\mathcal{L}_{i-1}, \mathcal{C}_{\tilde{s}}^\dagger)$ 是 $(\mathcal{L}_{i-1}, \mathcal{C}^\dagger)$ 的一个后代。结合这一事实与集合 $\mathcal{C}_{\tilde{s}}^\dagger$ 的性质可知，引理 4 成立。　　证毕。

在 $\mathcal{H}_i^{1+\epsilon} \cap \mathcal{G}_{i-1} = \emptyset$ 的情况下，引理 4 说明本节用递归的方式为 f_i^* 引导点的选取构造了较为理想的采样范围。本节基于此证明了 $\varphi(i)$ 在 $\mathcal{H}_i^{1+\epsilon} \cap \mathcal{G}_{i-1} = \emptyset$ 时的正确性。

引理 5. 如果 $\mathcal{H}_i^{1+\epsilon} \cap \mathcal{G}_{i-1} = \emptyset$，则以下事件成立的概率不低于 $\epsilon^2(20(k+\ell))^{-1}$：$(\mathcal{L}_{i-1}, \mathcal{C}^\dagger)$ 的一个后代节点 $(\mathcal{L}_{i-1} \uplus \{c_i\}, \mathcal{C}^\ddagger)$ 满足

$$\begin{aligned} &\sum_{j=1}^{i} \mid \mathcal{C}_j^* \mid d^{\min}(f_j^*, \mathcal{L}_{i-1} \uplus \{c_i\}) \\ &\leqslant (1+\epsilon) \sum_{j=1}^{i} opt_j + \epsilon i k^{-1} opt \end{aligned}$$

和

$$\begin{aligned} &\{c \in \mathcal{C} : d^{\min} c, \mathcal{L}_{i-1} \uplus \{c_i\} > \epsilon(k \mid \mathcal{C}_i^* \mid)^{-1} opt\} \\ &\subseteq \mathcal{C}^\ddagger。 \end{aligned}$$

证明. 令 $(\mathcal{L}_{i-1}, \mathcal{C}^\ddagger)$ 表示引理 4 中声明的后代节点。该节点有一个子节点 $(\mathcal{L}_{i-1} \uplus \{c_i\}, \mathcal{C}^\ddagger)$，其中，$c_i$ 是在 $\mathcal{C}^\ddagger$ 中均匀随机选取的用户。引理 4 和 $\mathcal{H}_i^{1+\epsilon} \cap \mathcal{G}_{i-1} = \emptyset$ 这一假设说明 $\mathcal{H}_i^{1+\epsilon} \subseteq \mathcal{C} \setminus \mathcal{G}_{i-1} \subseteq \mathcal{C}^\ddagger$，且 $\mid \mathcal{H}_i^{1+\epsilon} \mid \geqslant \epsilon^2(20(k+\ell))^{-1} \mid \mathcal{C}^\ddagger \mid$。由此可知，$c_i \in \mathcal{H}_i^{1+\epsilon}$ 成立的概率不低于 $\epsilon^2(20(k+\ell))^{-1}$。如果 $c_i \in \mathcal{H}_i^{1+\epsilon}$ 成立，则 $\mathcal{H}_i^{1+\epsilon}$ 的定义说明 $d(f_i^*, c_i) \leqslant (1+\epsilon) \mid \mathcal{C}_i^* \mid^{-1} opt_i$。结合该不等式与不等式 (5) 可知，

$$\begin{aligned} &\sum_{j=1}^{i} \mid \mathcal{C}_j^* \mid d^{\min}(f_j^*, \mathcal{L}_{i-1} \uplus \{c_i\}) \\ &\leqslant \mid \mathcal{C}_i^* \mid d(f_i^*, c_i) + \sum_{j=1}^{i-1} \mid \mathcal{C}_j^* \mid d^{\min}(f_j^*, \mathcal{L}_{i-1}) \\ &\leqslant 1 + \epsilon \sum_{j=1}^{i} opt_j + \frac{1}{k} \epsilon(i-1) opt \\ &< 1 + \epsilon \sum_{j=1}^{i} opt_j + \frac{1}{k} \epsilon i \cdot opt \end{aligned} \tag{16}$$

下面分析集合 $\{c \in \mathcal{C} : d^{\min}(c, \mathcal{L}_{i-1} \uplus \{c_i\}) > \epsilon(k \mid \mathcal{C}_i^* \mid)^{-1} opt\}$ 和 $\mathcal{C}^\ddagger$ 之间的包含关系。可以得出，

$$\begin{aligned} &\{c \in \mathcal{C} : d^{\min}(c, \mathcal{L}_{i-1} \uplus \{c_i\}) > \epsilon(k \mid \mathcal{C}_i^* \mid)^{-1} opt\} \\ &\subseteq \{c \in \mathcal{C} : d^{\min}(c, \mathcal{L}_{i-1}) > \epsilon(k \mid \mathcal{C}_i^* \mid)^{-1} opt\} \\ &= \mathcal{C} \setminus \mathcal{G}_{i-1} \\ &\subseteq \mathcal{C}^\ddagger \end{aligned} \tag{17}$$

