运筹学学报
Operations Research Transactions

k-均值问题的差分隐私算法综述

袁藩, 徐大川, 张冬梅

引用本文:
袁藩, 徐大川, 张冬梅. *k*-均值问题的差分隐私算法综述[J]. 运筹学学报, 2022, 26(3): 1-16.
YUAN Fan, XU Dachuan, ZHANG Dongmei. A survey of differential privacy algorithms for the *k*-means problem[J]. *Operations Research Transactions*, 2022, 26(3): 1-16.

DOI: 10.15960/j.cnki.issn.1007-6093.2022.03.001

k-均值问题的差分隐私算法综述*

袁 藩[1] 徐大川[1] 张冬梅[2,†]

摘要 k-均值问题是机器学习和组合优化领域十分重要的问题。它是经典的 NP-难问题, 被广泛的应用于数据挖掘、企业生产决策、图像处理、生物医疗科技等领域。随着时代的发展, 人们越来越注重于个人的隐私保护: 在决策通常由人工智能算法做出的情况下, 如何保证尽可能多地从数据中挖掘更多信息, 同时不泄露个人隐私。近十年来不断有专家学者研究探索带隐私保护的 k-均值问题, 得到了许多具有理论指导意义和实际应用价值的结果, 本文主要介绍关于 k-均值问题的差分隐私算法供读者参考。

关键词 k-均值问题, 差分隐私, 近似算法, 指数机制, 拉普拉斯机制

中图分类号 O221.7

2010 数学分类号 90C27, 90C59

A survey of differential privacy algorithms for the k-means problem*

YUAN Fan[1] XU Dachuan[1] ZHANG Dongmei[2,†]

Abstract The k-means problem is a very important problem in the field of machine learning and combinatorial optimization. It is a classic NP-hard problem, which is widely used in data mining, business production decision-making, image processing, biomedical technology, and more. As people in these fields pay more and more attention to personal privacy protection, which raises a question: In the case that decisions are usually made by artificial intelligence algorithms, how to ensure that as much information as possible is extracted from the data without revealing personal privacy? In the past ten years, experts and scholars have continuously studied and explored the k-means problem with privacy protection, and has also obtained many results with theoretical guiding significance and practical application value. In this paper we mainly introduce these differential privacy algorithms of the k-means problem for readers.

Keywords k-means problem, differential privacy, approximate algorithm, exponential mechanism, Laplace mechanism

Chinese Library Classification O221.7

2010 Mathematics Subject Classification 90C27, 90C59

收稿日期: 2021-12-20

* 基金项目: 国家自然科学基金 (Nos. 11871081, 12131003)

1. 北京工业大学北京科学与工程计算研究院, 北京 100124; Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing 100124, China

2. 山东建筑大学计算机科学与技术学院, 山东济南 250101; School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, Shandong, China

† 通信作者 E-mail: zhangdongmei@sdjzu.edu.cn

聚类是数据挖掘领域的重要分支, 在机器学习、组合优化、无监督学习、数据分析、图像处理等领域有广泛的应用, 是人工智能领域的热点问题之一。聚类问题的输入是度量空间上的数据点, 目标是将数据集根据某种相似度量自然分簇, 使簇内数据点相似性尽可能高, 簇间数据点相似性尽可能低。

基于划分的聚类因其能够处理数值数据、属性数据、文本数据、多媒体数据、离散序列、时间序列、流数据等具有非常广泛的应用, 而 k-均值则是划分聚类中的代表。k-均值问题是机器学习和组合优化领域的经典问题之一。相应的 Lloyd 算法是数据挖掘的十大经典算法之一, 在各种领域被广泛研究和应用, 特别是在图像处理和特征工程方面。k-均值问题的输入是 d 维欧氏空间 $\mathbb{R}^d$ 中的一些数据点集合 C, 目标是从 $\mathbb{R}^d$ 中选出 k 个中心点, 使得所有数据点到离其最近的中心点的距离的平方和最小。k-均值问题是 NP-难问题[1], 除了启发式算法, 还可以从近似算法的角度设计多项式时间算法求解。目前经典的 k-均值问题最好的结果是 Ahmadian 等[2] 给出的 6.357-近似算法。由于 k-均值问题在实际应用中遇到的情况多种多样, 不同的情况可能会对距离给出不同的定义, 对聚类中心采用不同的选取方式, 或采用不同的优化目标函数, 这样就引出了与 k-均值问题相关的各种各样的变形问题。有关 k-均值问题及其变形问题的最新研究结果, 建议读者参考综述论文[3]。

