用于 k中心带离群点聚类的随机贪心算法与可组合核心集*

Hu Ding huding@ustc.edu.cn
中国科学技术大学计算机科学与技术学院，中国安徽

Ruomin Huang hrm@mail.ustc.edu.cn
中国科学技术大学数据科学学院，中国安徽

Kai Liu liukai0010@mail.ustc.edu.cn
中国科学技术大学计算机科学与技术学院，安徽，中国

Haikuo Yu yhk7786@mail.ustc.edu.cn
中国科学技术大学计算机科学与技术学院，安徽，中国

Zixiu Wang wzx2014@mail.ustc.edu.cn
中国科学技术大学计算机科学与技术学院，安徽，中国

arXiv:2301.02814v1 [cs.LG] 7 Jan 2023

摘要

本文研究带离群点的 k-中心聚类问题。该问题在现实世界中有许多重要应用，但离群点的存在会显著增加计算复杂度。尽管过去几十年间已提出许多方法，但要为该问题设计一种复杂度低且具有质量保证的算法，仍颇具挑战。我们的思路受到用于求解普通 k-中心聚类问题的贪心方法 Gonzalez 算法的启发。基于一些新的观察，我们证明，这种简单的随机化贪心策略实际上能够高效处理离群点。我们进一步证明，在倍增度量空间中，这种随机贪心方法还能为该问题生成规模较小的核心集（即使倍增维数未知），从而大幅降低计算复杂度。此外，结合 Guha 等人（2019）提出的部分聚类框架，我们证明核心集方法可用于分布式数据，且通信复杂度较低。实验结果表明，与现有方法相比，我们的算法能够获得近似最优解，并实现更低的复杂度。

Ke关键词：k-中心聚类、离群点、核心集、倍增度量、分布式算法 s

*。本研究部分得到国家重点研发计划项目（资助编号：2021YFA1000900）的支持。本文的初步版本已发表于第 27 届欧洲算法年会（ESA2019）（Ding 等，2019）。

1. 引言

聚类是最基础的问题之一，广泛应用于机器学习和数据挖掘领域（Jain，2010）。给定一个元素集合，聚类的目标是根据元素之间的相似性或差异性，将输入集合划分为若干组。人们已广泛研究了多种聚类模型，例如 k-中心、k-中位数和 k-均值聚类（Awasthi 和 Balcan，2014）。在实践中，数据集往往包含离群点。具体而言，离群点可能出现在空间中的任意位置，例如，对抗性攻击者可以向数据中注入少量精心构造的样本（Biggio 和 Roli，2018）。即使只有少量离群点，也可能严重破坏最终的聚类结果（Chandola 等，2009）。带离群点的聚类问题还与稳健统计（Diakonikolas 等，2019）和离群点移除（Schubert 等，2017）等主题密切相关。与这些主题的关键区别在于，带离群点的聚类侧重于通过排除少量离群点来优化聚类目标函数。

本文聚焦于 k-中心带离群点聚类问题。给定一个含有 n 个顶点的度量空间，以及预先指定的离群点数量 $z < n$，该问题旨在寻找 k 个球，使其覆盖至少 $n - z$ 个顶点，并使球的最大半径最小化。该问题也可在欧几里得空间中定义，此时聚类中心可以是空间中的任意点（即不局限于从输入点中选取）。 k-中心带离群点聚类问题可视为普通 k-中心聚类问题的推广（即离群点数量为 $z = 0$）。普通 k-中心聚类在机器学习中有许多重要应用，例如深度学习（Coleman et al., 2020）、主动学习（Sener and Savarese, 2018）和公平性（Kleindessner et al., 2019）。Gonzalez（1985）以及Hochbaum和Shmoys（1985）提出了普通 k-中心聚类（不含离群点）的2-近似算法，其中“近似比”是所得半径与最优半径之比。另有证明表明，任何低于“2”的近似比都意味着 $P = NP$。

与普通的 k-中心聚类问题相比，解决带离群点的情形所面临的挑战会大大增加。例如，如果事先不知道哪些点是离群点，那么在优化目标时需要考虑 $\binom{n}{z}$种不同的情形。即使 z 是一个常数，这个数量 $\binom{n}{z}$ 也可能相当大。因此，现有算法往往面临计算复杂度高的问题。Charikar 等人（2001）提出了一种适用于任意度量空间的带离群点 k-中心聚类 3-近似算法。该算法的时间复杂度为$O(kn^2 \log\ n)$（在 D维欧几里得空间中则为 $O(kn^2 D \log n)$），相对于输入规模 n呈二次增长。随后，McCutchen 和 Khuller（2008）提出了一种流式（$4+\epsilon$）-近似算法。其时间复杂度为 $O\big(\frac{1}{\epsilon}(kzn + (kz)^2 \log\ \Phi)\big)$，其中 Φ是最优半径与顶点之间最小两两距离的比值（例如，如果 $z = 5\%n$，复杂度便相对于输入规模 n呈二次增长）。de Berg 等人（2021）基于 Charikar 等人（2001）的静态近似算法，提出了滑动窗口模型下的首个流式算法。最近，Chakrabarty 等人（2016）提出了一种度量空间中带离群点 k-中心聚类的 2-近似算法，但该算法需要求解复杂的线性规划模型，且未给出确切的时间复杂度。

显然，当输入数据规模较大时，这些现有算法在实践中难以高效实现。因此，从理论和实践两个角度来看，一个值得探讨的问题是：能否在保留聚类质量保证的同时，降低带离群点 k-中心聚类的计算复杂度?

1.1 我们的贡献

本文的贡献主要有三方面。

(1) 首先，我们证明，一种简单的随机化贪心数据选择策略能够在线性时间复杂度内得到具有质量保证的解（第3节）。我们的思路受到 Gonzalez（1985）提出的贪心方法启发，该方法用于解决普通的 k-中心聚类问题。Gonzalez 算法迭代地贪心选择 k 个点，每次迭代都选取与已选点集合距离最远的点。基于一些新颖的见解，我们证明，这种贪心方法的随机化版本同样适用于带离群点的问题。粗略地说，我们将每一步贪心选择替换为双层的“贪心选择+随机抽样”步骤：以较小的参数 $\epsilon \in (0, 1)$ 选出最远的 $(1+\epsilon)z$个点（而非单个最远点），然后从选出的集合中随机抽取一个样本。如果允许丢弃 $(1 + \epsilon)z$ （略多于预先指定数量 z）个离群点，我们的方法就能在聚类代价（即半径）方面达到“2”的近似比；此外，其时间复杂度与输入规模呈线性关系。我们方法的另一个优点是还可进一步改进为亚线性时间，即时间复杂度可以不依赖于输入数据规模 n。因此，我们的结果在时间复杂度方面显著优于以往的近似算法。

独立于我们的初步工作（Ding 等人，2019年），Bhaskara 等人（2019）提出了一种针对带离群点的 k-中心聚类的类似贪心算法，但其聚类近似比为 $2 + \delta\,(\delta \in (0, 1))$。此外，他们的运行时间能否改进为次线性尚不明确。

(2) 随后，我们研究带离群点的 k-中心聚类的核心集构造问题。给定一个大型数据集 X，“核心集”技术旨在生成一个小得多的集合 S，使其能够近似保留 X 的结构；因此，我们可以在 S 上运行任何现有算法，从而降低总体复杂度（Feldman，2020年）。

我们首先考虑用于构造核心集的均匀采样方法。Charikar 等人（2003年）证明，均匀随机采样技术可用于缩减度量空间中带离群点的 k-中心聚类的数据规模。近期，Huang 等人（2018年）针对欧几里得空间中的问题证明了类似结果。在 4.1节中，我们重新审视 Huang 等人（2018年）的结果，并给出更细致的分析。特别是，我们证明样本量可以缩减 $\frac{1}{\gamma}$ 倍，其中 $\gamma = \frac{z}{n}$。对于 $z \ll n$ 的情形，这一改进可能很重要，例如 $z = \sqrt{n}$。

虽然均匀采样方法非常容易实现，但它并不是标准的核心集，因为它在被丢弃的离群点数量上总会产生不可避免的误差。因此，我们进一步考虑构造一个能够解决这一问题的核心集，但首先需要加入一些较弱的假设。许多真实世界数据集的内在维度较低（Belkin，2003）。例如，图像集通常可以表示为低维流形，尽管图像向量的欧几里得维度可能非常高。“倍增

方法		Size	构造时间
Uniform sampling	Huang 等人（2018）	$\tilde{O}\left(\frac{n^2}{\epsilon^2 z^2}kD\right)$	
	本文（定理 14）	$\tilde{O}\left(\frac{n}{\epsilon^2 z}kD\right)$	$O\left((k+z)\left(\frac{24}{\mu}\right)^{\rho}n\right)$
μ-核心集	Ceccarello 等人（2019）	$O\left((k+z)\left(\frac{24}{\mu}\right)^{\rho}\right)$	
	本文（定理17）	$2z+\tilde{O}\left(\left(\frac{2}{\mu}\right)^{\rho}k\right)$	$\tilde{O}\left(\left(\frac{2}{\mu}\right)^{\rho}kn\right)$

表 1：现有的以及本文提出的用于含有 z 离群点的 k-中心聚类的数据压缩方法。“D”和“ρ”分别表示欧几里得空间的维数和倍增维数。

维数”常用于衡量数据集的内在维度（Talwar，2004）（其形式化定义见第 2 节）。在“低倍增维数”这一假设下，我们证明，前文所述的随机贪心方法可用于构造一种核心集，且不会在离群点数量上产生误差（第 4.2 节）。我们的核心集大小为 $2z+\tilde{O}\big((\frac{2}{\mu})^{\rho}k\big)$，其中 ρ 为倍增维数，$\mu \in (0,1)$ 为衡量核心集质量的小参数；构造时间为 $\tilde{O}((\frac{2}{\mu})^{\rho}kn)$。近期，Ceccarello 等人（2019）也为倍增度量中的 k-中心聚类（含 z个离群点）提供了一种核心集，其核心集大小为 $T=O((k+z)(\frac{24}{\mu})^{\rho})$，构造时间为 $O(nT)$。因此，无论在核心集大小还是构造时间方面，我们的结果都显著优于他们的结果。详情请见表 1。与 Ceccarello 等人（2019）的结果相比，我们方法的另一项优势是，我们只假设给定数据的内点具有较低的倍增维数 $\rho>0$。我们对离群点没有任何假设；也就是说，离群点可以在空间中任意分散（例如，离群点可能由对抗性攻击者添加（Biggio 和 Roli，2018））。我们认为，这一假设涵盖了实践中很大范围的高维实例。

(3) 随着现实世界数据量的快速增长，分布式计算研究受到了广泛关注。近期提出了多种用于带离群点的 k-中心聚类的分布式算法（Malkomes 等人，2015；Guha 等人（2019）；Ceccarello 等人，2019；Li 和 Guo（2018）），但其中大多数的近似比都很大，例如，Li 和 Guo（2018）的算法近似比为 >19。因此，有必要开发一种通信高效的可组合核心集（Indyk 等人，2014），以便在中央服务器上计算精度更高的近似解。也就是说，输入数据被划分并存储在 s 个站点上，每个站点都可以计算一个独立的核心集并将其发送至中央服务器；最后，中央服务器对收集到的核心集的并集计算近似结果。设 B 为对一个点进行编码的信息。直接实现我们在第 4.2 节提出的核心集，通信成本为 $s\Big(2z+O\big((\frac{2}{\mu})^{\rho}k\big)\Big)B$；当 z 较大时，这一成本可能过高（例如，若 $z=5\%n$ 且 $s=10$，成本可能超过 nB）。在第5节中，我们证明，通过采用 Guha 等人（2019）的部分聚类框架，通信成本可以降低至（大致）$\Big(4z+s\cdot O\big((\frac{2}{\mu})^{\rho}k\big)\Big)B$；因此，我们将“$2sz$”这一项降至“$4z$”。据我们所知，这是首个通信高效的可组合

	约	总通信量（B）	轮次	本地时间
Malkomes 等人（2015）	$3\alpha+2$	$s(k+z)$	1	$O((k+z)n_i)$
Guha 等人（2019）	$5\alpha+4$	$sk+z$	2	$O((k+z)n_i)$
Li 和 Guo（2018）	$((5\alpha+4)(1+\mu)\ 1+\mu)$	$O\left(\frac{sk}{\mu}\cdot\frac{\log\Delta}{\mu}\right)$	2	$O\left(n_i^2\cdot\frac{\text{对数}\,\Delta}{\mu}\right)$
Ceccarello 等人（2019）确定性	$3+\mu$	$s(k+z)(\frac{24}{\mu})^\rho$	1	$O((k+z)n_i(\frac{24}{\mu})^\rho)$
Ceccarello 等人（2019）随机化		$(sk+6z+s\text{对数}\ n)(\frac{24}{\mu})^\rho$	1	$\tilde{O}((k+z/s)n_i(\frac{24}{\mu})^\rho)$
本文（定理 21）	$\alpha\times\frac{1+2\mu}{1-2\mu}=\alpha\times(1+O(\mu))$	$4z+\tilde{O}((sk)(\frac{2}{\mu})^\rho)$	2	$\tilde{O}\left(k\left(\frac{2}{\mu}\right)^\rho n_i\log_2 z\right)$

表 2：现有结果以及我们关于带 z 个离群点的分布式 k-中心聚类的结果。“本地时间”列展示了每个站点上的运行时间，其中 n_i 是站点 i上的数据规模。“Δ”和“ρ”分别表示纵横比和倍增维数。“α”是中央服务器上对收集到的核心集的并集运行算法所得的近似比（例如，若运行 Charikar 等人（2001）的 3-近似算法，$\alpha=3$）。Li 和 Guo（2018）的结果是一种双准则近似，会丢弃 $(1+\mu)z$ 个离群点。

用于带离群点的 k-中心聚类的核心集，可保证 $(1+O(\mu))$-近似误差。详情请参见表 2。

1.2 其他相关工作

带离群点的聚类。除上述带离群点的 k-中心聚类相关研究外，近年来还提出了许多针对其他带离群点聚类问题的研究成果。例如，Charikar 等人（2001）、Chen（2008）、Krishnaswamy 等人（2018）和 Friggstad 等人（2018）提出了具有可证明保证的带离群点 k-均值/中位数聚类算法，但由于其复杂度较高，难以实现。此外，人们也研究了不具备可证明保证、但实用性较强的启发式算法，例如 Chawla 和 Gionis（2013）提出的方法。Gupta 等人（2017）利用局部搜索方法，为带离群点的 k-均值聚类提出了一种常数因子近似算法。此外，Bhaskara 等人（2019）和 Deshpande 等人（2020）分别表明，通过修改 k-均值++初始化方法可以提升效果。其他近期的带离群点聚类算法包括 Chen 等人（2018）、Im 等人（2020）和 Chakrabarty 等人（2022）提出的算法。

核心集。Agarwal 等人（2004）开创了对核心集的研究，这项技术已被广泛应用于多个领域的大规模数据集处理。例如，它可用于降低机器学习中聚类和回归问题的计算复杂度（Cohen-Addad 等人，2021；Munteanu 等人，2018）。为处理分布式数据问题，近期提出了“可合并摘要”（Agarwal 等人，2013）和“可组合核心集”（Indyk 等人，2014；Mirrokni 和 Zadimoghaddam，2015）等技术。Aghamolaei 和 Ghodsi（2018）也研究了倍增度量下的可组合核心集，但他们的方法仅适用于普通的 k-中心聚类问题（不含离群点）。