其中，第 1 步根据每对用户 $c, c_i \in \mathcal{C}$ 都满足 $d^{\min}(c, \mathcal{L}_{i-1} \uplus \{c_i\}) \leqslant d^{\min}(c, \mathcal{L}_{i-1})$ 这一事实得到，第 2 步基于 $\mathcal{G}_{i-1}$ 的定义得出，第 3 步基于引理 4 得出。由不等式 (16) 和不等式 (17) 可知，引理 5 正确。　　证毕。

给定整数 $i \in \{2, 3, \cdots, k\}$，引理 3 和引理 5 说明 $\varphi(i)$ 在 $\varphi(i-1)$ 正确的情况下成立。结合这一结论与 $\varphi(1)$ 成立这一事实可知，本节归纳地证明了每个整数 $i \in [k]$ 都满足 $\varphi(i)$。

6　基于引导点的候选解构造算法

令 $\mathcal{L}$ 表示 $\varphi(k)$ 中声明的引导点集合。给定整数 $i \in [k]$，令 $l_i = \arg\min_{c \in \mathcal{L}} d(c, f_i^*)$ 表示 f_i^* 在 $\mathcal{L}$ 中的引导点。本节基于引导点集合求解实例 $\mathcal{I}$。具体来说，我们在 6.1 节中围绕 $\mathcal{L}$ 中的引导点选取候选开设设施、在 6.2 节中选取用户集合中的异常点并构造从剩余用户到开设设施集合的连接映射。

6.1 开设设施选取

结合 $\mathcal{D}^*$ 中每个设施都与其在 $\mathcal{L}$ 中的引导点距离较近这一事实和三角不等式可知，我们可以开设与每个引导点距离最近的设施以得到同 $\mathcal{D}^*$ 较为接近的开设设施集合。通过移除 $\mathcal{O}^*$ 中的用户并将 $\mathcal{D}^*$ 中每个设施对应的簇连接到与其引导点距离最近的开设设施上，我们可以为实例 $\mathcal{I}$ 构造费用较低的近似解。然而，利用这一方式构造的解不一定能满足 $\mathcal{I}$ 的着色约束：在根据与引导点之间的距离选取开设设施时，可能存在被多次选取的设施；这些设施对应的簇会因为合并了最优解中的多个簇而包含颜色相同的用户。针对上述问题，本节基于彩色编码技术避免重复开设设施，如算法 Selection（算法 2）所述。该算法在第 5 步为每个设施随机分配 $[k]$ 中的一个标签，并在第 8 步选取标签互不相同的候选开设设施。具体来说，给定整数 $i \in [k]$ 以及设施 f_i^* 的引导点 l_i，算法 2 将标签为 i 且与 l_i 距离最近的设施作为候选开设设施。令 $\mathbb{D}$ 表示 Selection $(k,\mathcal{F},\mathcal{L},d)$ 构造的集合。以下引理给出了集合 $\mathbb{D}$ 包含满足要求的候选开设设施集合的概率。

算法 2. Selection$(k,\mathcal{F},\mathcal{L},d)$

输入：正整数 k、设施集合 $\mathcal{F}$、引导点集合 $\mathcal{L}=\{l_1,l_2,\cdots,l_k\}$ 和距离函数 d；

输出：一组候选开设设施集合 $\mathbb{D}$；

1. $\mathbb{D} \leftarrow \emptyset$；
2. FOR each $j \in [k^k]$ DO
3. $\quad \mathcal{D} \leftarrow \emptyset$；
4. $\quad$ FOR each $f \in \mathcal{F}$ DO
5. $\quad\quad$ 均匀随机地选取一个整数 $\sigma(f) \in [k]$；
6. $\quad$ FOR each $i \in [k]$ DO
7. $\quad\quad \mathcal{F}_i \leftarrow \{f \in \mathcal{F}:\sigma(f)=i\}$；
8. $\quad\quad \mathcal{D} \leftarrow \mathcal{D} \cup \{\arg\min_{f\in\mathcal{F}_i} d(f,l_i)\}$；
9. $\quad$ IF $|\mathcal{D}|=k$ THEN
10. $\quad\quad \mathbb{D} \leftarrow \mathbb{D} \cup \{\mathcal{D}\}$；
11. RETURN $\mathbb{D}$.