在当今的互联网大数据时代, 用户保护个人隐私的意识日益增长, 为各种数据挖掘算法带来了新的挑战, 即在尽可能从数据中挖掘更多价值的同时对隐私进行保护。在实际应用场景中, 当输入数据中有人际关系、病史、客户的位置等关于个人的敏感信息时, 我们希望算法可以保护用户的隐私信息, 这就引发了人们对可保护隐私的智能算法的研究, 如带隐私的大数据问题[4-7]、带隐私的社交网络问题[8-10]、带隐私的无线网络问题[11, 12]、带隐私的深度学习和医疗问题[13-15]、带隐私的次模优化问题[16-18] 等。

由于 k-均值问题是目前应用最广泛的聚类问题之一, 加上人们对用户隐私保护的日益增长的认识和需求, 从而激发了对可保护隐私的 k-均值问题算法的研究。带差分隐私的组合优化问题的研究始于 2010 年 Gupta[19] 的工作, 是机器学习和组合优化领域的前沿方向。本文首先介绍 k-均值问题和差分隐私概念以及两种不同的差分隐私模型, 然后介绍解决差分隐私问题的两种常用机制, 最后介绍几种 k-均值问题的差分隐私算法。

1 差分隐私 k-均值问题

1.1 k-均值问题

首先我们介绍 k-均值问题及一些相关的符号。k-均值问题的输入是 d 维欧氏空间 $\mathbb{R}^d$ 上的数据点集合 C, 以及正整数 k, 其中 C 中数据点的个数为 n。我们的目标是从 $\mathbb{R}^d$ 中选出 k 个点的中心点集合 S, 然后将所有的用户点都分配到离其最近的中心点上, 使得所有用户点到其对应的中心点的距离的平方和最小。对于 $\mathbb{R}^d$ 上的任意两点 $i, j \in \mathbb{R}^d$, 我们用 $\delta(i, j)$ 来表示他们之间的距离。而对于任意一点 j 和任意集合 S 之间的距离我们用 $\delta(S, j) = \min_{i \in S} \delta(i, j)$ 来表示。这样 k-均值问题可以描述为, 在 $\mathbb{R}^d$ 上找到基数为 k 的中心点集合 S, 使得下式取得最小：

$$\mathrm{cost}(S) = \sum_{j \in C} \delta(S, j)^2 。$$

由于 k-均值问题是 NP-难问题[1], 可以使用近似算法找到形式为 $\alpha \times \mathrm{OPT} + \beta$ 的解, 其中 OPT 表示问题的最优解的值。$\alpha \times \mathrm{OPT} + \beta$ 中的 $\alpha \times \mathrm{OPT}$ 称为乘法项误差, α 为乘法项误差的系数, β 称为加法项误差。由于原始的 k-均值问题的中心点集合是在 $\mathbb{R}^d$ 上选取的, 而 $\mathbb{R}^d$ 上的点是无穷多的, 这就导致了许多启发式算法和近似算法无法直接应用于原始的 k-均值问题, 因而目前大家主要研究的都是离散 k-均值问题。在这个问题中, 中心点集合 S 不是从 $\mathbb{R}^d$ 上选取的, 而是从 $\mathbb{R}^d$ 的有限子集 F 上选取的。Matoušek[20] 于 2000 年提出了在 $\mathbb{R}^d$ 上寻找近似中心点集合的方法, 运用他的方法可以在多项式时间内找到一大小为 $O(n\varepsilon^{-d}\log(1/\varepsilon))$ 的近似中心点集合 $F \subseteq \mathbb{R}^d$, 在 F 上选取中心点和在 $\mathbb{R}^d$ 上选取中心点, 两个问题的近似比只会相差 $(1+\varepsilon)$。根据他的结论可以知道离散 k-均值问题和原始的 k-均值问题的解相差不大, 而离散 k-均值问题的设定可以让我们利用众多的启发式算法和近似算法。