2. 预备知识

我们研究任意度量和欧几里得空间$\mathbb{R}^D$中的带离群点 k-中心问题。设 (X, d) 为一个抽象度量，其中 X 包含 n 个顶点，且 $\mathsf{d}(\cdot,\cdot)$ 为距离函数；稍微滥用一下记号，我们也用函数 d 表示两个子集 $X_1, X_2 \subseteq X$ 之间的最短距离，即 $\mathsf{d}(X_1, X_2) = \min_{p\in X_1,\ q\in X_2} \mathsf{d}(p,q)$。在 $\mathbb{R}^D$ 中，我们用 $||p-q||$ 表示任意两点 p 和 q 之间的欧几里得距离。为简便起见，我们假设可以在 $O(1)$ 时间内求得 X 中任意一对顶点之间的距离；对于欧几里得空间中的问题，计算任意一对点之间距离需要 $O(D)$ 时间。下面，我们介绍本文中会用到的几个重要定义。

定义 1（带离群点的 k-中心聚类）给定一个度量 (X, d) 和两个正整数 k 与 $z < n$，带离群点的 k-中心聚类问题是寻找一个子集 $X' \subseteq X$，其中 $|X'| \geq n - z$，以及 k 个中心 $\{c_1, \cdots, c_k\} \subseteq X$，使得

$$\max_{p\in X'} \min_{1\leq j\leq k} \mathsf{d}(p, c_j)$$

最小化。如果给定 $\mathbb{R}^D$ 中由 n 个点组成的集合 P，该问题是寻找一个子集 $P' \subseteq P$，其中 $|P'| \geq n - z$，以及 k 个中心 $\{c_1, \cdots, c_k\} \subset \mathbb{R}^D$，使得 $\max_{p\in P'} \min_{1\leq j\leq k} ||p - c_j||$ 最小化。

本文始终用 $X_{\mathtt{opt}}$ 表示 X 的一个大小为 $n - z$ 的子集，该子集可得到最优解。此外，设 $\{C_1, \cdots, C_k\}$ 为构成 $X_{\mathtt{opt}}$ 的 k 个簇，所得聚类代价为 $r_{\mathtt{opt}}$；也就是说，每个 C_j 都由一个半径为 $r_{\mathtt{opt}}$ 的独立球覆盖。

通常，带离群点的优化问题很难求解。因此，我们通常会放宽目标，允许移除略多于预先指定数量的离群点。事实上，许多关于带离群点的聚类问题的研究此前也采用了相同的放宽思想（Charikar 等，2003；Huang 等，2018；Li 和 Guo，2018）。因此，我们引入定义 2。为方便起见，下面的定义 2 和定义 3 仅针对度量空间进行描述。事实上，这些定义可以很容易地针对欧几里得空间中的问题进行修改。

定义 2（$(k, z)_\epsilon$-中心聚类）设 (X, d) 是一个带有 z 个离群点的 k-中心聚类实例，且 $\epsilon \geq 0$。$(k, z)_\epsilon$-中心聚类旨在从 X 中寻找一个子集 X'，其中$|X'| \geq n - (1 + \epsilon)z$，使得定义 1 中在 X' 上对应的聚类代价最小。

(i) 给定一组聚类中心（其数量可能大于 k），我们定义聚类代价

$$\phi_\epsilon(X, A) := \min\Big\{ \max_{p\in X'} \min_{c\in A} \mathsf{d}(p, c) \mid X' \subseteq X, |X'| \geq n - (1+\epsilon)z \Big\}.$$

(ii) 若 $|A| = k$ 和 $\phi_\epsilon(X, A) \leq \alpha r_{\mathtt{opt}}$ 满足 $\alpha > 0$[1]，则集合 A 称为 α-近似；若 $|A| = \beta k$ 满足 $\beta > 1$，则集合 A 称为 (α, β)-近似。

1. 由于我们舍弃了超过 z 个离群点，因此近似比可能为 $\alpha < 1$，即 $\phi_\epsilon(X, A) < r_{\mathtt{opt}}$。

显然，定义 1 中的问题是 $(k, z)_\epsilon$-中心聚类（其中$\epsilon = 0$）的一个特例。此外，定义 1 和定义 2 可以自然地扩展到加权情形：每个顶点 p 都有一个非负权重 w_p，离群点的总权重应等于 z。由此，我们得到核心集的以下定义。

定义 3（核心集）给定一个小参数 $\mu \in (0, 1)$ 和一个包含 z 个离群点的 k-中心聚类实例 (X,d)，若集合 $S \subseteq X$ 中的每个顶点都被赋予一个非负权重，且对于任意由 k 个顶点构成的集合 $H \subseteq X$，$\phi_0(S, H) \in (1 \pm \mu)\phi_0(X, H)$，则称 $S \subseteq X$ 是 X 的 μ-核心集。

给定一个大规模实例 (X, d)，我们可以在其核心集 S 上运行现有算法，以计算 X 的近似解。如果 $|S| \ll n$，则可以显著缩短运行时间。形式化地，我们有如下命题（证明见附录A）。

命题 4 若集合 H 能为 μ 的 α-核心集 S 产生一个 α-近似解，则它能为 X 产生一个 $\alpha \times \frac{1+\mu}{1-\mu}$-近似解。

如前所述，我们还考虑倍增维数较低的情形。粗略地说，倍增维数描述了度量的扩张率。对于任意 $p \in X$ 和 $r \geq 0$，我们用 $\mathrm{Ball}(p, r)$ 表示以 p 为中心、半径为 r 的球。

定义 5（倍增维数）度量空间 (X, d) 的倍增维数是满足以下条件的最小数 $\rho > 0$：对于任意 $p \in X$ 和 $r \geq 0$，$X \cap \mathrm{Ball}(p, 2r)$ 总能被至多 2^ρ 个半径为 r 的球覆盖。

本文其余部分组织如下。第3节中，我们介绍带离群点的 k-中心聚类的常数因子近似，其主要思想是基于 Gonzalez 算法的随机贪心方法。第4节中，我们分别研究欧几里得空间和倍增度量中带离群点的 k-中心聚类核心集。第5节中，我们证明，第4节提出的核心集可以通过一种通信高效方式在分布式环境下构建。第6节中，我们进行实验，以评估所提出的方法。

3. 用于（k, z）$_\epsilon$-中心聚类的随机贪心算法

为完整起见，我们先简要介绍 Gonzalez（1985）提出的普通 k-中心聚类算法。首先，从 X中任意选取一个顶点，然后迭代地选取接下来的 $k-1$ 个顶点：每隔 j步（$2 \leq j \leq k$）选取与已选 $j-1$ 个顶点的最小距离最大的顶点；最后，将每个输入顶点分配给这些已选 k 个顶点中距离它最近的顶点。这种贪心策略可为 k-中心聚类提供 2-近似；该算法也适用于欧几里得空间中的问题，并能达到相同的近似比。在本节中，我们将证明，Gonzalez 算法的随机化版本能够以质量保证求解（k, z）$_\epsilon$-中心聚类问题。

算法1 双准则近似算法

输入：带有 z 个离群点的度量 k-中心聚类实例（X, d），以及 $|X| = n$；参数 $\epsilon > 0$、$\eta \in$ （0, 1/2）和 $t \in \mathbb{Z}^+$。1. 令 $\gamma = z/n$，并初始化集合 $E = \emptyset$。2. 首先，$j = 1$；从 X 中随机选取 $\frac{1}{1-\gamma}\log\frac{1}{\eta}$ 个顶点，并将它们加入 E。3. 重复执行以下步骤，直到 $j = t$：(a) 更新 $j = j + 1$，并令 Q_j 为 X 中距离 E 最远的（$1+\epsilon$）z 个顶点组成的子集（对于每个顶点 $p \in X$，其到 E 的距离定义为 $\min_{q\in E} \mathrm{d}(p, q)$）。(b) 从 Q_j 中随机选取 $\frac{1+\epsilon}{\epsilon}\log\frac{1}{\eta}$ 个顶点，并将它们加入 E。

输出 E。

3.1 $(2, O(\frac{1}{\epsilon}))$-近似

我们考虑返回多于 k 个簇中心的双准则近似。我们的总体思路如下。

实现 Gonzalez 算法的主要挑战在于，离群点和内点在 X中混杂在一起。例如，所选顶点（即与已选顶点的最小距离最大的顶点）很可能是离群点，因而聚类质量可能会变得任意差。为了解决这一问题，我们将每一步贪心选择替换为双层“贪心选择+随机抽样”步骤：使用一个较小的参数 $\epsilon \in (0, 1)$，选出距离最远的 $(1+\epsilon)z$个点（而非最远的单个点），然后从这个选出的集合中随机抽取一个样本。这种组合策略能够确保我们从 k 个最优簇中成功抽取足够数量的内点，同时限制抽取的离群点数量。我们在算法1中实现这一思路。为简单起见，在算法中用 γ 表示 z/n 。

定理 6 令 $\epsilon > 0$ 和 $\eta \in (0, 1/2)$。如果我们在算法1中设置 $t = \frac{ck}{1-\eta}$ with $c = 2 + \frac{2}{(1-\eta)}\ln\frac{1}{\eta}$，则其概率至少为 $1 - 2\eta$，$\phi_\epsilon(X, E) \le 2r_{\mathrm{opt}}$。

如果$\frac{1}{\eta}$和 $\frac{1}{1-\gamma}$是常数，则大小为 $|E| = \frac{1}{1-\gamma}\log\frac{1}{\eta} + \frac{1+\epsilon}{\epsilon}\log\frac{1}{\eta} \times t = O(\frac{k}{\epsilon})$。因此，定理 6 表明，$E$是一个 $\left(2, O(\frac{1}{\epsilon})\right)$-近似算法，用于对 X进行 $(k, z)_\epsilon$-中心聚类，其概率至少为 $1 - 2\eta$。为证明定理 6，我们首先需要引理 7 和引理 10。

引理 7：以至少 $1 - \eta$ 的概率，算法1的步骤 2 中的集合 E 包含来自 X_{opt} 的至少一个点。

由于 $|X_{\mathrm{opt}}|/|X| = 1 - \gamma$，通过以下断言即可轻松得到引理 7。

断言 8 设 U 是一个元素集合，$V \subseteq U$ 且 $\frac{|V|}{|U|} = \tau > 0$。给定 $\eta \in (0, 1)$，如果从 U 中随机抽取 $\frac{1}{\tau}\log\frac{1}{\eta}$ 个元素，那么该样本以至少 $1 - \eta$ 的概率包含至少一个来自 V 的元素。

事实上，断言 8 是一个众所周知的结果，此前已在多篇论文中提出（例如 Ding 和 Xu 2014）。由于每个抽取的元素落在 V 中的概率为 τ，我们知道，该样本 S 以概率 $1-(1-\tau)^{|S|}$ 至少包含一个来自 V 的元素。因此，如果我们希望 $1-(1-\tau)^{|S|} \geq 1-\eta$，$|S|$ 至少应为 $\frac{\log 1/\eta}{\log 1/(1-\tau)} \leq \frac{1}{\tau} \log \frac{1}{\eta}$。

回顾一下，$\{C_1, C_2, \cdots, C_k\}$ 是构成 X_{opt}的 k 个簇。用 $\lambda_j(E)$表示在算法1的步骤 3 中第 j轮开始时，与 E 有非空交集的簇的数量。例如，根据引理 7，我们知道 $\lambda_1(E)$至少为 1。显然，如果 $\lambda_j(E) = k$，即对于任意 $1 \leq l \leq k$，$C_l \cap E \neq \emptyset$ 都会利用三角不等式得到一个 2-近似解。

断言 9 如果 $\lambda_j(E) = k$，那么 $\phi_0(X, E) \leq 2r_{\mathsf{opt}}$。

引理 10 在算法1步骤 3的每一轮中，事件 (1) $\mathsf{d}(Q_j, E) \leq 2r_{\mathsf{opt}}$发生；或者以至少 $1-\eta$的概率，事件 (2) $\lambda_j(E) \geq \lambda_{j-1}(E) + 1$ 发生。

证明：假设事件 (1) 不发生，即 $d(Q_j, E) > 2r_{opt}$，接下来证明事件 (2) 发生的概率至少为 $1-\eta$。令 $\mathcal{J}$ 包含所有满足 $E \cap C_l \neq \emptyset$ 的索引 $l \in \{1, 2, \cdots, k\}$。我们声称，对于每个 $l \in \mathcal{J}$，都有 $Q_j \cap C_l = \emptyset$。否则，任意选取 $p \in Q_j \cap C_l$ 和 $p' \in E \cap C_l$；根据三角不等式可知，$\mathsf{d}(p, p') \leq 2r_{\mathsf{opt}}$，这与假设 $\mathsf{d}(Q_j, E) > 2r_{\mathsf{opt}}$ 矛盾。因此，$Q_j \cap X_{\mathsf{opt}}$ 只包含来自 C_l 且满足 $l \notin \mathcal{J}$ 的顶点。注意，离群点的数量为 z。因此，我们有 $|Q_j \setminus X_{\mathsf{opt}}| \leq z$ 和 $\frac{|Q_j \cap X_{\mathsf{opt}}|}{|Q_j|} \geq \frac{\epsilon}{1+\epsilon}$。根据断言 8，如果从 Q_j 中随机选取 $\frac{1}{\eta}$ 个顶点作为样本，概率至少为 $1-\eta$，则样本中至少包含一个来自 $Q_j \cap X_{\mathsf{opt}}$ 的顶点；此外，该顶点必定来自 $\cup_{l \notin \mathcal{J}} C_l$。也就是说，事件 (2) $\lambda_j(E) \geq \lambda_{j-1}(E) + 1$ 发生。

■

I若引理 10 中的事件 (1) 发生，即 $\mathsf{d}(Q_j, E) \leq 2r_{\mathsf{opt}}$，则可推出 t

$$\max_{p \in X \setminus Q_j} \mathsf{d}(p, E) \leq 2r_{\mathsf{opt}};$$

此外，由于 $|Q_j| = (1+\epsilon)z$，我们有 $\phi_\epsilon(X, E) \leq 2r_{\mathsf{opt}}$。接下来，我们假设引理 10 中的事件 (1) 从未发生，并证明当 $j = \Theta(k)$ 时，$\lambda_j(E) = k$ 以常数概率成立。以下思路受 Aggarwal 等人（2009）的启发，该方法为 k-均值聚类实现了双准则近似。我们定义一个随机变量 x_j：若 $\lambda_j(E) = \lambda_{j-1}(E)$，则取值为 $x_j = 1$；若 $\lambda_j(E) \geq \lambda_{j-1}(E) + 1$，则取值为 $x_j = 0$，其中 $j = 1$、2、$\cdots$。因此，根据引理 10 和

$$\sum_{1 \leq s \leq j} (1 - x_s) \leq \lambda_j(E). \tag{1}$$

此外，令 $J_j = \sum_{1 \leq s \leq j} (x_s - \eta)$ 和 $J_0 = 0$。那么，$\{J_0, J_1, J_2, \cdots\}$ 是一个上鞅，且$J_{j+1} - J_j < 1$。根据 Azuma-Hoeffding 不等式（Alon 和 Spencer，2004 年），对于任意 $t \in \mathbb{Z}^+$ 和 $h > 0$，我们有 $\mathrm{Prob}(J_t \geq J_0 + h) \leq e^{-\frac{h^2}{2t}}$。令 $t = \frac{ck}{1-\eta}$满足 $c = 2 + {}_{k}{}^{2}{}_{(1-\eta)} \ln \frac{1}{\eta}$

算法2 2-近似算法

输入：一个度量 k-中心聚类实例 (X,d)，其中有 z 个离群点，且 $|X|=n$；一个参数 $\epsilon>0$。

1. 初始化一个集合 $E=\emptyset$。

2. 令 $j=1$；从 X 中随机选择一个顶点，并将其加入 E。

3. 重复执行以下步骤，直到 $j=k$:

   (a) 更新 $j=j+1$，并令 Q_j 为 X 中距离 E 最远的 $(1+\epsilon)z$ 个顶点组成的子集。(b) 从 Q_j 中随机选择一个顶点，并将其加入 E。