引理 6. 以下事件成立的概率不低于 $1-e^{-1}$：$\mathbb{D}$ 中存在满足 $d(f_i,l_i) \leqslant d(f_i^*,l_i)\ \forall\, i \in [k]$ 且不包含重复元素的集合 $\{f_1,f_2,\cdots,f_k\}$。

证明. Selection $(k,\mathcal{F},\mathcal{L},d)$ 在第 5 步为每个设施 $f \in \mathcal{F}$ 随机生成一个标签 $\sigma(f) \in [k]$。每个设施 $f_i^* \in \mathcal{D}^*$ 都满足 $\sigma(f_i^*)=i$ 的概率为 k^{-k}。鉴于 Selection $(k,\mathcal{F},\mathcal{L},d)$ 将第 3－10 步的操作循环执行 k^k 次，等式 $\sigma(f_i^*)=i\ \forall\, i \in [k]$ 在至少一次循环中成立的概率为 $1-(1-k^{-k})^{k^k} > 1-e^{-1}$。对于每个整数 $i \in [k]$，Selection $(k,\mathcal{F},\mathcal{L},d)$ 在第 7 步将标签为 i 的设施划分到集合 $\mathcal{F}_i$ 中，并在第 8 步将 $\mathcal{F}_i$ 中与 l_i 距离最近的设施 f_i 作为候选开设设施；在 $\sigma(f_i^*)=i$ 的情况下，由等式 $d(f_i,l_i)=\min_{f\in\mathcal{F}_i} d(f,l_i)$ 可知 $d(f_i,l_i) \leqslant d(f_i^*,l_i)$。此外，每个整数 $i \in [k]$ 和每个设施 $f \in \mathcal{F}_i$ 都满足 $\sigma(f)=i$ 这一事实说明，集合 $\mathcal{F}_1,\mathcal{F}_2,\cdots,\mathcal{F}_k$ 之间不存在交集。由此可知，集合 $\{f_1,f_2,\cdots,f_k\}$ 不包含重复元素。因此，引理 6 成立。 证毕。

6.2 候选解构造

给定实例 $\mathcal{I}$ 的开设设施集合，本节基于最小费用循环问题的求解算法确定异常点集合与用户连接映射。该问题的定义如下。

定义 2.（最小费用循环问题）

输入：最小费用循环问题的一个实例 $(\mathcal{V},\mathcal{A},\mu_1,\mu_2,g)$ 包含以 $\mathcal{V}$ 为点集、$\mathcal{A}$ 为边集的有向图 $(\mathcal{V},\mathcal{A})$，其中，每条边 $e(u,v) \in \mathcal{A}$ 都有一个取值为非负整数的需求 $\mu_1(u,v)$、一个不小于 $\mu_1(u,v)$ 的整数容量 $\mu_2(u,v)$ 和一个非负费用 $g(u,v)$。

输出：实例 $(\mathcal{V},\mathcal{A},\mu_1,\mu_2,g)$ 的最小费用可行解。该实例的一个可行解 h 为每条边 $e(u,v) \in \mathcal{A}$ 分配一个取值为非负整数的流量 $h(u,v) \in [\mu_1(u,v),\mu_2(u,v)]$，使得每个点 $u \in \mathcal{V}$ 都满足 $\sum_{u:e(u,v)\in\mathcal{A}} h(u,v) = \sum_{u:e(v,u)\in\mathcal{A}} h(v,u)$。可行解 h 的费用为 $\sum_{e(u,v)\in\mathcal{A}} g(u,v)h(u,v)$。

给定实例 $\mathcal{I}$ 的一个开设设施集合 $\mathcal{D}=\{f_1,f_2,\cdots,f_k\}$，本节将确定异常点集合 $\mathcal{O}\subseteq\mathcal{C}$ 与用户连接映射 $\tau:\mathcal{C}\setminus\mathcal{O}\to\mathcal{D}$ 的任务归约为最小费用循环问题的以下实例。

（1）令 m 表示 $\mathcal{C}$ 中用户颜色的数量。令 $\mathcal{C}^1,\mathcal{C}^2,\cdots,\mathcal{C}^m$ 表示根据颜色的不同划分 $\mathcal{C}$ 得到的 m 个用户子集。令 $\mathcal{D}^1=\{f_1^1,f_2^1,\cdots,f_k^1\},\cdots,\mathcal{D}^m=\{f_1^m,f_2^m,\cdots,f_k^m\}$ 为复制集合 $\mathcal{D}=\{f_1,f_2,\cdots,f_k\}$ 得到的 m 个设施集合。点集 $\mathcal{V}$ 由 $\mathcal{C}$ 中的用户、$\biguplus_{i=1}^m \mathcal{D}^i$ 中的设施以及额外的三个点 u_1、u_2 和 u_3 组成。

（2）边集 $\mathcal{A}$ 包含边 $e(u_3,u_1)$，其中，$\mu_1(u_3,u_1)=\mu_2(u_3,u_1)=|\mathcal{C}|$，且 $g(u_3,u_1)=0$。

（3）给定用户 $c \in \mathcal{C}$，边集 $\mathcal{A}$ 包含边 $e(c,u_2)$，其中，$\mu_1(c,u_2)=0$，$\mu_2(c,u_2)=1$，且 $g(c,u_2)=0$。$h(c,u_2)=1$ 说明我们将 c 作为异常点（即 $c \in \mathcal{O}$）。