1.2 差分隐私概念

在实际应用中, 当输入数据 (例如人际关系、病史、客户的位置等) 是关于个人的敏感信息时, 在利用经典算法来处理这些数据时, 会使个人敏感信息有泄露和被攻击的风险。假设有 n 个点 $c_1, c_2, \cdots, c_n$, 根据不带隐私的算法其中心的计算方式为:

$$s = \frac{\sum_{j=1}^{n} c_j}{n}。$$

此时如果存在一个攻击者, 其拥有最大限度的额外知识, 即知道 $c_1, c_2, \cdots, c_{n-1}$ 个点的位置, 那么其可以很轻松地通过调用算法两次得到 s 和 s_1

$$s_1 = \frac{\sum_{j=1}^{n-1} c_j}{n-1}。$$

再通过 s 和 s_1 之间的差距得到 c_n 位置为:

$$c_n = sn - s_1(n-1)。$$

可以看到攻击者只需重复调用算法, 计算相邻数据集的输出差就可以很轻易的得到用户的隐私信息。为了防止被攻击, 如何设计出既能充分利用数据的整体统计性质来进行计算, 又能保证不泄露数据信息的隐私算法就显得格外重要了。接下来介绍 Dwork[21] 于 2006 年提出的差分隐私概念。

定义 1[22] 给定一个随机算法 M, 记算法的输出空间为 $\mathrm{Range}(M)$, 当算法的输入是两个只相差一个元素的数据 $A = \{x_1, x_2, \cdots, x_n, x_{n+1}\}$ 和 $B = \{x_1, x_2, \cdots, x_n\}$ 时, 对于任意输出集合 $N \subseteq \mathrm{Range}(M)$, 总有下式成立, 则称这个随机算法 M 满足 ε-差分隐私:

$$\Pr[M(A) \in N] \leqslant \exp(\varepsilon) \times \Pr[M(B) \in N]。$$

简单来说, 满足差分隐私的算法保证当输入只有一个元素发生改变时, 整个算法的计算结果的分布不会发生明显变化。差分隐私算法可以保证当攻击者想得到用户 x_{n+1} 的位置信息时, 即使攻击者拥有一切除 x_{n+1} 用户位置以外的信息 (如算法每次的输出, 其他用户的信息等), 其也不可能得到用户 x_{n+1} 的真实信息, 攻击者甚至不知道用户 x_{n+1} 是否在数据库中, 因为差分隐私定义下的算法对于个人信息的保护是十分严密的。这个

定义一方面保证了个人敏感数据的隐私, 防止了差分攻击的发生, 另一方面其允许算法在大量个人更改他们的数据时做出相应的调整, 这就保证了算法的灵敏性和有效性, 使得算法在隐私保护和算法的性能两方面取得了平衡。

对于复杂的隐私保护问题, 可能会多次应用差分隐私保护算法以得到更好的隐私保护效果。Mcsherry 等[23, 24] 提出了隐私保护算法的两个组合性质：序列组合性与并行组合性。

性质 1 (序列组合性[23]) 设有 t 个差分隐私算法 $M_1, M_2, \cdots, M_t$, 其中算法 $M_i(1 \leqslant i \leqslant t)$ 满足 ε_i-差分隐私。对于同一个数据集 C, 顺序使用这些算法 M_i 所构成的组合算法 $M(C) = \{M_i(C)\}$ 满足 $(\sum_{i=1}^{t} \varepsilon_i)$-差分隐私。