输出 E。

且 $h=(c-1)k$，该不等式意味着

$$\begin{aligned} & \texttt{Prob}\Big(\sum_{1\le s\le t}(1-x_s)\ge t(1-\eta)-h\Big)\ge 1-e^{-\frac{h^2}{2t}} \\ \Longrightarrow\quad & \texttt{Prob}\Big(\sum_{1\le s\le t}(1-x_s)\ge k\Big)\ge 1-e^{-\frac{k(c-1)^2(1-\eta)}{2c}}\ge 1-e^{-(c/2-1)k(1-\eta)} \\ \Longrightarrow\quad & \texttt{Prob}\Big(\sum_{1\le s\le t}(1-x_s)\ge k\Big)\ge 1-\eta. \end{aligned} \tag{2}$$

结合 (1) 和 (2)，我们知道 $\lambda_t(E)\ge k$ 的概率至少为 $1-\eta$。此外，当 $\lambda_t(E)=k$时，根据断言 9，很容易知道 E 是一个 2-近似解。结合引理 7，我们立即得到定理 6，其中总体成功概率至少为 $(1-\eta)^2>1-2\eta$。

时间复杂度。在步骤 3 的每一轮中，都会向 E 添加 $O(\frac{1}{\epsilon})$ 个新顶点，因此更新从 X 中的顶点到 E 的距离需要 $O(\frac{1}{\epsilon}n)$ 时间；要选出集合 Q_j，我们可以使用 Blum 等人（1973）的线性时间选择算法。总的来说，算法1的运行时间为 $O(\frac{k}{\epsilon}n)$。如果给定实例属于 $\mathbb{R}^D$，运行时间将为 $O(\frac{k}{\epsilon}nD)$。

3.2 常数 k下的 2-近似

如果 k 是一个常数，我们证明可以实现单准则 2-近似。实际上，我们采用与第 3.1 节相同的策略，但只运行 k 轮次，每轮仅采样一个顶点。详情请参见算法2。

用 $\{v_1,\cdots,v_k\}$ 表示从 E 中采样的 k 个顶点。实际上，定理 11 的证明与第 3.1 节中的分析类似。唯一的区别在于，引理 10 中事件 (2) $\lambda_j(E)\ge\lambda_{j-1}(E)+1$ 发生的概率变为至少 $\frac{\epsilon}{1+\epsilon}$。另请注意，$v_1\in X_{\mathsf{opt}}$ 以概率 $1-\gamma$ 成立（因为 $\gamma=z/n$）。如果所有这些事件都发生，则要么在 k 步之前得到一个 2-近似（即，对于某个 $j<k$，$\mathsf{d}(E,X\setminus Q_j)\le 2r_{\mathsf{opt}}$），要么 $\{v_1,\cdots,v_k\}$ 分别落入 k 个最优簇 C_1、C_2、$\cdots$、C_k（即，

算法3 算法1的次线性时间实现

输入：一个度量 k-中心聚类实例 (X, d)，含有 z 个离群点，且 $|X| = n$；参数 $\epsilon > 0$、$\eta \in (0, 1/2)$ 和 $t \in \mathbb{Z}^+$。

1. 令 $\gamma = z/n, \sigma = \frac{2}{1+\sqrt{1+\frac{4(1+\epsilon)}{3\epsilon}}}, n' = \sigma^2 \frac{3}{(1+\epsilon)\gamma} \log \frac{4}{\eta}$，并初始化一个集合 $E = \emptyset$。

2. 初始时，$j = 1$；从 X 中随机选取 $\frac{1}{1-\gamma} \log \frac{1}{\eta}$ 个顶点，并将它们添加到 E。

3. 重复执行以下步骤，直到 $j = t$:

   (a) 更新 $j = j+1$；从 X 中均匀抽取 n' 个顶点，并将抽取出的集合记为 A_j。
   (b) 令 $\hat{r}_j$ 为从 A_j 到 E的第 $(1+\sigma)(1+\epsilon)\gamma n'$ 个最远距离。令 $\hat{A}_j = \{p \in A_j \mid \mathrm{d}(p, E) \geq \hat{r}_j\}$。(c) 将 $\hat{A}_j$ 添加到 E。

输出 E。

$\lambda_k(E) = k$)。无论出现哪种情况，我们都能针对 $(k, z)_\epsilon$-中心聚类问题得到一个 2-近似。因此，我们得到以下定理 11。

定理 11 算法2以至少 $(1-\gamma)(\frac{\epsilon}{1+\epsilon})^{k-1}$ 的概率，为 X 上的 $(k, z)_\epsilon$-中心聚类问题返回一个 2-近似。时间复杂度为 $O(kn)$。如果给定实例属于 $\mathbb{R}^D$，时间复杂度将为 $O(knD)$。

要提高定理 11 中的成功概率，只需重复运行该算法。通过取并集界，可以轻松计算成功概率。
推论 12 如果我们运行算法 $2\, O\big(\frac{1}{1-\gamma}(\frac{1+\epsilon}{\epsilon})^{k-1}\big)$ 次，则以常数概率，至少有一次该算法会为 $(k, z)_\epsilon$- 中心聚类问题返回一个 2- 近似。

3.3 算法1的次线性时间实现

实际应用中的输入数据规模 n 可能非常大，因此本节考虑以更低的时间复杂度实现算法 1。我们提出了算法1的一个修改版本，其时间复杂度仅为 $O(\frac{k^2}{\gamma\epsilon^2})$，其中 $\gamma = \frac{z}{n}$。当数据包含强噪声且离群点占 n 的常数比例（例如 $z = 5\%n$ 和 $\frac{1}{\gamma} = 20$）时，该算法的时间复杂度为次线性，且与 n 无关。此外，可以精确保证定理 6 中算法1的质量。

关键观察与分析。回顾算法1的每一轮，我们都需要扫描整个数据集，并选出距离 E 最远的 $(1+\epsilon)z$ 个顶点。因此，每一轮都需要线性时间。我们的关键观察是，实际上可以通过简单的随机抽样来避免这一步。新思路如算法3（步骤 3）所示。我们仍令Q_j 表示第 j 轮中距离 E 最远的 $(1+\epsilon)z$ 个顶点（如算法1中的步骤 3(a)），其中 X

我们从 X 中随机抽取 n' 个顶点，并用 A_j 表示这个样本集合。我们可以将 A_j 中的每个抽样顶点视为一个独立的随机变量 $\in \{0,1\}$：如果某个抽样顶点属于 Q_j，则将其标记为“1”；否则，标记为“0”。令 $\sigma \in (0,1)$。根据切尔诺夫界，我们有

$$\texttt{Prob}\left[|A_j \cap Q_j| \in (1\pm\sigma)(1+\epsilon)\gamma n'\right] \geq 1 - 2e^{-\frac{1}{3}(\sigma^2(1+\epsilon)\gamma n')}. \tag{3}$$

我们从 A_j 中选取距离 E 最远的 $(1+\sigma)(1+\epsilon)\gamma n'$ 个顶点，所选集合记为 $\hat{A}_j$。我们设参数为 $\sigma = \frac{2}{1+\sqrt{1+\frac{4(1+\epsilon)3\epsilon}{\eta}}}$，于是 (3) 右侧变为 $1-\frac{\eta}{2}$。因此，以至少 $1-\frac{\eta}{2}$ 的概率，$|A_j \cap Q_j| \leq (1+\sigma)(1+\epsilon)\gamma n'$。由于 $|\hat{A}_j| = (1+\sigma)(1+\epsilon)\gamma n'$，我们有

$$|\hat{A}_j| \geq |A_j \cap Q_j|. \tag{4}$$

用 r_j 表示距离 $\mathsf{d}(A_j \cap Q_j, E)$。由于 Q_j 是从 X 到 E 距离最远的 $(1+\epsilon)z$ 个顶点的集合，我们有

$$\{p \in A_j \mid \mathsf{d}(p,E) \geq r_j\} = A_j \cap Q_j; \tag{5}$$
$$\{p \in A_j \mid \mathsf{d}(p,E) > r_j\} \subsetneq A_j \cap Q_j. \tag{6}$$

现在我们声称 $\hat{r}_j \leq r_j$，其中 $\hat{r}_j$ 是从 A_j 到 E 的第 $(1+\sigma)(1+\epsilon)\gamma n'$ 远距离（定义见算法3的步骤 3(b)）。否则，若 $\hat{r}_j > r_j$，则由 (6) 以及 $\hat{A}_j = \{p \in A_j \mid d(p,E) \geq \hat{r}_j\}$ 这一事实，可推出 $\hat{A}_j \subseteq \{p \in A_j \mid d(p,E) > r_j\} \subsetneq A_j \cap Q_j$；因此我们有 $|\hat{A}_j| < |A_j \cap Q_j|$，这与 (4) 矛盾。因此有 $\hat{r}_j \leq r_j$；结合 (5)，可得

$$A_j \cap Q_j \subseteq \{p \in A_j \mid \mathsf{d}(p,E) \geq \hat{r}_j\} = \hat{A}_j.$$

因此 $A_j \cap Q_j \subseteq \hat{A}_j \cap Q_j$。另一方面，由于 $A_j \subseteq \hat{A}_j$，很容易得知 $\hat{A}_j \cap Q_j \subseteq A_j \cap Q_j$。因此，

$$A_j \cap Q_j = \hat{A}_j \cap Q_j. \tag{7}$$

此外，再次根据 (3) 可知

$$|A_j \cap Q_j| \geq (1-\sigma)(1+\epsilon)\gamma n' = \frac{1+\epsilon}{\epsilon}\log\frac{2}{\eta} \tag{8}$$

其概率至少为 $1-\frac{\eta}{2}$，其中我们令 $n' = \sigma^2 \frac{3}{(1+\epsilon)\gamma \log\frac{4}{\eta}}$。根据 (7) 和 (8)，可知

$$|\hat{A}_j \cap Q_j| = |A_j \cap Q_j| \geq \frac{1+\epsilon}{\epsilon}\log\frac{2}{\eta}.$$

因此，$\hat{A}_j$ 中至少包含 $\frac{1+\epsilon}{\epsilon}\log\frac{2}{\eta}$ 个来自 Q_j 的顶点。于是我们可以得到与引理 10 类似的结果：在每一轮中，事件“要么 (1) $\mathsf{d}(Q_j,E) \leq 2r_{\texttt{opt}}$，要么 (2) $\lambda_j(E) \geq \lambda_{j-1}(E)+1$”发生的概率至少为 $1-\frac{\eta}{2}-\frac{\eta}{2} = 1-\eta$（回想一下，(3) 中的概率为 $1-\frac{\eta}{2}$，因此总概率至少为 $1-\frac{\eta}{2}-\frac{\eta}{2}$）。

结合定理 6 中相同的超鞅论证，我们得到以下结果。

定理 13 令 $\epsilon > 0$。如果我们为算法3设置 $t = \frac{ck}{1-\eta}$，其中$\underline{c = 2} + k^{2}{}_{(1-\eta)} \ln \frac{1}{\eta}$，则以至少 $1 - 2\eta$ 的概率，$\phi_\epsilon(X, E) \le 2r_{\mathtt{opt}}$。

质量与时间复杂度。我们假设 γ 和 $1/\eta$ 是常数。算法3在每次迭代中向 E 添加 $O(\frac{1}{\sigma^2})$ 个顶点。注意，$\sigma = \Theta(\sqrt{\epsilon})$，这意味着每次迭代向 E添加的顶点数为 $O(\frac{1}{\epsilon})$。因此，算法结束时 $|E| = O(\frac{k}{\epsilon})$。接着，根据定理 13，$E$是一个 $\big(2、 O(\frac{1}{\epsilon})\big)$-近似算法，能够以常数概率求解 X 的 $(k, z)_\epsilon$-中心聚类问题。每一轮中，我们计算A_j的顶点到 E的距离，并选出 $\hat{} A_j$。由于 $|A_j| = O(\frac{1}{\gamma\epsilon})$ 且 $|E| = O(\frac{k}{\epsilon})$，因此算法3每轮计算距离的时间复杂度为$O(\frac{k}{\gamma\epsilon^2})$。选出 $\hat{} A_j$ 需要 $O(|A_j|)$ 时间。总的来说，算法 3 的时间复杂度为 $O(\frac{k^2}{\gamma\epsilon^2})$，与 n 无关。如果给定的距离属于 $\mathbb{R}^D$，时间复杂度将为 $O(\frac{k^2}{\gamma\epsilon^2}D)$。如果输入数据规模 n 较大，算法3可以显著降低时间复杂度，同时保持与算法1相同的聚类质量。

4. 带离群点的 k-中心聚类核心集

本节研究带离群点的 k-中心聚类的核心集构造问题。首先，我们证明，简单的均匀采样方法可以为欧几里得空间中的 $(k, z)_\epsilon$-中心聚类构造一个稍弱的核心集，其中被丢弃的离群点数量从 $(1 + \epsilon)z$ 增加到 $(1 + O(\epsilon))z$。随后，我们考虑倍增度量中的核心集构造。我们证明，即使倍增维数未知，也可以扩展算法1的思路，高效地构造核心集。

4.1 欧几里得空间中的均匀采样

给定一个度量空间 (X, d)，Charikar 等人（2003）证明，我们可以使用随机样本 S 来替代 X。回顾 $\gamma = z/n$。令 $|S| = O(\frac{k}{\epsilon^2\gamma} \ln n)$，并设 E 是 $(k, z)_\epsilon$-中心聚类在 (S, d) 上的一个 α-近似解，则 E 以常数概率是 $(k, z)_{O(\epsilon)}$-中心聚类在 (X, d) 上的一个 α-近似解。在 D 维欧几里得空间中，Huang 等人（2018）证明了一个类似的结果，其中样本量为 $|S| = \tilde{O}(\frac{1}{\epsilon^2\gamma^2}kD)$[2]。本节将证明，Huang 等人（2018）的样本量还可以进一步缩小 $\frac{1}{\gamma}$ 倍，新的样本量为 $\tilde{O}(\frac{1}{\epsilon^2\gamma}kD)$。这一改进对于 $z \ll n$ 的情形可能很重要，例如 $z = \sqrt{n}$。下面我们先重新审视他们的思路，再通过更细致的分析实现这一改进。

设 P 是 $\mathbb{R}^D$ 中一个由 n 个点组成的集合。考虑范围空间 $\Sigma = (P, \Pi)$，其中每个范围 $\pi \in \Pi$ 都是 $\mathbb{R}^D$ 中 k 个球的并集的补集。我们知道，球的 VC 维数为 $O(D)$（Alon 和 Spencer, 2004 年），因此 k 个球的并集的 VC 维数为 $O(kD \log k)$（Blumer 等人，1989 年）。也就是说，范围空间 Σ 的 VC 维数为 $O(kD \log k)$。令 $\epsilon \in (0, 1)$，并将 P 的“ϵ-样本” S 定义如下：

$$\forall \pi \in \Pi, \quad \Big|\frac{|\pi \cap P|}{|P|} - \frac{|\pi \cap S|}{|S|}\Big| \le \epsilon.$$

2. 渐近记号 $\tilde{O}(f) = O\big(f \cdot \mathrm{polylog}(\frac{kD}{\epsilon\gamma})\big)$。

粗略地说， S 是 P 的一种近似，在每个范围 π 内的加性误差不超过给定值。对于 VC 维数为 $d_{\mathtt{vc}}$ 的范围空间，可以通过均匀采样轻松获得一个 ϵ-样本（Alon 和 Spencer，2004 年），其成功概率为 $1-\lambda$，样本量为 $O\big(\frac{1}{\epsilon^2}(d_{\mathtt{vc}}\log\frac{d_{\mathtt{vc}}}{\epsilon}+\log\frac{1}{\lambda})\big)$，适用于任意 $0<\lambda<1$。对于我们的问题，需要将“ϵ-样本”中的“ϵ”替换为 $\epsilon\gamma$，以保证未覆盖点的数量不超过 $\big(1+O(\epsilon)\big)\gamma n$（具体细节见下文）。由于 $d_{\mathtt{vc}}=O(kD\log k)$，样本量为 $\tilde{O}(\frac{1}{\epsilon^2\gamma^2}kD)$（Huang 等，2018）。