（4）为了保证异常点数量不超过 ℓ（即 $|\mathcal{O}|=\sum_{c\in\mathcal{C}} h(c,u_2) \leqslant \ell$），边集 $\mathcal{A}$ 包含满足 $\mu_1(u_2,u_3)=0$、$\mu_2(u_2,u_3)=\ell$ 和 $g(u_2,u_3)=0$ 的边 $e(u_2,u_3)$。

(5) 给定整数 $i\in[m]$ 和 $j\in[k]$ 以及用户 $c\in\mathcal{C}^i$，边集 $\mathcal{A}$ 包含边 $e(c,f_j^i)$，其中，$\mu_1(c,f_j^i)=0$，$\mu_2(c,f_j^i)=1$，且 $g(c,f_j^i)=d(c,f_j)$。$h(c,f_j^i)=1$ 说明我们将 c 连接到 f_j 上（即 $\tau(c)=f_j$）。

(6) 为了保证每个用户 $c\in\mathcal{C}$ 都被连接到一个设施上或被标记为异常点（即 $\sum_{f\in\uplus_{i=1}^m\mathcal{D}^i\cup\{u_2\}}h(c,f)=1$），边集 $\mathcal{A}$ 包含满足 $\mu_1(u_1,c)=\mu_2(u_1,c)=1$ 和 $g(u_1,c)=0$ 的边 $e(u_1,c)$。

(7) 给定整数 $i\in[m]$ 和 $j\in[k]$，为了保证 f_j 对应的簇不包含颜色相同的用户（即 $\sum_{c\in\mathcal{C}^i}h(c,f_j^i)\leqslant 1$），边集 $\mathcal{A}$ 包含满足 $\mu_1(f_j^i,u_3)=0$、$\mu_2(f_j^i,u_3)=1$ 和 $g(f_j^i,u_3)=0$ 的边 $e(f_j^i,u_3)$。

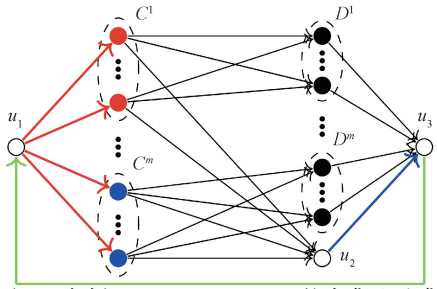


图 7　实例 $(\mathcal{V},\mathcal{A},\mu_1,\mu_2,g)$ 的点集和边集

图 7 中给出了实例 $(\mathcal{V},\mathcal{A},\mu_1,\mu_2,g)$ 的点集和边集，其中：$\mathcal{C}^1,\mathcal{C}^2,\cdots,\mathcal{C}^m$ 为根据用户颜色的不同划分 $\mathcal{C}$ 得到的 m 个子集，$\mathcal{D}^1,\mathcal{D}^2,\cdots,\mathcal{D}^m$ 为复制 $\{f_1,f_2,\cdots,f_k\}$ 得到的 m 个设施集合，与节点 u_2 之间的边流量为 1 的用户被标记为异常点；红色边与绿色边的流量分别被固定为 1 和 $|\mathcal{C}|$，蓝色边的容量和需求分别为 ℓ 和 0，剩余边的容量和需求分别为 1 和 0。本节基于点集 $(\uplus_{i=1}^m\mathcal{C}^i\cup\mathcal{D}^i)\cup\{u_2\}$ 所对应的二部图中边的流量刻画异常点的选取和用户的连接方式。我们可以利用 Orlin[30] 提出的多项式时间算法在 $(|\mathcal{C}|+|\mathcal{D}|)^{O(1)}=(nk)^{O(1)}$ 时间内得到该实例的最优解，并基于此构造令实例 $\mathcal{I}$ 的费用达到最小值的异常点集合与用户连接映射，如引理 7 所述。对于算法 2 输出的每个候选设施集合，本节都基于引理 7 构造对应的异常点集合与用户连接映射以得到实例 $\mathcal{I}$ 的候选解。这一过程在算法 3 中给出。

引理 7. 给定满足 $|\mathcal{C}\cup\mathcal{F}|=n$ 的实例 $\mathcal{I}=((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$ 以及规模为 k 的设施集合 $\mathcal{D}\subseteq\mathcal{F}$，我们可以在 $(nk)^{O(1)}$ 时间内构造满足

$$\sum_{c\in\mathcal{C}\setminus\mathcal{O}}d(c,\tau(c))=\min_{(D,O',\tau')\in\mathrm{Fea}(\mathcal{I},D)}\sum_{c\in\mathcal{C}\setminus\mathcal{O}'}d(c,\tau'(c))$$

的异常点集合 $\mathcal{O}\subseteq\mathcal{C}$ 和映射 $\tau:\mathcal{C}\setminus\mathcal{O}\to\mathcal{D}$，其中，$\mathrm{Fea}(\mathcal{I},\mathcal{D})$ 表示以 $\mathcal{D}$ 为开设设施集合的实例 $\mathcal{I}$ 可行解集合。