事实上，通过更仔细的分析，可以将样本量的前置系数 $\frac{1}{\epsilon^2\gamma^2}$ 进一步降低至 $\frac{1}{\epsilon^2\gamma}$。我们注意到，无须保证每个范围 π 都满足加性误差（如 ϵ-样本的定义所要求的那样）。相反，只要覆盖至少 γn 个点的范围满足乘法误差即可。请注意，当一个范围覆盖的点更多时，乘法误差比加性误差更宽松，因此样本量也会减少。为此，我们采用相对近似（Har-Peled 和 Sharir，2011；Li 等，2001）：令 $S\subseteq P$ 为一个大小为 $\tilde{O}$（$\frac{1}{\epsilon^2\gamma}kD$）的子集，且该子集是均匀随机选取的，则以常数概率，

$$\forall\pi\in\Pi,\ \Big|\frac{|\pi\cap P|}{|P|}-\frac{|\pi\cap S|}{|S|}\Big|\le\epsilon\times\max\Big\{\frac{|\pi\cap P|}{|P|},\gamma\Big\}. \tag{9}$$

我们在下面正式陈述结果。定理 14 表明，如果我们有一个 α-近似算法，就可以在样本 S 上运行该算法，从而得到一个解 E；该解也是针对 P 上的 $(k,z)^{O(\epsilon)}$-中心聚类的 α-近似解。由于 $|S|\ll|P|$，我们可以大幅减少运行时间。

定理 14 设 P 是定义 1 中所述带离群点的 k-中心聚类问题的一个实例，且 $S\subseteq P$ 是从中均匀随机选取的、大小为 $\tilde{O}$（$\frac{1}{\epsilon^2\gamma}kD$）的子集。假设 $\epsilon\le0.5$。令 S 为带离群点的 k-中心聚类问题的一个新实例，其中离群点数量设为 $z'=(1+\epsilon)\gamma|S|$。若 E 是 S 上（k，z'）ϵ-中心聚类的一个 α-近似解，则 E 是 P 上（k，z）$^{O(\epsilon)}$-中心聚类的一个 α-近似解，且这一结论以常数概率成立。

证明 我们假设 S 是 P 的相对近似，且式 (9) 成立（这种情况以常数概率发生）。令 $\mathbb{B}_{\mathtt{opt}}$ 为覆盖由 P 的最优解所确定的 $(1-\gamma)n$ 个点的 k 个球的集合，令 $\mathbb{B}_S$ 为由 S 上（k，z'）ϵ-中心聚类的一个 α-近似解所确定的 k 个球的集合。假设 $\mathbb{B}_{\mathtt{opt}}$（分别为 $\mathbb{B}_S$）中每个球的半径为 $r_{\mathtt{opt}}$（分别为 r_S）。我们分别将 $\mathbb{B}_{\mathtt{opt}}$ 和 $\mathbb{B}_S$ 的补集记为 $\pi_{\mathtt{opt}}$ 和 π_S。

首先，由于 $\mathbb{B}_{\mathtt{opt}}$ 覆盖了 P 中的 $(1-\gamma)n$ 个点，且 S 是 P 的相对近似，因此我们有

$$\frac{\big|\pi_{\mathtt{opt}}\cap S\big|}{|S|}\le\frac{\big|\pi_{\mathtt{opt}}\cap P\big|}{|P|}+\epsilon\times\max\Big\{\frac{|\pi_{\mathtt{opt}}\cap P|}{|P|},\gamma\Big\}=(1+\epsilon)\gamma$$

根据（9）。也就是说，集合球 $\mathbb{B}_{\mathtt{opt}}$ 至少覆盖 $\big(1-(1+\epsilon)\gamma\big)|S|$ 个 S 中的点，因此它是关于带有 z' 个离群点的 k-中心聚类问题的实例 S 的一个可行解。由于 $\mathbb{B}_S$ 是 S 上 $(k,z')_\epsilon$-中心聚类的一个 α-近似解，我们有

$$r_S\le\alpha r_{\mathtt{opt}};\quad |\pi_S\cap S|\le(1+\epsilon)z'=(1+\epsilon)^2\gamma|S|. \tag{10}$$

现在，我们声称

$$\left|\pi_S \cap P\right| \leq \frac{(1+\epsilon)^2}{1-\epsilon}\gamma|P|. \tag{11}$$

假设 (11) 不成立，则由 (9) 可得

$$\left|\frac{|\pi_S \cap P|}{|P|} - \frac{|\pi_S \cap S|}{|S|}\right| \leq \epsilon \times \max\left\{\frac{|\pi_S \cap P|}{|P|}, \gamma\right\} = \epsilon\frac{|\pi_S \cap P|}{|P|}.$$

因此 $\frac{|\pi_S\cap S|}{|S|} \geq (1-\epsilon)\frac{|\pi_S\cap P|}{|P|} > (1+\epsilon)^2\gamma$，这与 (10) 的第二个不等式矛盾，故 (11) 成立。我们假设 $\epsilon \leq 0.5$，因此 $\frac{1}{1-\epsilon} \leq 1+2\epsilon$ and $(1+\epsilon)^2 = 1+O(\epsilon)$。由此，(11) 与 (10) 的第一个不等式共同推出，$\mathbb{B}_S$ 是 P 上 $(k,z)_{O(\epsilon)}$-中心聚类的一个 α-近似解。

■

4.2 倍增度量中的核心集构造

实际上，定理 14 中得到的样本并非定义 3 所述的标准核心集，因为它总会在丢弃离群点的数量上产生误差。本节考虑构造严格满足定义 3 的核心集。我们首先引入以下假设。

假设 15 给定一个带离群点的 k-中心聚类实例 (X, d)，度量 $(X_{\mathtt{opt}}, \mathsf{d})$（即由内点集合构成的度量）具有常数倍增维数$\rho > 0$。

我们对离群点 $X \setminus X_{\mathtt{opt}}$ 没有任何限制。因此，与假设整个 (X, d) 具有常数倍增维数相比，上述假设更加宽松且实用（例如，Ceccarello 等人（2019）先前提出的核心集构造算法假设整个 (X, d) 具有常数倍增维数 ρ）。根据定义 5，我们可以直接得知，每个最优簇 C_j of $X_{\mathtt{opt}}$ 都可以用半径为 $r_{\mathtt{opt}}/2$ 的 2^ρ 个球覆盖（见图 1 左图）。因此，我们可以认为实例 (X, d) 有 $2^\rho k$ 个簇，最优半径至多为 $r_{\mathtt{opt}}/2$。所以，只需在算法 1 中将 k 替换为 $2^\rho k$，即可降低近似比（即所得半径与 $r_{\mathtt{opt}}$ 的比值），从 2 降至 1。

定理 16 若我们为算法1设置 $t = \frac{2^\rho ck}{1-\eta}$ with $c = 2 + \frac{2}{(1-\eta)}\ln\frac{1}{\eta}$，则以至少 $1-2\eta$ 的概率，$\phi_\epsilon(X, E) \leq r_{\mathtt{opt}}$。因此，集合 E 是一个 $\left(1, O(\frac{2^\rho}{\epsilon})\right)$-近似，可用于解决 $(k,z)_\epsilon$-中心聚类问题，时间复杂度为 $O((k+\ln\frac{1}{2\eta})\frac{2^\rho}{\epsilon}n\ln\frac{1}{2\eta})$。

定理 16 是一个预热结果，我们还可以进一步构造用于 k-中心带离群点聚类的核心集。令 $\mu \in (0,1)$，为简便起见，假设 $\log\frac{2}{\mu}$ 是一个整数。若递归应用定义 5，我们知道每个 C_j都被半径为$\frac{\mu}{2}r_{\mathtt{opt}}$的 $2^{\rho\log 2/\mu} = (\frac{2}{\mu})^\rho$ 个球覆盖，而 $X_{\mathtt{opt}}$ 总共被 $(\frac{2}{\mu})^\rho k$ 个这样的球覆盖。见图 1 右图。基于这一观察，我们得到算法 4。

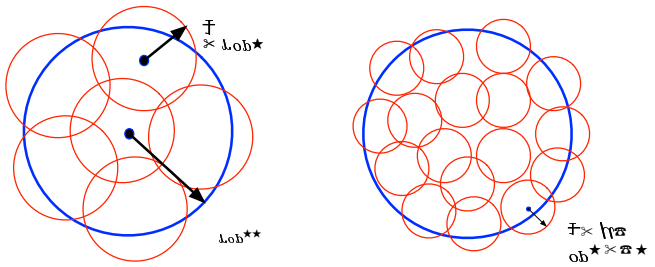


图 1：定理 16 和定理17的示意图。

算法 4 倍增度量中的核心集构造 输入：度量 k 下带有 z 个离群点的 X中心聚类实例 (X, d)，以及 $|X| = n$；参数 $\eta \in (0, 1/2)$ 和 $\mu \in (0, 1)$。1. 令 $l = (\frac{2}{\mu})^\rho k$, $c = 2 + {}^{2}{}_{(1-\eta)}$ln $\frac{1}{\eta}$。2. 设置 $\epsilon = 1$，并运行算法1，进行 $t = \frac{cl}{1-\eta}$ 轮。将最终一轮算法1之后，排除最远的 $2z$ 个顶点后，E 与 X 之间的最大距离记为 $\tilde{r} = \phi_1(X, E)$。3. 令 $X_{\tilde{r}} = \{p \mid p \in X$，且 $\mathsf{d}(x, E) \leq \tilde{r}\}$。4. 对每个顶点 $p \in X_{\tilde{r}}$，将其分配给 E中与之最近的邻居；对每个顶点 $q \in E$，令其权重为分配给它的顶点数量。5. 将 $X \setminus X_{\tilde{r}}$ 加入 E；$X \setminus X_{\tilde{r}}$ 中的每个顶点的权重均为 1。输出 E 作为核心集。

定理 17 令 $\eta \in (0, 1/2)$。以至少 $1 - 2\eta$ 的概率，算法 4 返回一个用于带有 z个离群点的 k- 中心聚类的 μ- 核心集 E。E 的大小至多为 $2z + O\big((\frac{2}{\mu})^\rho (k + \ln \frac{1}{2\eta}) \ln \frac{1}{2\eta}\big)$，构造时间为 $O(n(\frac{2}{\mu})^\rho (k + \ln \frac{1}{2\eta}) \ln \frac{1}{2\eta})$。

证明 与定理 16 类似，我们知道 $|X_{\tilde{r}}| = n - 2z$和 $\tilde{r} \leq 2 \times \frac{\mu}{2} r_{\mathtt{opt}} = \mu r_{\mathtt{opt}}$ 的概率至少为 $1 - 2\eta$。E 的大小为

$$|X \setminus X_{\tilde{r}}| + O\big((\frac{2}{\mu})^\rho (k + \ln \frac{1}{2\eta}) \ln \frac{1}{2\eta}\big) = 2z + O\big((\frac{2}{\mu})^\rho (k + \ln \frac{1}{2\eta}) \ln \frac{1}{2\eta}\big).$$

此外，很容易看出，算法 4 的运行时间为 $O\big((\frac{2}{\mu})^\rho (k + \ln \frac{1}{2\eta}) n \ln \frac{1}{2\eta}\big)$。接下来，我们证明 E 是 X 的一个合格的 μ- 核心集。

对于每个顶点 $q \in E$，用 $w(q)$ 表示 q 的权重；为方便证明，我们将每个 q 视为由 $w(q)$ 个相互重叠的单位权重顶点组成的集合。因此，根据 E 的构造可知，X 与 E 之间存在双射 f，其中

$$\mathsf{d}\left(p, f(p)\right) \leq \tilde{r} \leq \mu r_{\mathtt{opt}}, \quad \forall p \in X. \tag{12}$$

令 $H = \{c_1, c_2, \cdots, c_k\}$ 为 X 的任意 k 个顶点。假设 H 在针对 E （分别为 X）上的含 z 个离群点的 k-中心聚类问题中，分别诱导出 k 个簇 $\{A_1, A_2, \cdots, A_k\}$（分别为 $\{B_1, B_2, \cdots, B_k\}$ 个簇），其中每个 A_j （分别为 B_j）的簇中心为 c_j，对应于 $1 \le j \le k$。分别令 $r_E = \phi_0(E, H)$ 和 $r_X = \phi_0(X, H)$。另外，令 r'_E（分别为 r'_X）为满足以下条件的最小值 r：对于任意 $1 \le j \le k$，$f(B_j) \subseteq \text{Ball}(c_j, r)$（分别为 $f^{-1}(A_j) \subseteq \text{Ball}(c_j, r)$）。我们需要以下断言（证明见 B节）。

命题 18 $|r'_E - r_X| \le \mu r_{\text{opt}}$ 和 $|r'_X - r_E| \le \mu r_{\text{opt}}$。

此外，由于 $\{f(B_1) \cdots, f(B_k)\}$ 也为实例 E形成了 k个簇，其固定簇中心为 H，因此我们知道 $r'_E \ge \phi_0(E, H) = r_E$。同理，我们有 $r'_X \ge r_X$。结合命题18，可得

$$r_X - \mu r_{\text{opt}} \le \underbrace{r'_X - \mu r_{\text{opt}} \le r_E}_{\text{by Claim 18}} \le \underbrace{r'_E \le r_X + \mu r_{\text{opt}}}_{\text{by Claim 18}}.$$

因此 $|r_X - r_E| \le \mu r_{\text{opt}}$，即 $\phi_0(E, H) \in \phi_0(X, H) \pm \mu r_{\text{opt}} \subseteq (1 \pm \mu)\phi_0(X, H)$。因此， E是 (X, d) 的 μ-核心集。 ■

注释19 (1) 值得强调的是，第4.1节中的均匀采样思路无法避免被排除离群点数量上的误差；如果不允许移除超过 z 个离群点（即$\frac{1}{\epsilon} = \infty$），样本量将变为无穷大。但我们在定理17中提出的核心集方法能够保证恰好排除 z 个离群点时的聚类质量。

(2) 核心集大小中的系数“2”实际上可以进一步降低，只需修改算法 4 的步骤 2 中 ϵ 的取值即可（为简便起见，我们设 $\epsilon = 1$ ）。一般来说， E 的大小为

$$(1+\epsilon)z + O\Big(\frac{1}{\epsilon}(\frac{2}{\mu})^{\rho}(k + \ln\frac{1}{2\eta})\ln\frac{1}{2\eta}\Big)$$

构造时间为 $O(n\frac{1}{\epsilon}(\frac{2}{\mu})^{\rho}(k + \ln\frac{1}{2\eta})\ln\frac{1}{2\eta})$。

4.3 未知倍增维数 ρ 时

在算法 4 中，我们运行算法 1 $t = \frac{cl}{1-\eta}$ 轮次。但当倍增维数 ρ 未知时，我们无法确定 l 和 t的取值。我们知道有几种方法可以估计给定数据集的倍增维数（Har-Peled 和 Mendel, 2006）。Ceccarello 等人（2019）也提到，他们的核心集构造方法可以应用于甚至未给出 ρ 的情况。这些思路主要依赖这样一个事实：如果在数据上运行 Gonzalez 的 k- 中心聚类算法，由于倍增度量的性质，所得半径可以显著减小。然而，我们需要强调，在假设 15 下，这些倍增维数估计方法不适用于我们的问题，因为离群点和内点混杂在一起，且只有内点具有倍增度量的良好性质。我们对算法 4 做如下修改。粗略地说，我们将算法 4 的步骤 2 分解为两个子步骤。

(1) 首先，我们运行算法 $1\ \tilde{k}=\frac{ck}{1-\eta}$ 轮次，然后得到半径 $\tilde{r}=\phi_1(E,X)\ \leq 2r_{\mathtt{opt}}$。现在，$X$ 被划分为 k 个簇 $H_1,H_2,\cdots,H_{\tilde{k}}$，并排除 $2z$ 个离群点。每个$H_j\cap X_{\mathtt{opt}}$ 的倍增维数 ρ 均为常数（注意，每个 H_j 也可能包含一些来自 $X\setminus X_{\mathtt{opt}}$的点）。此外，大小为 $\left|\left(\cup_{j=1}^{\tilde{k}}H_j\right)\setminus X_{\mathtt{opt}}\right|\leq z$，这意味着

$$\left|\left(\cup_{j=1}^{\tilde{k}}H_j\right)\cap X_{\mathtt{opt}}\right|=\left|\cup_{j=1}^{\tilde{k}}H_j\right|-\left|\left(\cup_{j=1}^{\tilde{k}}H_j\right)\setminus X_{\mathtt{opt}}\right|\geq n-2z-z=n-3z.$$

˜因此，若将实例 (*X*,d) 视为一个含有 $3z$ 个离群点的 k-中心聚类实例，则最优半径（记为 $r_{\mathtt{opt}}^{(-3z)}$）应至多为 $\tilde{r}$。总之，我们得到 $\tilde{r}$ 的上下界：

$$r_{\mathtt{opt}}^{(-3z)}\leq\tilde{r}\leq 2r_{\mathtt{opt}}.$$

(2) 接着，若我们至多进行 $t=\frac{cl'}{1-\eta}$ 轮算法1的步骤 3（将“z”替换为“$3z$”），其中

$$l'=\left(\frac{r_{\mathtt{opt}}^{(-3z)}}{\frac{1}{4}\mu\tilde{r}}\right)^{\rho}\tilde{k}\leq\left(\frac{r_{\mathtt{opt}}^{(-3z)}}{\frac{1}{4}\mu r_{\mathtt{opt}}^{(-3z)}}\right)^{\rho}\tilde{k}=O\left(\left(\frac{4}{\mu}\right)^{\rho}k\right),$$

所得半径（不包括最远的 $6z$ 个顶点）应至多为

$$2\times\frac{1}{4}\mu\tilde{r}\leq 2\times\frac{1}{2}\mu r_{\mathtt{opt}}=\mu r_{\mathtt{opt}}.$$

接着，我们可以采用定理17证明中的类似思路，证明所得集合E 是一个合格的 μ-核心集。

总之，对于未给定倍增维数 ρ 的情形，我们有算法 5。时间复杂度为 $O(\frac{cl'}{1-\eta}n)$ $=O\big(n(\frac{4}{\mu})^{\rho}(k+\ln\frac{1}{2\eta})\ln\frac{1}{2\eta}\big)$，核心集大小为 $6z+O\big((\frac{4}{\mu})^{\rho}(k+\ln\frac{1}{2\eta})\ln\frac{1}{2\eta}\big)$。

定理20：以至少 $1-2\eta$ 的概率，算法 5 输出一个关于具有 z 离群点的 k-中心聚类的 μ-核心集 E。E的大小至多为 $6z+O\big((\frac{4}{\mu})^{\rho}(k+\ln\frac{1}{2\eta})\ln\frac{1}{2\eta}\big)$，构造时间为 $O((\frac{4}{\mu})^{\rho}(k+\ln\frac{1}{2\eta})$ $n\ln\frac{1}{2\eta})$。

我们可以采用注释19 (2)中的相同思路，将核心集大小缩减至 $3(1+\epsilon)z+O\big(\frac{1}{\epsilon}(\frac{4}{\mu})^{\rho}(k+$ $\ln\frac{1}{2\eta})\ln\frac{1}{2\eta}\big)$，同时，时间复杂度变为 $O\big(\frac{n}{\epsilon}(\frac{4}{\mu})^{\rho}(k+\ln\frac{1}{2\eta})\ln\frac{1}{2\eta}\big)$。

5. 分布式数据的核心集

本节考虑协调器模型下分布式聚类的核心集（˘Duri˘s 和 Rolim，1998）。假设数据 $X=\sqcup_{i=1}^{s}X_i$ 不相交地分布在 $s\geq 2$ 个站点上；所有站点都可以与中央服务器通信。令 $O=\sqcup_{i=1}^{s}O_i$ 为最优解中的离群点集合，其中 $O_i\subset X_i$；并且假设每个 $|O_i|=z_i^*$。令 $X\setminus O=\sqcup_{j=1}^{k}C_j$ 为 k 最优聚类。注意，每个 z_i^* 的值都是未知的。

算法 5 倍增度量中具有未知 ρ 的核心集构造输入：度量 k 上的 z 离群点 X-中心聚类实例 (X, d)，以及 $|X| = n$；参数 $\mu \in (0, 1)$ 和 $\eta \in (0, 1/2)$。1. 令 $\epsilon = 1$，并运行算法 1 $t = \frac{ck}{1-\eta}$ 轮，其中 $c = 2 + k^{2}(1-\eta) \ln \frac{1}{\eta}$。在算法 1 的最后一轮之后，排除最远的 $2z$ 个顶点，记 $\tilde{r} = \phi_1(X, E)$ 为 E 与 X 之间的最大距离。2. 继续执行算法 1 的步骤 3（但将“z”替换为“$3z$”），直到 $\phi_5(X, E) \leq \frac{1}{2}\mu\tilde{r}$（即排除 $6z$ 个离群点）。3. 令 $\tilde{r}' = \phi_5(X, E)$。令 $X_{\tilde{r}'} = \{p \mid p \in X$，且 $\mathsf{d}(p, E) \leq \tilde{r}'\}$。4. 对每个顶点 $p \in X_{\tilde{r}'}$，将其分配给 E 中最近的邻居；对于每个顶点 $q \in E$，令其权重为分配给它的顶点数。5. 将 $X \setminus X_{\tilde{r}'}$ 添加到 E；$X \setminus X_{\tilde{r}'}$ 中的每个顶点权重均为 1。输出 E 作为核心集。

因此，一种直接的方法是在每个 X_i 上计算带 z 个离群点的 k-中心聚类核心集，并将得到的核心集直接发送给中央服务器。令 B 表示编码一个点所需的信息量。显然，这种方法的通信成本为

$$\Big(2sz + s \cdot O\big((\frac{2}{\mu})^{\rho}(k + \log\frac{1}{2\eta})\ln\frac{1}{2\eta}\big)\Big)B, \tag{13}$$

当 z 较大时，该成本可能过高（例如，如果 $z = 5\%n$ 且 $s = 10$，成本可能大于 nB）。

本节将说明，Guha 等人（2019）提出的分布式带离群点的 k-中位数/均值聚类框架，也可以结合我们在第 4.2 节提出的核心集方法，用于解决带 z 个离群点的 k-中心聚类问题；具体而言，（13）中的“$2sz$”这一项可以缩减为“$4z$”。Guha 等人（2019）的总体思路如下。首先，我们需要设计一组数值 $\{z_1, z_2, \ldots, z_s\}$，其中每个 z_i 都是 z_i^* 的上界，适用于 $1 \leq i \leq s$，且它们的和为 $\sum_{i=1}^{s} z_i \leq 2z$。请注意，“$\sum_{i=1}^{s} z_i \leq 2z$”这一要求对于限制总通信成本至关重要。每个站点 i 都在 X_i 上运行带 z_i 个离群点的 k-中心聚类核心集算法，以构造局部核心集。随后，每个站点将局部核心集中的加权点发送给中央服务器。最后，中央服务器汇总这些加权点，形成全局核心集。关键挑战在于计算适用于我们核心集方法的集合 $\{z_1, z_2, \ldots, z_s\}$。

下面我们先介绍一些记号。

1. $r_{\mathsf{opt}}(X_i, k, z_i^*)$：在 X_i 上，带有 z_i^* 个离群点的 k-中心聚类的最优半径。

2. 给定一组二维点 $A = \{(x_1, y_1) \ldots, (x_l, y_l)\} \subset \mathbb{R}^2$，其中 $x_1 < x_2 < \cdots < x_l$，我们定义相应的分段函数 $h_A(\cdot)$，从 $[x_1, \infty)$ 到 $\mathbb{R}$：$h_A(x) = y_i$ 当 $x_i \leq x < x_{i+1}$ 时。这里我们定义 $x_{l+1} = \infty$。示意图见图 2。

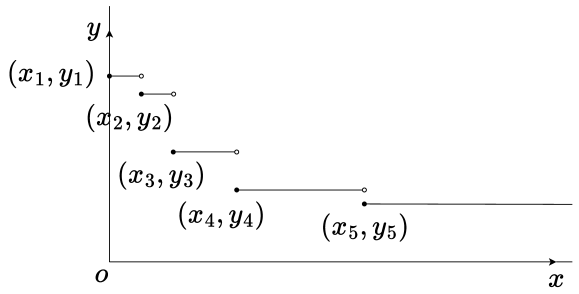


图 2：非递增分段函数 $h_A(\cdot)$ 的示意图，其中 $A = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5)\}$。

对于任意一对 (x_i, y_i) 和 (x_j, y_j)，我们定义它们的字典序：

$$(x_i, y_i) \prec (x_j, y_j) \quad \text{if} \begin{cases} x_i < x_j; \text{ or} \\ x_i = x_j \text{ and } y_i < y_j. \end{cases}$$

定理 21 以至少 $1 - 2s(2 + \log_2 z)\eta$ 的概率，算法 6 返回分布式输入 $X = \sqcup_{i=1}^{s} X_i$ 的 k-中心聚类的 2μ-核心集 E，其中包含 z 个离群点。总通信复杂度为 $\left(4z + O((\frac{2}{\mu})^\rho s(k + \ln \frac{1}{2\eta}) \ln \frac{1}{2\eta})\right) B$，共 2 轮，其中 B 是发送一个点的通信成本。每个站点的运行时间为 $O((\frac{2}{\mu})^\rho (k + \ln \frac{1}{2\eta}) n_i \ln \frac{1}{2\eta} \log_2 z)$。

注 22 在算法 6 中，我们可以将 η 替换为 $\frac{\eta}{2s(2+\log_2 z)}$，从而获得成功概率 $1 - \eta$。通信复杂度将为

$$\left(4z + O\big((\frac{2}{\mu})^\rho s(k + \ln \frac{s \log_2 z}{\eta})(\ln \frac{s \log_2 z}{\eta})\big)\right) B.$$

在证明定理 21 之前，我们先介绍引理 23 和引理 24。

引理 **23** $2r_{\mathtt{opt}} \geq \max_{1 \leq i \leq s} r_{\mathtt{opt}}(X_i, k, z_i^*)$。

请注意，尽管 $X_i \subset X$，站点 i 的最优半径 $r_{\mathtt{opt}}(X_i, k, z_i^*)$ 不一定是 $\leq r_{\mathtt{opt}}$，因为 X 的聚类中心可能不属于 X_i（但对于欧几里得空间中的问题，$r_{\mathtt{opt}}(X_i, k, z_i^*)$ 总是不大于 $r_{\mathtt{opt}}$，因为聚类中心可以是空间中的任意点）。$X_i \setminus O_i$ 是站点 i的内点集合。对于每个最优簇 C_j，$1 \leq j \leq k$，我们可以从 $(X_i \setminus O_i) \cap C_j$ 中任意选取一个内点作为代理簇中心（如果 $(X_i \setminus O_i) \cap C_j = \emptyset$，我们就忽略这个簇）。根据三角不等式，我们知道 $r_{\mathtt{opt}}(X_i, k, z_i^*) \leq 2r_{\mathtt{opt}}$，由此得到引理 23。

在以下分析中，我们假设对于每个 $i = 1$、2、$\cdots$、s，步骤 2(a) 中的函数 h_{A_i} 都是非递增的。实际上，这一假设很容易满足。如果存在一对 $q < q'$，使得 $\tilde{r}_{i,q} > \tilde{r}_{i,q'}$，我们只需将核心集 $E_{i,q}$ 替换为核心集 $E_{i,q'}$，并在算法 6 的步骤 2 中令 $\tilde{r}_{i,q} = \tilde{r}_{i,q'}$。下面的引理说明了算法 6 中所得值 z_1、z_2、$\cdots$、z_s 的关键性质。

算法 6 分布式核心集构造

输入：一个含有 z 个离群点的分布式度量 k-中心聚类实例 (X, d)，以及$X = \sqcup_{i=1}^{s} X_i$；参数 $\eta \in (0, 1/2)$ $\mu \in (0, 1)$。

1. 令 $[z] = \{0, 1, 2, \cdots, z\}$ 和 $\Gamma = \{2^r : 1 \le r \le \lfloor \log_2 z \rfloor, r \in \mathbb{Z}\} \cup \{0, z\}$。在各站点与中央服务器之间执行以下两轮通信。

2.（第 1 轮）在每个站点 i：运行算法 4，获取半径 $\tilde{r}_{i,q}$ 以及核心集 $E_{i,q}$，对每个 $q \in \Gamma$，其中 $\tilde{r}_{i,q}$ 是算法 4 中的半径“$\tilde{r}$”，且将离群点数量设为 q。

(a) 定义集合 $A_i = \{(q, \tilde{r}_{i,q}) \mid q \in \Gamma\}$，并构造相应的分段函数 h_{A_i}；(b) 将函数 h_{A_i} 发送至中央服务器。

3.（第 1 轮）在中央服务器：按字典序降序对 $s(z+1)$ 对 $\{(h_{A_i}(q), i) \mid i = 1, 2, \cdots, s; q \in [z]\}$ 排序；选取第 $(2z+1)$ 大的项，记作“$(h_{A_{i_0}}(q_0), i_0)$”，并将其广播给所有站点。

4.（第 2 轮）在每个站点 i：

(a) If $i \neq i_0$, let $z_i = \min\{q \in [z] : (h_{A_i}(q), i) \prec (h_{A_{i_0}}(q_0), i_0)\}$ (if the set is $\emptyset$, let $z_i = z$);

(b) Else, $i = i_0$, let $z_{i_0} = \min\{q \in \Gamma : h_{A_{i_0}}(q) = h_{A_{i_0}}(q_0)\}$;

(c) Send E_{i,z_i} to the central server.

5.（第 2 轮）在中央服务器：取并集 $E = \cup_{i=1}^{s} E_{i,\ z_i}$ 作为最终核心集。

引理 24 算法 6 的第 4 步中得到的集合 $\{z_1, \ldots, z_s\}$ 是以下极小极大问题的最优解：

$$\begin{aligned} & \min_{q_1,\ldots,q_s} \max_{1 \le i \le s} h_{A_i}(q_i) \\ s.t. \quad & \sum_{i=1}^{s} q_i \le 2z, \\ & q_i \in [z], \quad i = 1, \ldots, s. \end{aligned} \tag{14}$$

证明 W　我们考虑以下两种情况。回顾一下，“i_0”是第.....步中得到的索引　3.