算法 3. Construction$(\mathcal{I},\mathbb{D})$

输入：着色 (k,ℓ)-中值问题的实例 $\mathcal{I}=((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$ 和一组候选开设设施集合 $\mathbb{D}$；

输出：实例 $\mathcal{I}$ 的候选解集合 $\mathbb{S}$；

1. $\mathbb{S}\leftarrow\emptyset$；
2. FOR each $\mathcal{D}\in\mathbb{D}$ DO
3. 　令 $\mathcal{O}$ 和 $\tau:\mathcal{C}\setminus\mathcal{O}\to\mathcal{D}$ 分别表示利用引理 7 为实例 $\mathcal{I}$ 和集合 $\mathcal{D}$ 构造的异常点集合和用户连接映射；
4. 　$\mathbb{S}\leftarrow\mathbb{S}\cup\{(\mathcal{D},\mathcal{O},\tau)\}$；
5. RETURN $\mathbb{S}$.

7　着色 (k,ℓ)-中值问题的求解算法

本节结合第 5 节中的引导点挖掘算法和第 6 节中基于引导点的候选解构造算法提出着色 (k,ℓ)-中值问题的近似算法，如算法 4 所述。该算法首先循环调用算法 Sampling 以构造一组候选引导点集合 $\mathbb{L}$。对于 $\mathbb{L}$ 中的每个集合，算法 4 都基于算法 Selection 和 Construction 为实例 $\mathcal{I}$ 构造对应的候选解。最后，算法 4 返回候选解集合中费用最低的解。引理 8 中给出了该算法的性能保证。

算法 4. 着色 (k,ℓ)-中值问题的求解算法

输入：着色 (k,ℓ)-中值问题的实例 $\mathcal{I}=((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$ 和常数 $\epsilon\in(0,1)$；

输出：实例 $\mathcal{I}$ 的近似解 $(\mathcal{D}^\dagger,\mathcal{O}^\dagger,\tau^\dagger)$；

1. $\mathbb{L}\leftarrow\emptyset,\mathbb{S}^\dagger\leftarrow\emptyset$;
2. FOR each $i\in[(20(k+\ell)\epsilon^{-2})^k]$ DO
3. 　Sampling $(k,\emptyset,\mathcal{C},\mathbb{L},d)$；
4. FOR each $\mathcal{L}\in\mathbb{L}$ DO
5. 　令 $\mathbb{D}$ 为 Selection $(k,\mathcal{F},\mathcal{L},d)$ 构造的集合；
6. 　令 $\mathbb{S}$ 为 Construction $(\mathcal{I},\mathbb{D})$ 构造的候选解集合；
7. 　$\mathbb{S}^\dagger\leftarrow\mathbb{S}^\dagger\cup\mathbb{S}$；
8. RETURN $(\mathcal{D}^\dagger,\mathcal{O}^\dagger,\tau^\dagger)\leftarrow\arg\min_{(\mathcal{D},\mathcal{O},\tau)\in\mathbb{S}^\dagger}\sum_{c\in\mathcal{C}\setminus\mathcal{O}}d(c,\tau(c))$.

引理 8. 给定满足 $|\mathcal{C}\cup\mathcal{F}|=n$ 的实例 $\mathcal{I}=((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$ 和常数 $\epsilon\in(0,1)$，算法 4 的时间复杂度为 $((k+\ell)\epsilon^{-1})^{O(k)}n^{O(1)}$，其构造的解 $(\mathcal{D}^\dagger,\mathcal{O}^\dagger,\tau^\dagger)$ 满足 $\sum_{c\in\mathcal{C}\setminus\mathcal{O}^\dagger}d(c,\tau^\dagger(c))\leqslant(3+4\epsilon)opt$ 的概率不低于 $(1-\mathrm{e}^{-1})^2$，其中，opt 为实例 $\mathcal{I}$ 最优解的费用。

证明. 给定实例 $\mathcal{I}=((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$ 和常数 $\epsilon\in(0,1)$，令 $\mathbb{S}^\dagger$ 表示算法 4 构造的候选解集

合，并令 $(\mathcal{D}^{\dagger},\mathcal{O}^{\dagger},\tau^{\dagger})$ 表示算法 4 返回的解。我们首先分析算法 4 的近似比。由 $\varphi(k)$ 可知，Sampling $(k,\emptyset,\mathcal{C},\emptyset,d)$ 构造的集合 $\mathbb{L}$ 中包含满足

$$\sum_{i=1}^{k}|\mathcal{C}_i^*|d^{\min}(f_i^*,\mathcal{L})\leqslant(1+2\epsilon)opt \quad (18)$$

的引导点集合 $\mathcal{L}$ 的概率不低于 $(20(k+\ell)\epsilon^{-2})^{-k}$。由于算法 4 在第 3 步将 Sampling 循环调用 $(20(k+\ell)\epsilon^{-2})^{k}$ 次，满足不等式 (18) 的引导点集合被其成功构造的概率可以被提升 $1-(1-(20(k+\ell)\epsilon^{-2})^{-k})^{(20(k+\ell)\epsilon^{-2})^{k}}>1-\mathrm{e}^{-1}$。给定满足该不等式的引导点集合 $\mathcal{L}$ 和整数 $i\in[k]$，令 l_i 表示 $\mathcal{L}$ 中与 f_i^* 距离最近的引导点。算法 4 在第 5 步和第 6 步调用 Selection 和 Construction 为 $\mathcal{L}$ 构造对应的候选解集合 $\mathbb{S}$。引理 6 说明，$\mathbb{S}$ 有不低于 $1-\mathrm{e}^{-1}$ 的概率包含对于每个整数 $i\in[k]$ 都满足

$$d(f_i,l_i)\leqslant d(f_i^*,l_i) \quad (19)$$

的解 $(\mathcal{D}',\mathcal{O}',\tau')$，其中，$\mathcal{D}'=\{f_1,f_2,\cdots,f_k\}$。可以得出，

$$\begin{aligned}
&\sum_{c\in\mathcal{C}\setminus\mathcal{O}'}d(c,\tau'(c))\\
&\leqslant\sum_{i=1}^{k}d^{\mathrm{sum}}(\mathcal{C}_i^*,f_i)\\
&\leqslant\sum_{i=1}^{k}(d^{\mathrm{sum}}(\mathcal{C}_i^*,f_i^*)+|\mathcal{C}_i^*|d(f_i^*,l_i)\\
&\quad+|\mathcal{C}_i^*|)d(l_i,f_i)\\
&\leqslant\sum_{i=1}^{k}(d^{\mathrm{sum}}(\mathcal{C}_i^*,f_i^*)+2|\mathcal{C}_i^*|d(f_i^*,l_i))\\
&\leqslant(3+4\epsilon)opt
\end{aligned} \quad (20)$$

其中，第 1 步基于引理 7 得出，第 2 步根据三角不等式得出，第 3 步基于不等式 (19) 得出，第 4 步基于 $\sum_{i=1}^{k}d^{\mathrm{sum}}(\mathcal{C}_i^*,f_i^*)=opt$ 这一事实和不等式 (18) 得出。因此，

$$\begin{aligned}
&\sum_{c\in\mathcal{C}\setminus\mathcal{O}^{\dagger}}d(c,\tau^{\dagger}(c))\\
&=\min_{(D,\mathcal{O},\tau)\in\mathbb{S}^{\dagger}}\sum_{c\in\mathcal{C}\setminus\mathcal{O}}d(c,\tau(c))\\
&\leqslant\sum_{c\in\mathcal{C}\setminus\mathcal{O}'}d(c,\tau'(c))\\
&\leqslant(3+4\epsilon)opt
\end{aligned} \quad (21)$$

其中，第 1 步根据算法 4 的第 8 步操作得出，第 3 步基于不等式 (20) 得出。不等式 (21) 说明，在满足不等式 (18) 的引导点集合和满足不等式 (19) 的解存在的情况下 (其概率不低于 $(1-\mathrm{e}^{-1})^2$)，算法 4 所得解的近似比为 $3+4\epsilon$。

下面分析算法 4 的时间复杂度。由引理 2 可知，算法 4 在第 3 步将 Sampling 循环调用 $(20(k+\ell)\epsilon^{-2})^{k}$ 次所需时间不超过 $((k+\ell)\epsilon^{-1})^{O(k)}n$，其构造的集合 $\mathbb{L}$ 规模不超过 $((k+\ell)\epsilon^{-1})^{O(k)}n$。对于 $\mathbb{L}$ 中的每个候选引导点集合，算法 4 通过调用 Selection 在 $O(nk^{k+1})$ 时间内构造 k^k 个候选开设设施集合，并通过调用 Construction 在不超过 $n^{O(1)}k^{O(k)}$ 时间内 (由引理 7 得出) 构造规模为 k^k 的候选解集合。因此，算法 4 构造 $\mathbb{S}^{\dagger}$ 所需时间为 $((k+\ell)\epsilon^{-1})^{O(k)}n^{O(1)}$，且 $\mathbb{S}^{\dagger}$ 的规模不超过 $((k+\ell)\epsilon^{-1})^{O(k)}n$。此外，算法 4 寻找 $\mathbb{S}^{\dagger}$ 中费用最低的解所需时间为 $O(|\mathcal{C}||S^{\dagger}|)\leqslant((k+\ell)\epsilon^{-1})^{O(k)}n^{O(1)}$。综上所述，算法 4 的时间复杂度为 $((k+\ell)\epsilon^{-1})^{O(k)}n^{O(1)}$。由此可知，引理 8 成立。 证毕。

引理 8 说明算法 4 是时间复杂度为 $((k+\ell)\epsilon^{-1})^{O(k)}n^{O(1)}$ 的 $(3+4\epsilon)$-近似算法。令 $\epsilon=\varepsilon/4$，则该算法的时间复杂度和近似比分别为 $((k+\ell)\varepsilon^{-1})^{O(k)}n^{O(1)}$ 和 $3+\varepsilon$。