情况（i）：$h_{A_{i_0}}(z_{i_0}) = \max_i h_{A_i}(z_i)$。在这种情况下，根据 z_i的定义以及 $h_{A_i}(\cdot)$非递增这一事实，对于每个 $q = 0$、.....、$z_i - 1$以及每个 $i = 1$、.....、s，都有（$h_{A_{i_0}}(q_0)$ i_0）$\prec (h_{A_i}(q)\ i)$。因此，有 $\sum_{i=1}^{s} z_i$对（$h_{A_i}(q)$ i），其字典序大于（$h_{A_{i_0}}(q_0)$ i_0）。因此，根据（$h_{A_{i_0}}(q_0)$ i_0）的定义，我们

知道 $z_{i_0} \le q_0$，这意味着

$$\sum_{i=1}^{s} z_i \le q_0 + \sum_{i \neq i_0} z_i.$$

右侧的“$q_0 + \sum_{i \neq i_0} z_i$”恰好等于 $2z$，因为它是在算法 6 的步骤 3 中排序序列里排在（$h_{A_{i_0}}$（q_0） i_0）之前的项数。假设引理 24 不成立，那么问题 (14) 存在另一个解 $\{z'_1, \cdots, z'_s\}$，使得

$$\max_i h_{A_i}(z'_i) < h_{A_{i_0}}(z_{i_0}). \tag{15}$$

根据 z_{i_0} 的定义，右侧“$h_{A_{i_0}}$（z_{i_0}）”等于 $h_{A_{i_0}}$（q_0）。因此可得 $h_{A_{i_0}}$（z'_{i_0}） $< h_{A_{i_0}}$（q_0）；由于 $h_{A_{i_0}}(\cdot)$ 是非递增的，我们知道 $z'_{i_0} > q_0$。不失一般性，假设 $\sum_i z'_i = 2z$（同样，由于 $h_{A_i}(\cdot)$ 是非递增的，我们总可以增大 z'_is，直到 $\sum_i z'_i = 2z$）。注意 $q_0 + \sum_{i \neq i_0} z_i = 2z$。因此应存在一个索引 $j \neq i_0$，使得 $z'_j < z_j$。根据 z_j 的定义，我们有（$h_{A_{i_0}}$（q_0） i_0） $\prec$（h_{A_j}（z'_j） j）。因此 h_{A_j}（z'_j） $\ge h_{A_{i_0}}$（q_0） $= h_{A_{i_0}}$（z_{i_0}）。于是 $\max_i h_{A_i}$（z'_i） $\ge h_{A_j}$（z'_j） $\ge h_{A_{i_0}}$（z_{i_0}），这与 (15) 矛盾。

情形（ii）：假设 $h_{A_{i_1}}(z_{i_1}) = \max_i h_{A_i}(z_i)$ 且 $h_{A_{i_1}}(z_{i_1}) > h_{A_{i_0}}(z_{i_0})$。在此情形下，我们在步骤 4(a) 中有 $z_{i_1} = z$。与情形（i）的分析类似，对于 $q = 0, \ldots, z_i - 1$, $i \neq i_0$，有 $(h_{A_{i_0}}(q_0)\ i_0) \prec (h_{A_i}(q)\ i)$；同时，对于 $q = 0, \ldots, z_{i_1}$，有 $(h_{A_{i_0}}(q_0)\ i_0) \prec (h_{A_{i_1}}(q)\ i_1)$。因此，同理可得 $1 + \sum_{i=1}^{s} z_i \le q_0 + 1 + \sum_{i \neq i_0} z_i = 2z$。对于任意可行解 $\{z'_1, \ldots, z'_s\}$，由于 $z'_{i_1} \le z = z_{i_1}$ 且 $h_{A_{i_1}}(\cdot)$ 非递增，我们有

$$\max_i h_{A_i}(z'_i) \ge h_{A_{i_1}}(z'_{i_1}) \ge h_{A_{i_1}}(z_{i_1}) = \max_i h_{A_i}(z_i),$$

这意味着 $\{z_1, \ldots, z_s\}$ 优于解 $\{z'_1, \ldots, z'_s\}$。因此，$\{z_1, \ldots, z_s\}$ 应为问题 (14) 的最优解。 ■

证明（定理 21）我们定义 $\hat{z}_i = \min\{q \in \Gamma : q \ge z^*_i\}$ for $i = 1, 2, \cdots, s$。于是可直接得到 $\max_i r_{\mathtt{opt}}(X_i, k, z^*_i) \ge \max_i r_{\mathtt{opt}}(X_i, k, \hat{z}_i)$，因为 $\hat{z}_i \ge z^*_i$。由于“$r_{i,\ \hat{z}_i}$”是步骤 2 中核心集的半径，因此我们有

$$\tilde{r}_{i,\hat{z}_i} \le \mu r_{\mathtt{opt}}(X_i, k, \hat{z}_i).$$

此外，根据引理 23，我们知道 $2r_{\mathtt{opt}} \ge \max_i r_{\mathtt{opt}}(X_i, k, z^*_i)$。因此我们有 $2r_{\mathtt{opt}} \ge \frac{1}{\mu} \max_i \tilde{r}_{i,\ \hat{z}_i}$。注意，由于 $\hat{z}_i \in \Gamma$，$h_{A_i}(\hat{z}_i) = \tilde{r}_{i,\ \hat{z}_i}$。因此我们有

$$2r_{\mathtt{opt}} \ge \frac{1}{\mu} \max_i h_{A_i}(\hat{z}_i). \tag{16}$$

$\hat{z}_i$ 和 Γ 的定义共同推出 $\sum_i \hat{z}_i \le \sum_i 2z^*_i = 2z$。因此，集合 $\{\hat{z}_1, \ldots, \hat{z}_s\}$ 是问题 (14) 的一个可行解。由引理 24 可得

$$\max_i h_{A_i}(\hat{z}_i) \ge \max_i h_{A_i}(z_i). \tag{17}$$

对于每个 $i \neq i_0$，根据分段函数 $h_{A_i}(\cdot)$的定义，我们知道 $z_i \in \Gamma$ 。由算法 6 的步骤 4(b)，我们知道 $z_{i_0} \in \Gamma$。因此可得

$$h_{A_i}(z_i) = \tilde{r}_{i,z_i} = \phi_1(X_i, E_{i,z_i}),\ i = 1, \ldots, s. \tag{18}$$

结合不等式 (16)、(17) 和 (18)，可得

$$\max_i \phi_1(X_i, E_{i,z_i}) \leq 2\mu \cdot r_{\mathtt{opt}}.$$

与定理17证明中的不等式 (12) 类似，我们可以定义从 X到 E的双射“f”（回忆一下，$E = \cup_i^s E_{i,\ z_i}$），使得

$$\mathtt{d}(p, f(p)) \leq 2\mu \cdot r_{\mathtt{opt}},\ \forall p \in X. \tag{19}$$

上述不等式 (19) 表明，集合 E 是一个合格的 2μ-核心集。

成功概率至少为 $1 - s(2 + \log_2 z)2\eta$，因为每个站点运行算法 4 的次数不超过 $(2 + \log_2 z)$。由于 $\sum_i^s z_i \leq 2z$，从站点发送到中央服务器的点总数不超过 $4z + O((\frac{2}{\mu})^\rho s(k + \ln \frac{1}{2\eta}) \ln \frac{1}{2\eta})$。因此，通信复杂度为 $\left(4z + O((\frac{2}{\mu})^\rho s(k + \ln \frac{1}{2\eta}) \ln \frac{1}{2\eta})\right) B$。每个站点 i的运行时间为 $O(|\Gamma|(\frac{2}{\mu})^\rho(k + \ln \frac{1}{2\eta}) n_i \ln \frac{1}{2\eta}) = O((\frac{2}{\mu})^\rho(k + \ln \frac{1}{2\eta}) n_i \ln \frac{1}{2\eta} \log_2 z)$。 ■

6. 实验

所有实验均在一台 Ubuntu 工作站上进行，该工作站配备 2.40GHz Intel(R) Xeon(R) CPU E5-2680 处理器和 256GB 内存。算法使用 MATLAB R2019b 实现。代码可在 https://github.com/OpsTreadstone/randomized-k-center 获取。

基线方法。我们将算法与两种知名的基线方法“CKM+”（Charikar 等人，2001）和“MK”（McCutchen 和 Khuller，2008）进行比较，同时也与近期提出的算法“BVX”（Bhaskara 等人，2019）进行比较。对于核心集构造问题，我们将算法与“CPP”（Ceccarello 等人，2019）以及均匀采样方法“Uniform”进行比较。

对于分布式环境，我们以“CPP”、“MKC+”（Malkomes 等人，2015）、“GLZ”（Guha 等人，2019）和“LG”（Li 和 Guo，2018）作为基线方法。

所有实验均重复进行了 10 次，我们报告平均结果及标准差。

数据集。我们在四个来自 UCI KDD 档案的真实分类数据集（Dua 和 Graff，2017）上评估算法：Shuttle、Covertype、KDD Cup 1999 和 Poker Hand。Shuttle 数据集（King 等，1995）包含 43, 500 个实例，分属 7 个类别，具有 9 个数值属性。Covertype 数据集包含 581, 012 个实例，分属 7 个类别，具有 54 个连续型和分类型属性。KDD Cup 1999 数据集包含 4, 898, 431 个实例，分属 23 个类别，具有 41 个属性。Poker Hand 数据集包含 1, 025, 010 个实例，分属 10 个类别，具有 10 个属性。对于后三个数据集 Covertype、KDD Cup 1999 和 Poker Hand，我们随机选取 100, 000 个实例，并在所选实例上运行算法。

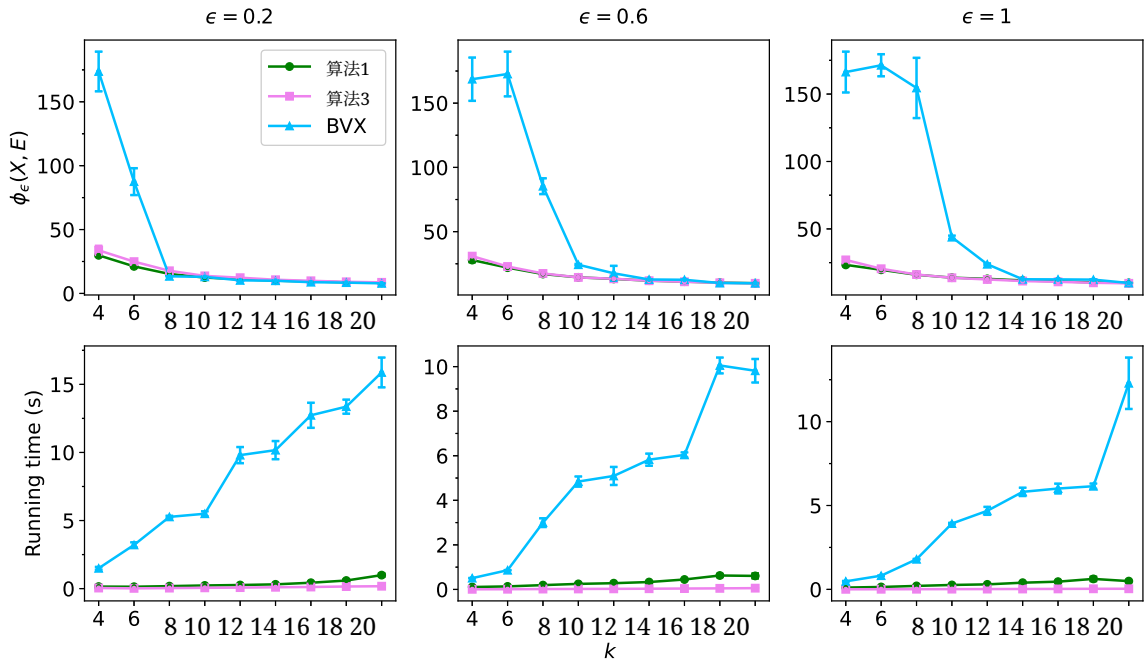


图 3：算法1和算法3在 Shuttle 上的性能。

为生成离群点，我们对每个数据集使用 B˘adoiu 和 Clarkson（2003年）的算法计算整个数据集的最小包围球；令 $r_{\texttt{meb}}$ 和 $c_{\texttt{meb}}$分别为其半径和中心。随后，我们在以 $c_{\texttt{meb}}$为中心、半径为 $1.1 \times r_{\texttt{meb}}$ 的球内随机添加 1% 的点作为离群点。

6.1 双准则算法

我们将算法1及其次线性版本算法3与 BVX进行比较。对于算法1，我们设定 $\epsilon = 0.2$, 0.6,1，并相应地修改算法3和 BVX 的参数，使它们输出相同数量的中心。我们将 k 从 4 变化到 20。实验结果如图 3、图 4、图 5和图6所示。与 BVX相比，算法1和算法3的运行时间显著更短，同时实现了相近或更低的聚类代价 $\phi_\epsilon(X, E)$。

为了更清楚地比较算法1和算法3，我们放大查看了 $\epsilon = 1$ 不含 BVX 时的实验结果（见图7）。可以看到，随着 k 增大，算法3的运行时间增长速度远慢于算法1。这一结果也符合我们的理论分析，因为算法3的时间复杂度仅为次线性。

对于较小的 k，我们还将算法2与 CKM+、 MK和 BVX 进行了比较。我们令 $k = 2$、3、4、5。对于算法2，我们设置 $\epsilon = 1$，并按照推论 12 的建议运行$\ln\frac{10}{1-\gamma}(\frac{1+\epsilon}{\epsilon})^{k-1}$次。实验结果如图8所示。总体而言，算法2的聚类代价与 CKM+ 和 MK相当，但运行速度快于这两种基线方法。BVX 速度更快，但聚类代价高于算法2。

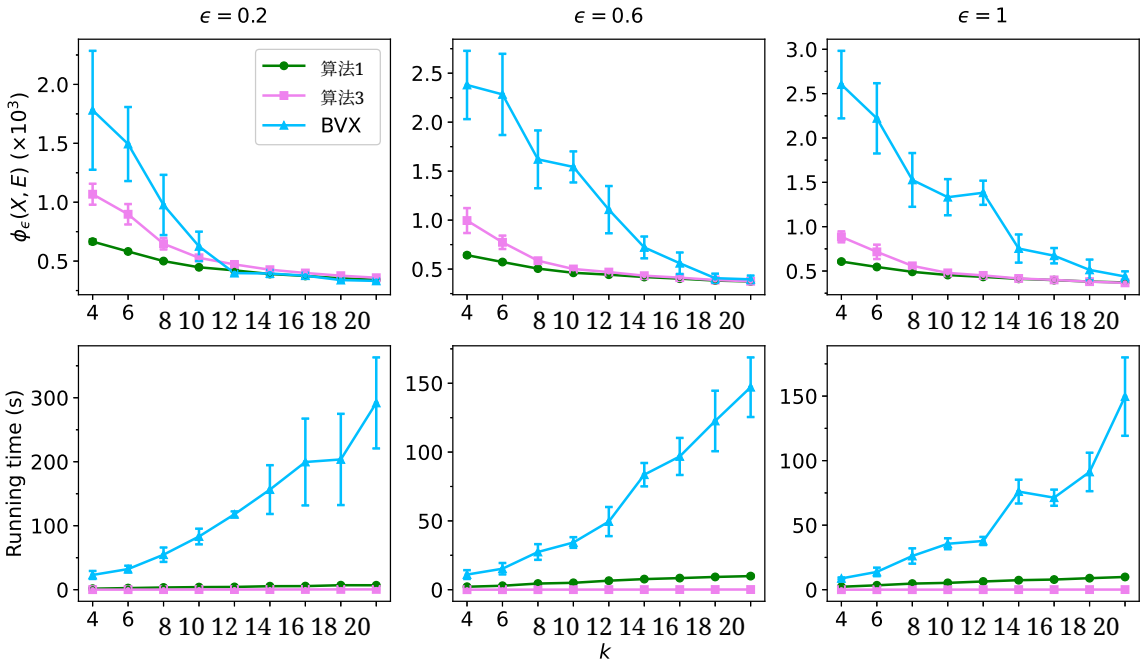


图 4：算法1和算法3在 Covertype 上的性能。

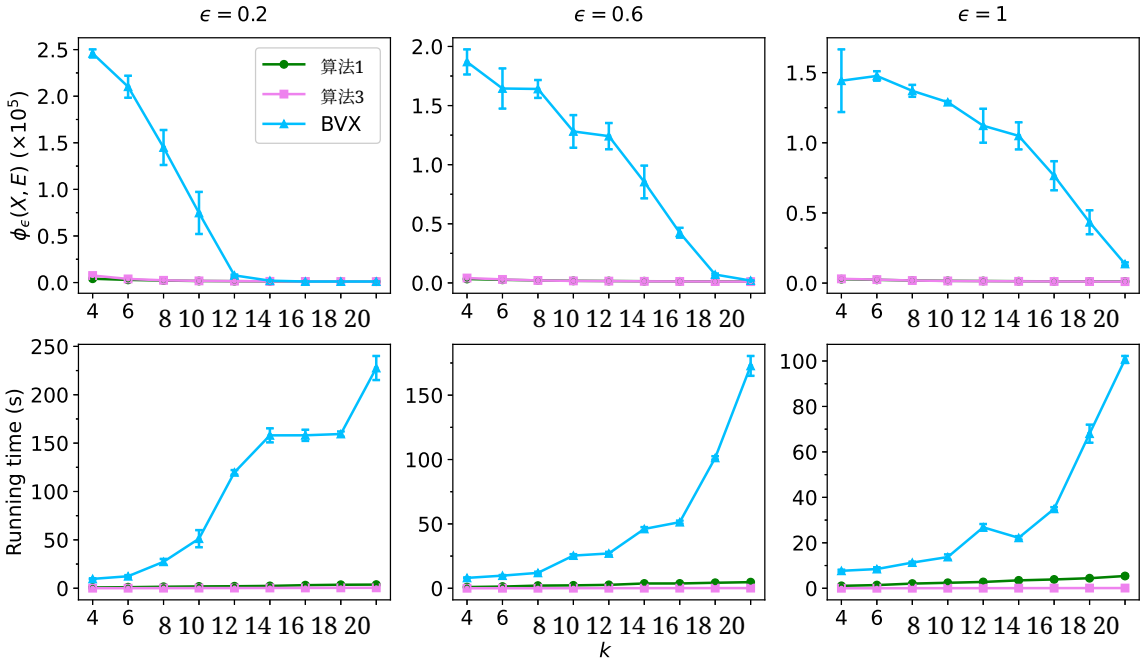


图5：算法1和算法3在 KDD Cup 1999 上的性能。

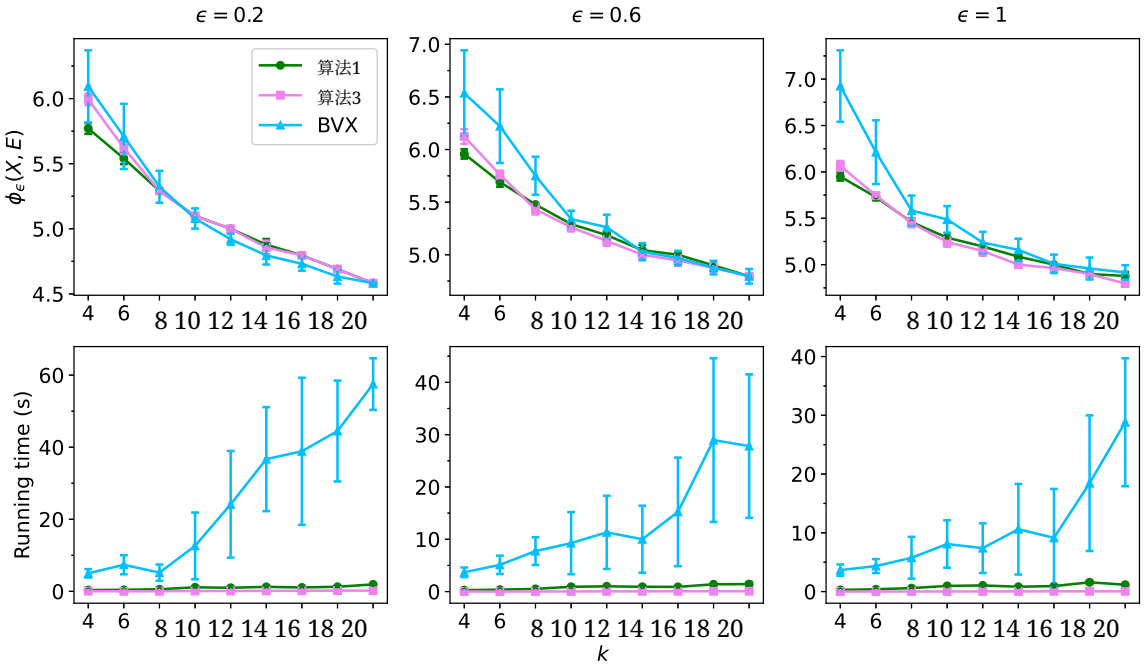


图6：算法1和算法3在 Poker Hand 上的性能。

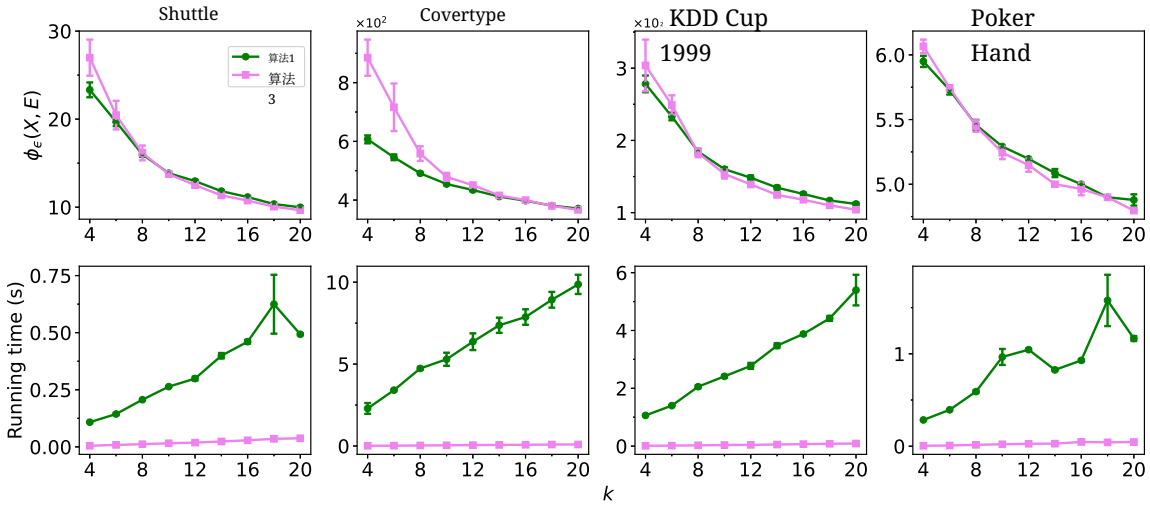


图7：算法1和算法3的比较。

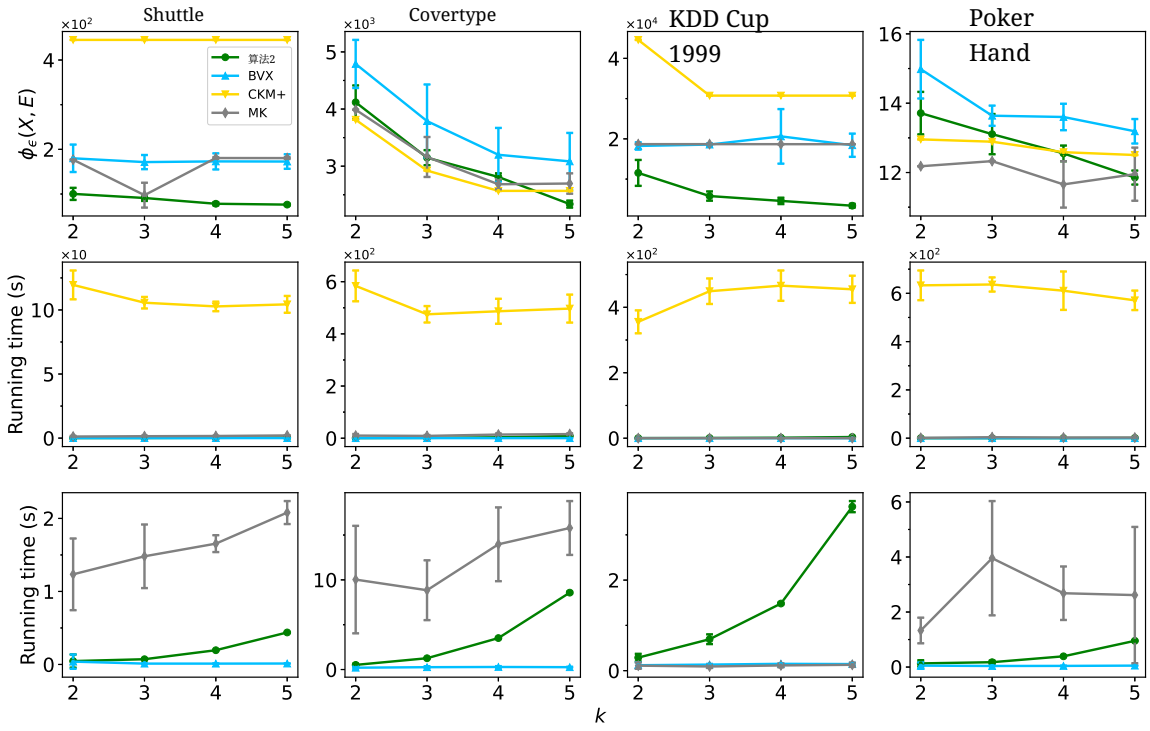


图 8：算法2的性能。第三行移除了 CKM+ ，以便更清晰地展示另外三种算法的运行时间。

6.2 核心集算法

我们将算法 5 与核心集方法 CPP 和 UNIFORM进行比较。对于这三种方法，我们将核心集的大小设为 $\{4\%n、8\%n、12\%n、16\%n$和 $20\%n\}$ ，其中 n 表示点的数量（包括离群点）。我们将 Malkomes 等人（2015）提出的算法 CLUSTER （它是对 CKM+ 的一种改进）作为“宿主”算法，运行在由算法 5 和 UNIFORM构造得到的核心集上。我们用 $\mathrm{RT}_{\texttt{coreset}}$ 表示核心集构造时间，用 $\mathrm{RT}_{\texttt{total}}$ 表示总运行时间（包括核心集构造时间，以及在核心集上运行带离群点的 k-中心算法所需的时间）。为研究核心集的优势，我们还将其与 CKM+ 和 MK 进行比较；我们直接在整个数据集（不使用核心集）上运行这两种算法，以计算聚类结果。

实验结果如图 9 所示。请注意，我们在图 9（以及第6.3节的图 10）第一行展示的是聚类代价 $\phi_0(X,E)$（而不是 $\phi_\epsilon(X,E)$），也就是说，我们恰好舍弃 z 个离群点，而不是 $(1+\epsilon)z$。 UNIFORM 始终是运行速度最快的核心集方法，因为它只是简单的均匀采样，无需任何构造过程；但在大多数情况下，其聚类代价高于算法 5 和 CPP 。算法 5 和 CPP 的聚类代价均低于 CKM+ 和 MK。与 CPP相比，算法 5 在 Covertype 和 Poker Hand 上的聚类代价更低；算法 5 的 $\mathrm{RT}_{\texttt{coreset}}$ 和 $\mathrm{RT}_{\texttt{total}}$也更低。实验结果表明，算法 5

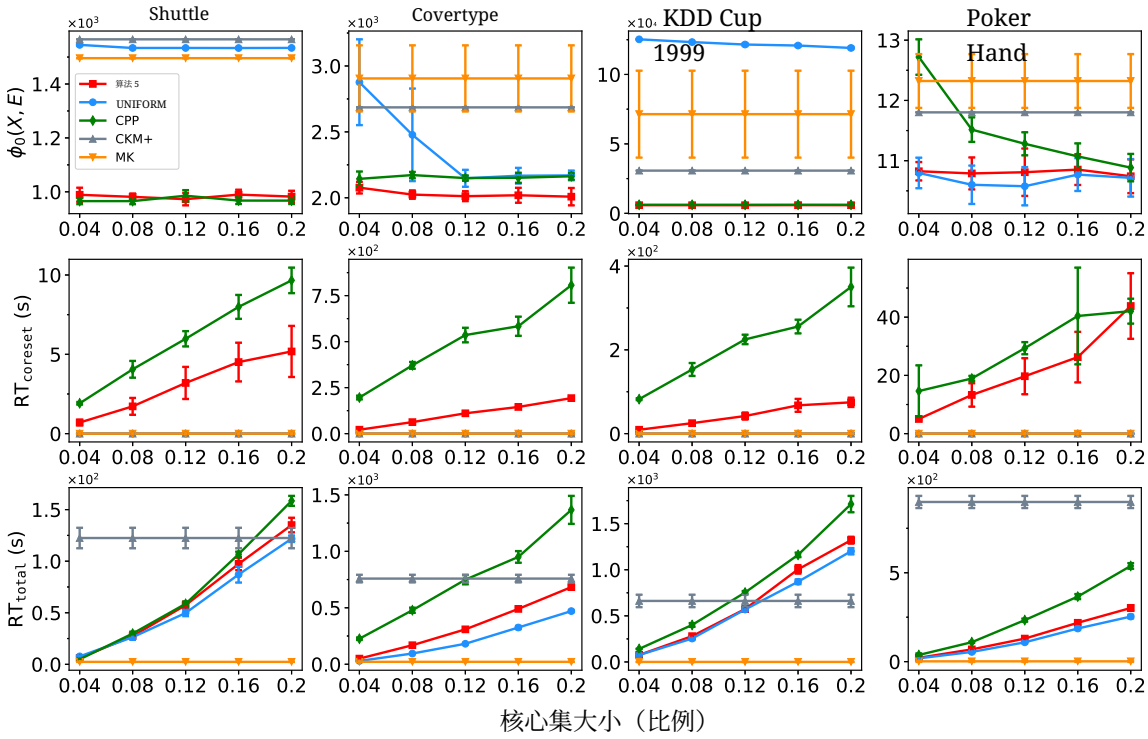


图 9：核心集方法算法 5 的性能。

能够显著缩短运行时间（当核心集大小设为 $\leq 12\%$ 时），同时也能取得良好的聚类质量。

6.3 分布式算法

我们通过改变站点数量 s，将算法 6 与 CPP、MKC+、GLZ 和 LG 进行比较。对于算法 6，在步骤 2 中，我们运行算法 5 而非算法 4，因为这四个数据集的倍增维数未知。与第 6.2 节类似，我们也在算法 6 构造的核心集上运行 CLUSTER算法。对于 CPP，按照 Ceccarello 等人（2019）的设置，每个站点向中央服务器发送一个大小为 $\lambda(k+z)$ 的核心集，其中 $\lambda = 1$、2、4。LG 返回一个 (k, z)-中心解，并按照他们的论文（Li 和 Guo，2018）的建议，在算法中设置 $\epsilon = 0.1$、0.99。

四个数据集上的聚类代价和通信成本实验结果如图 10 所示。通信成本以站点与中央服务器之间传输的浮点数总量来衡量。GLZ 和 LG 的通信成本较低，但聚类代价要高得多。算法 6 可以实现相当低的聚类代价，但与 GLZ 和 LG相比，通信成本较高。

7. 未来工作

在我们的工作之后，未来还有几个值得研究的有趣问题。例如，能否改进算法 4 的核心集构造时间，例如采用 Har-Peled 和 Mendel（2006）提出的快速网构造