定理 1. 给定满足 $|\mathcal{C}\cup\mathcal{F}|=n$ 的着色 (k,ℓ)-中值问题实例 $((\mathcal{X},d),\mathcal{C},\mathcal{F},k,\ell,\eta)$ 以及常数 $\varepsilon\in(0,1)$，存在时间复杂度为 $((k+\ell)\varepsilon^{-1})^{O(k)}n^{O(1)}$ 且近似比为 $3+\varepsilon$ 的随机近似算法。

8 总 结

本文以 k 和 ℓ 为固定参数，为着色 (k,ℓ)-中值问题提出了时间复杂度为 $((k+\ell)\varepsilon^{-1})^{O(k)}n^{O(1)}$ 的 $(3+\varepsilon)$-近似算法。这是关于该问题的第一个具有近似保证的固定参数时间求解算法。目前，人们只在以 k 为唯一固定参数的情况下证明了着色 (k,ℓ)-中值问题的固定参数时间近似下界[14]。我们还不能排除在 ℓ 不作为固定参数时得到相同近似结果的可能性。因此，一个值得探索的方向是证明着色 c-中值问题相对于 k 和 ℓ 的多元参数复杂性，或在以 k 为唯一固定参数的情况下尝试设计着色 (k,ℓ)-中值问题的固定参数近似算法。鉴于本文算法与人们在不考虑着色约束的情况下为 (k,ℓ)-中值问题提出的固定参数近似算法[9]有相同的时间复杂度和近似比，这一方向上的突破还依赖于对 (k,ℓ)-中值问题的进一步探索。

参 考 文 献

[1] Guha S, Khuller S. Greedy strikes back: Improved facility location algorithms. Journal of Algorithms, 1999, 31(1): 228-248

[2] Gowda K N, Pensyl T W, Srinivasan A, et al. Improved bi-point rounding algorithms and a golden barrier for k-median//

Proceedings of the 34th ACM-SIAM Symposium on Discrete Algorithms. Florence, Italy, 2023: 987-1011

[3] Li S, Svensson O. Approximating k-median via pseudo-approximation. SIAM Journal on Computing, 2016, 45(2): 530-547

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

[5] Ding H, Xu J H. Chromatic k-mean clustering in high dimensional space. CoRR, 2012, abs/1204.6699

[6] Ding H, Xu J H. Solving the chromatic cone clustering problem via minimum spanning sphere//Proceedings of the 38th International Colloquium on Automata, Languages and Programming. Zurich, Switzerlan, 2011: 773-784

[7] Li J, Yi K, Zhang Q. Clustering with diversity//Proceedings of the 37th International Colloquium on Automata, Languages and Programming. Bordeaux, France, 2010: 188-200

[8] Arkin E M, Díaz-Báñez J M, Hurtado F, et al. Bichromatic 2-center of pairs of points. Computational Geometry, 2015, 48(2): 94-107

[9] Chen X R, Han L, Xu D C, et al. k-median/means with outliers revisited: A simple FPT approximation//Proceedings of the 29th International Conference on Computing and Combinatorics. Hawaii, USA, 2023: 295-302

[10] Feldman D, Schulman L J. Data reduction for weighted and outlier-resistant clustering//Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms. Kyoto, Japan, 2012: 1343-1354

[11] Agrawal A, Inamdar T, Saurabh S, et al. Clustering what matters: Optimal approximation for clustering with outliers. Journal of Artificial Intelligence Research, 2023, 78: 143-166

[12] Bhattacharya A, Goyal D, Jaiswal R. On sampling based algorithms for k-means//Proceeding of the 40th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science. Virtual, 2020: 13:1-13:17

[13] Zhang Z, Huang J Y, Feng Q L. Faster approximation schemes for (constrained) k-means with outliers//Proceedings of the 49th International Symposium on Mathematical Foundations of Computer Science. Bratislava, Slovakia, 2024: 84:1-84:17

[14] Cohen-Addad V, Gupta A, Kumar A, et al. Tight FPT approximations for k-median and k-means//Proceedings of the 46th International Colloquium on Automata, Languages and Programming. Patras, Greece, 2019: 42:1-42:14

[15] Gupta A, Moseley B, Zhou R. Structural iterative rounding for generalized k-median problems//Proceedings of the 48th International Colloquium on Automata, Languages, and Programming. Glasgow, Scotland, 2021: 77:1-77:18

[16] Charikar M, Khuller S, Mount D M, et al. Algorithms for facility location problems with outliers//Proceedings of the 12th Annual Symposium on Discrete Algorithms. Washington, USA, 2001: 642-651

[17] Chen K. A constant factor approximation algorithm for k-median clustering with outliers//Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms. San Francisco, USA, 2008: 826-835