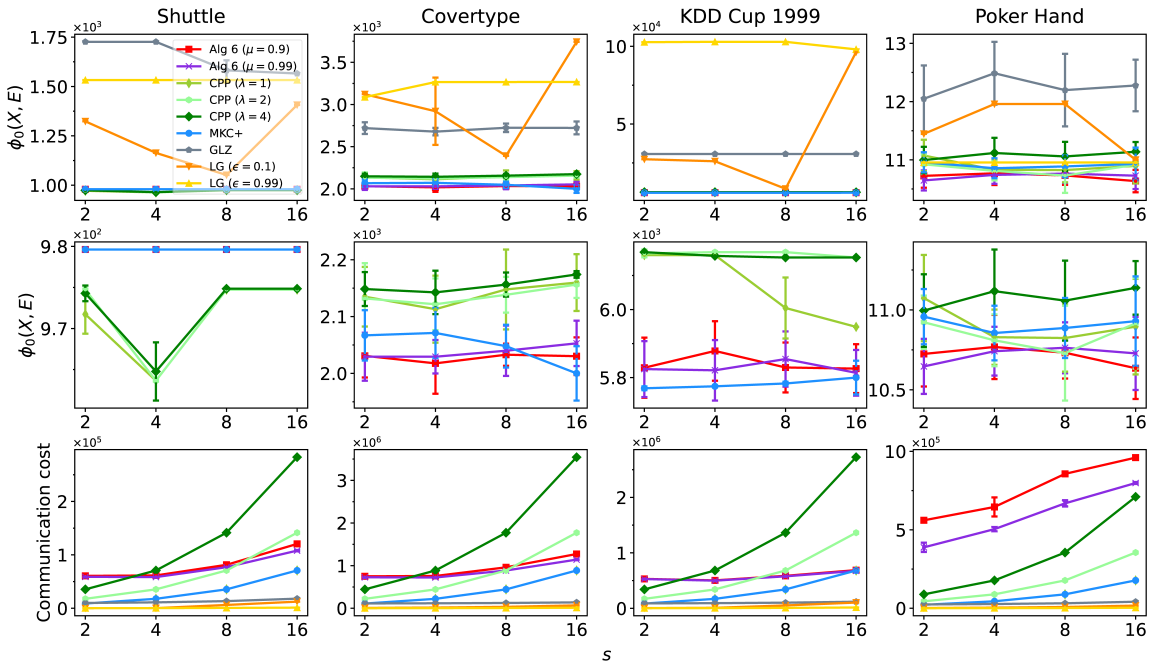


图 10：算法 6 的性能。第二行中，我们移除 GLZ 和 LG（因为它们的聚类代价远高于其他算法），并放大查看其他算法之间的比较。

方法，应用于倍增度量？从理论上说，研究使用贪心策略解决其他涉及离群点的优化问题也很有趣。此外，如果我们将 k-中心聚类替换为带离群点的 k-中心聚类，就有可能提升其在深度学习（Coleman 等，2020）、主动学习（Sener 和 Savarese，2018）以及公平性（Kleindessner 等，2019）应用中的鲁棒性。

附录 A. 命题 4 的证明

假设 H 是实例（核心集） S的一个 α-近似。令 $H_{\mathtt{opt}}$ 为使 X达到最优解的 k个聚类中心集合。于是我们有

$$\begin{aligned} \phi_0(S,H) &\leq \alpha\phi_0(S,H_{\mathtt{opt}}); && (20)\\ \phi_0(S,H) &\in (1\pm\mu)\phi_0(X,H); && (21)\\ \phi_0(S,H_{\mathtt{opt}}) &\in (1\pm\mu)\phi_0(X,H_{\mathtt{opt}}). && (22) \end{aligned}$$

结合上述不等式，我们可直接得到

$$\phi_0(X,H) \leq \frac{1}{1-\mu}\phi_0(S,H) \leq \frac{\alpha}{1-\mu}\phi_0(S,H_{\mathtt{opt}}) \leq \frac{\alpha(1+\mu)}{1-\mu}\phi_0(X,H_{\mathtt{opt}}). \quad (23)$$

因此 H 是 $\frac{\alpha}{(1+\mu)}{}^{1-\mu}$ 的 X-近似。

附录B。命题18的证明

我们只需证明第一个不等式，因为另一个可以用同样的方法得到。根据式 (12)，每个 $B_j \subseteq$球(c_j, r_X) 以及每个顶点 p移动的距离至多为$\mu r_{\texttt{opt}}$，因此我们知道 $f(B_j) \subseteq$球$(c_j, r_X + \mu r_{\texttt{opt}})$，即 $r'_E \leq r_X + \mu r_{\texttt{opt}}$。

设 p_0 为实现 $r_X = \phi_0(X, H)$的顶点，即存在某个 $1 \leq j_0 \leq k$使得 $\mathsf{d}(c_{j_0}, p_0) = r_X$。三角不等式与（12）共同推出 $\mathsf{d}(c_{j_0}, f(p_0)) \geq r_X - \mu r_{\texttt{opt}}$。因此，$r'_E \geq r_X - \mu r_{\texttt{opt}}$。

总的来说，我们有 $|r'_E - r_X| \leq \mu r_{\texttt{opt}}$。

参考文献

P. K. Agarwal、S. Har-Peled 和 K. R. Varadarajan。点的范围度量近似。《ACM期刊》，51(4):606–635，2004 年。P. K. Agarwal、G. Cormode、Z. Huang、J. M. Phillips、Z. Wei 和 K. Yi。可合并摘要。《ACM 数据库系统汇刊》(TODS)，38(4):1–28，2013 年。A. Aggarwal、A. Deshpande 和 R. Kannan。用于 k-均值聚类的自适应采样。载于《近似、随机化与组合优化：算法与技术》，第 15–28 页。Springer，2009 年。S. Aghamolaei 和 M. Ghodsi。倍增度量空间中 k-中心问题的可组合核心集。载于第 30 届加拿大计算几何会议论文集（CCCG 2018），2018年8月8日至10日，曼尼托巴大学，温尼伯，曼尼托巴省，加拿大，第 165–171 页，2018 年。N. Alon 和 J. H. Spencer。《概率方法》。约翰·威立父子出版公司，2004 年。P. Awasthi 和 M.-F. Balcan。基于中心的聚类：基础性视角。2014 年。M. Belkin。流形上的学习问题。芝加哥大学，2003年。A. Bhaskara、S. Vadgama 和 H. Xu。存在离群点时近似聚类的贪心采样。载于《神经信息处理系统进展》，第 32 卷。Curran Associates, Inc.，2019 年。B. Biggio 和 F. Roli。野生模式：对抗机器学习兴起十年之后。《模式识别》，84:317–331，2018 年。M. Blum、R. W. Floyd、V. Pratt、R. L. Rivest 和 R. E. Tarjan。选择问题的时间界。《计算机与系统科学杂志》，7(4):448–461，1973 年。A. Blumer、A. Ehrenfeucht、D. Haussler 和 M. K. Warmuth。可学习性与 Vapnik–Chervonenkis 维数。《ACM期刊》(JACM)，36(4):929–965，1989 年。M. B˘adoiu 和 K. L. Clarkson。更小的球核集。载于 ACM-SIAM 离散算法研讨会（ SODA ）论文集，第 801–802 页，2003 年。

M. Ceccarello、A. Pietracaprina 和 G. Pucci。在 MapReduce 和流式处理中求解 k-中心聚类（含离群点），精度几乎与顺序算法相同。PVLDB，12(7):766–778，2019 年。D. Chakrabarty、P. Goyal 和 R. Krishnaswamy。非均匀 k-中心问题。载于第 43 届国际自动机、语言与程序设计研讨会（ICALP 2016），2016年7月11日至15日，意大利罗马，第 67:1–67:15 页，2016 年。D. Chakrabarty、M. Negahbani 和 A. Sarkar。连续聚类与设施选址问题的近似算法。载于 S. Chechik、G. Navarro、E. Rotenberg 和 G. Herman 编，《第 30 届欧洲算法年度研讨会》（ESA 2022），LIPIcs 第 244 卷，第 33:1–33:15 页，2022 年。V. Chandola、A. Banerjee 和 V. Kumar。异常检测：综述。《ACM计算调查》(CSUR)，41(3):15，2009 年。M. Charikar、S. Khuller、D. M. Mount 和 G. Narasimhan。带离群点的设施选址问题算法。载于第十二届 ACM-SIAM 离散算法研讨会论文集，第 642–651 页。工业与应用数学学会，2001 年。M. Charikar、L. O’Callaghan 和 R. Panigrahy。更优的聚类问题流式算法。载于第三十五届 ACM 计算理论研讨会论文集，第 30–39 页。ACM，2003年。S. Chawla 和 A. Gionis。k-means–：一种统一的聚类与离群点检测方法。载于 2013 年 SIAM 国际数据挖掘会议论文集，第 189–197 页。SIAM，2013 年。J. Chen、E. S. Azer 和 Q. Zhang。分布式聚类与离群点检测的实用算法。载于《神经信息处理系统进展》，第 2253–2262 页，2018 年。K. Chen。带离群点的 k-中位数聚类的常数因子近似算法。载于第十九届 ACM-SIAM 离散算法研讨会论文集，第 826–835 页。工业与应用数学学会，2008 年。V. Cohen-Addad、D. Saulpic 和 C. Schwiegelshohn。一种新的聚类核心集框架。载于第 53 届 ACM SIGACT 计算理论研讨会论文集，第 169–182 页，2021 年。C. Coleman、C. Yeh、S. Mussmann、B. Mirzasoleiman、P. Bailis、P. Liang、J. Leskovec 和 M. Zaharia。代理选择：用于深度学习的高效数据选择。载于第 8 届国际学习表征会议（ICLR 2020），埃塞俄比亚亚的斯亚贝巴，2020年4月26日至30日。OpenReview.net，2020 年。M. de Berg、M. Monemizadeh 和 Y. Zhong。滑动窗口模型中的带离群点的 k-中心聚类。载于 P. Mutzel、R. Pagh 和 G. Herman 编，《第 29 届欧洲算法年度研讨会》（ESA 2021），2021年9月6日至8日，葡萄牙里斯本（线上

会议），LIPIcs 第 204 卷，第 13:1–13:13 页。Schloss Dagstuhl - Leibniz 信息学中心，2021 年。

A. Deshpande、P. Kacham 和 R. Pratap。鲁棒 k-均值++。载于第 36 届人工智能不确定性会议（UAI）论文集，机器学习研究论文集第 124 卷，第 799–808 页。AUAI 出版社，2020 年。

I. Diakonikolas、G. Kamath、D. Kane、J. Li、A. Moitra 和 A. Stewart。无需面对计算难题的高维鲁棒估计量。SIAM 计算杂志，48(2)：742–864，2019 年。

H. Ding 和 J. Xu。最小截尾平方估计量及相关问题的亚线性时间混合近似。载于 S. Cheng 和 O. Devillers 编，第 30 届计算几何年度研讨会（SoCG'14），日本京都，2014 年 6 月 8–11 日，第 110 页。ACM，2014 年。

H. Ding、H. Yu 和 Z. Wang。贪心策略适用于带离群点的 k 中心聚类和核心集构造。载于第27届欧洲算法年会（ESA 2019），2019年9月9日至11日，德国慕尼黑/加兴，第40:1–40:16页，2019年。

D. Dua 和 C. Graff。UCI 机器学习库，2017。网址：http://archive.ics.uci.edu/ml。ˇP. Duriˇs 和 J. D. Rolim。多方通信复杂度的下界。《计算机与系统科学杂志》，56(1):90–95，1998。ISSN 0022-0000。doi：https://doi.org/10.1006/jcss.1997.1547。网址：https://www.sciencedirect.com/science/article/pii/S0022000097915476。D. Feldman。核集：最新综述。《Wiley 跨学科评论：数据挖掘与知识发现》，10(1)，2020。Z. Friggstad、K. Khodamoradi、M. Rezapour 和 M. R. Salavatipour。带离群点聚类的近似方案。载于《第二十九届 ACM-SIAM 离散算法年会论文集》，第398–414页。SIAM，2018。T. F. Gonzalez。使最大簇间距离最小化的聚类。《理论计算机科学》，38:293–306，1985。S. Guha、Y. Li 和 Q. Zhang。分布式部分聚类。《ACM 并行计算汇刊》，6(3):11:1–11:20，2019。S. Gupta、R. Kumar、K. Lu、B. Moseley 和 S. Vassilvitskii。带离群点 k-均值的局部搜索方法。《VLDB 基金会论文集》，10(7):757–768，2017。S. Har-Peled 和 M. Mendel。低维度量中网的快速构造及其应用。《SIAM 计算杂志》，35(5):1148–1184，2006。S. Har-Peled 和 M. Sharir。几何中的相对 (p，ε)-近似。《离散与计算几何》，45(3):462–496，2011。

D. S. Hochbaum 和 D. B. Shmoys。k-中心问题的最优启发式算法。《运筹学数学》，10(2):180–184，1985。L. Huang、S. Jiang、J. Li 和 X. Wu。倍增度量中聚类（含离群点）的 ε-核集。载于《2018 IEEE 第59届计算机科学基础研讨会（FOCS）》，第814–825页。IEEE，2018。S. Im、M. M. Qaem、B. Moseley、X. Sun 和 R. Zhou。k-均值聚类的快速噪声去除。载于《第二十三届国际人工智能与统计会议论文集》，《机器学习研究论文集》。PMLR，2020。P. Indyk、S. Mahabadi、M. Mahdian 和 V. S. Mirrokni。用于多样性与覆盖最大化的可组合核集。载于 R. Hull 和 M. Grohe（编），《第33届 ACM SIGMOD-SIGACT-SIGART 数据库系统原理研讨会论文集》，PODS'14，美国犹他州斯诺伯德，2014年6月22–27日，第100–108页。ACM，2014。A. K. Jain。数据聚类：超越 k-均值的50年。《模式识别快报》，31(8):651–666，2010。R. D. King、C. Feng 和 A. Sutherland。Statlog：大型真实问题中分类算法的比较。《应用人工智能国际期刊》，9(3):289–333，1995。M. Kleindessner、P. Awasthi 和 J. Morgenstern。用于数据摘要的公平 k-中心聚类。载于 K. Chaudhuri 和 R. Salakhutdinov（编），《第36届国际机器学习会议论文集》，ICML 2019，2019年6月9–15日，美国加利福尼亚州长滩，第97卷《机器学习研究论文集》，第3448–3457页。PMLR，2019。R. Krishnaswamy、S. Li 和 S. Sandeep。通过迭代舍入实现带离群点的 k- 中位数和 k- 均值的常数近似。载于《第50届 ACM SIGACT 计算理论年会论文集》，第646–659页。ACM，2018。S. Li 和 X. Guo。强噪声数据的分布式 k 聚类。载于《神经信息处理系统进展》，第7849–7857页，2018。Y. Li、P. M. Long 和 A. Srinivasan。学习样本复杂度的改进界。《计算机与系统科学杂志》，62(3):516–527，2001。G. Malkomes、M. J. Kusner、W. Chen、K. Q. Weinberger 和 B. Moseley。海量数据上带离群点的快速分布式 k-中心聚类。载于《神经信息处理系统进展》，第1063–1071页，2015。R. M. McCutchen 和 S. Khuller。带匿名性的带离群点 k-中心聚类流式算法。载于《近似、随机化与组合优化：算法与技术》，第165–178页。Springer，2008。V. Mirrokni 和 M. Zadimoghaddam。用于分布式子模最大化的随机化可组合核集。载于《第四十七届 ACM 计算理论研讨会年会论文集》，第153–162页，2015。

A. Munteanu、C. Schwiegelshohn、C. Sohler 和 D. Woodruff。关于逻辑回归的核心集。神经信息处理系统进展，第 31 卷，2018 年。E. Schubert、J. Sander、M. Ester、H. P. Kriegel 和 X. Xu。重新审视 DBSCAN：为什么以及如何（仍然）应当使用 DBSCAN。ACM 数据库系统汇刊（TODS），42(3)：1–21，2017 年。O. Sener 和 S. Savarese。卷积神经网络的主动学习：一种核心集方法。载于第六届国际学习表征会议，ICLR 2018，温哥华，不列颠哥伦比亚省，加拿大，2018年4月30日至5月3日，会议论文集。Open-Review.net，2018 年。K. Talwar。绕过嵌入：低维度量的算法。载于第三十六届 ACM 计算理论研讨会论文集，第 281–290 页，2004 年。