[18] Krishnaswamy R, Li S, Sandeep S. Constant approximation for k-median and k-means with outliers via iterative rounding//Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. Los Angeles, USA, 2018: 646-659

[19] Böhm M, Fazzone A, Leonardi S, et al. Fair clustering with multiple colors. CoRR, 2020, abs/2002.07892

[20] Wu D, Feng Q L, Wang J X. Approximation algorithms for fair k-median problem without fairness violation. Theoretical Computer Science, 2024, 985: 114332

[21] Dickerson J P, Esmaeili S A, Morgenstern J H, et al. Doubly constrained fair clustering//Proceedings of the 37th Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 13267-13293

[22] Feng Q L, Zhang Z, Huang Z Y, et al. A unified framework of FPT approximation algorithms for clustering problems//Proceedings of the 31st International Symposium on Algorithms and Computation. Hong Kong, China, 2020: 5:1-5:17

[23] Goyal D, Jaiswal R, Kumar A. FPT approximation for constrained metric k-median/means//Proceedings of the 15th International Symposium on Parameterized and Exact Computation. Hong Kong, China, 2020: 14:1-14:19

[24] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding//Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. New Orleans, USA, 2007: 1027-1035

[25] Bhattacharya A, Jaiswal R, Kumar A. Faster algorithms for the constrained k-means problem. Theory of Computing Systems, 2018, 62(1): 93-115

[26] Ding H, Xu J H. A unified framework for clustering constrained data without locality property. Algorithmica, 2020, 82(4): 808-852

[27] Cohen-Addad V, Li J. On the fixed-parameter tractability of capacitated clustering//Proceedings of the 46th International Colloquium on Automata, Languages, and Programming. Patras, Greece, 2019: 41:1-41:14

[28] Bandyapadhyay S, Fomin F V, Simonov K. On coresets for fair clustering in metric and Euclidean spaces and their applications. Journal of Computer and System Sciences, 2024, 142: 103506

[29] Kong X Y, Zhang Z. Fixed-parameter tractability of capacitated k-facility location. Frontiers of Computer Science, 2023, 17(6): 176408

[30] Orlin J B. A faster strongly polynomial minimum cost flow algorithm. Operations Research, 1993, 41(2): 338-350

CHEN Xiao-Hong, Ph. D., professor, Academician of the Chinese Academy of Engineering. Her research interests include data intelligence and decision intelligence.

ZHANG Zhen, Ph. D., associate professor. His research interests include combinatorial optimization and approximation algorithms.

XU Xue-Song, Ph. D., professor. His research interests include complex system optimization and algorithm optimization.

CHEN Jie, Ph. D., associate professor. His research interests include privacy-preserving computing and computational intelligence.

YUAN Han-Chun, Ph. D., lecturer. His research interests include fixed-parameter tractable algorithms and kernelization.

SHI Feng, Ph. D., associate professor. His research interests include graph theory and fixed-parameter tractable algorithms.

Background

Clustering with outliers generalizes the standard clustering formulation in that it allows the removal of a specified number of outliers from the set of points to be clustered, which is crucial in many domains where data is affected by noise or corruption. However, the added task of identifying which points to exclude significantly increases the complexity compared to the outlier-free counterpart. For instance, there remains a considerable gap between the best-known polynomial-time approximation ratios for the k-median problem and its outlier variant. A commonly adopted strategy to simplify such outlier-aware clustering problems is to assume that both the number of cluster centers (also referred to as opened facilities) and the number of outliers are small relative to the size of the input. More formally, these two quantities are treated as fixed parameters, and the related outlier-aware problems are solved in fixed-parameter tractable (FPT) time.

In this paper, we focus on an extension of the k-median with outliers problem under the chromatic constraint, known as the chromatic (k,ℓ)-median problem. The objective is to open at most k facilities, remove up to ℓ outliers from the client set, and assign each remaining client to an opened facility, such that clients sharing the same color are assigned to distinct facilities, and the total assignment cost is minimized. We give a sampling-based approach to identify a set of clients located near the facilities in an optimal solution. Around these clients, we construct a carefully selected set of candidate facilities and derive a bounded collection of candidate solutions. This yields a $(3+\varepsilon)$-approximation algorithm that runs in $((k+\ell)\varepsilon^{-1})^{O(k)}n^{O(1)}$ time for the chromatic (k,ℓ)-median problem. To the best of our knowledge, this is the first algorithm with a provable approximation ratio for the problem, which matches the approximation and runtime guarantees previously achieved in the case without the chromatic constraint.

This work was supported by National Natural Science Foundation of China under Grant No. 72088101, Major Project of Xiangjiang Laboratory under Grant No. 24XJJCYJ01003, National Natural Science Foundation of China under Grant Nos. 62202161 and 62202160, National Key Research and Development Program of China under Grant No. 2022YFC3302302, Natural Science Foundation of Hunan Province under Grant No. 2023JJ40240, and Scientific Research Project of the Hunan Provincial Department of Education under Grant No. 23B0597.