性质 2 (并行组合性[24]) 设有 t 个差分隐私算法 $M_1, M_2, \cdots, M_t$, 其中算法 $M_i(1 \leqslant i \leqslant t)$ 满足 ε_i-差分隐私。对于不相交的数据集 $C_1, C_2, \cdots, C_t$, 分别使用这些算法所构成的组合算法 $M(C_1, C_2, \cdots, C_t) = \{M_1(C_1), M_2(C_2), \cdots, M_t(C_t)\}$ 满足 $(\max\{\varepsilon_i\})$-差分隐私。

1.3 差分隐私模型

在隐私聚类领域, 主要的隐私模型分为两种。一种称为集中式差分隐私 (centralized differential privacy), 在这个模型下, 存在一个可信赖的数据中心, 其收集所有用户的真实信息, 对这些真实信息进行计算并发布满足差分隐私的信息给公众以供查询。在这个模型下, 数据中心发布的信息是满足差分隐私的, 因而不会泄露客户的真实信息。但在现实情况中用户真实信息的泄露可能出现在数据中心的存储环节, 即使是著名公司也经常出现泄露客户信息的情况。例如, 2018 年谷歌的 Google+ 社交网络数十万用户隐私数据被曝光。2019 年, 超过 2.67 亿的 Facebook 用户的用户 ID、电话号码和姓名在网上被公开[25]。而在 2020 年 2 月有 650 万以色列选民的个人数据由于选举人应用程序的缺陷而被暴露[26]。为了解决以上问题人们提出另一种差分隐私模型, 局部差分隐私 (local differential privacy) 模型, 在这个模型下人们不再信任数据中心, 人们在将数据交给中心的时候已经对数据进行扰动和加密处理, 只有数据所有者才能访问自己的原始数据, 而数据中心得到的数据是经过扰动的不真实数据, 这样可以为用户提供更强的隐私保护。但其也同样存在一些问题, 例如数据中心的算法在处理数据时, 由于算法输入就已经是非真实信息, 那么最后算法的输出自然也会带有一定程度的偏差, 所以集中式差分隐私 (centralized differential privacy) 模型下, 算法更注重的是用户隐私的保护, 而在局部差分隐私 (local differential privacy) 模型下, 算法更注重的是算法性能的维持。这就导致了两个模型下, 差分隐私算法的不同。

1.4 差分隐私 k-均值问题

问题 1 (差分隐私 k-均值问题): 输入 d 维欧氏空间 $\mathbb{R}^d$ 上的一个用户点集合 C, 以及一个正整数 k。我们希望设计一个算法从 $\mathbb{R}^d$ 选出 k 个点的中心点集合 S, 算法满足 ε-差分隐私, 且使得下式取得最小:

$$\mathrm{cost}(S) = \sum_{j \in C} \delta(S, j)^2。$$

根据 Bassily 等[27] 的结果我们可以知道, 与不带隐私的聚类问题不同的是, 任何差分隐私聚类算法的近似比中的 β 均不为 0。我们在差分隐私下的 k-均值问题中想要保护的

是用户点集合 C 的位置信息。

2 差分隐私问题常用机制

设计出满足差分隐私的 k-均值问题的算法的一般思路是对于传统的 k-均值问题的算法进行随机化, 在传统的局部搜索、k-均值++ 等 k-均值问题算法的基础上, 运用隐私保护中的指数机制和拉普拉斯机制对其进行改造, 从而设计出合适的满足差分隐私的 k-均值问题的算法。我们在这节中主要介绍指数机制和拉普拉斯机制。

2.1 指数机制

McSherry 等[23] 于 2007 年提出了指数机制 (Exponential Mechanism)。对于问题的输入域 D 的一组数量为 n 的输入 d, 隐私机制的目标是随机的将 d 映射到输出域 $\mathbb{R}$ 中的某个输出。在指数机制中存在一个打分函数 $q: D^n \times \mathbb{R} \to \mathbb{R}$, 此函数对于任意的 $D^n \times \mathbb{R}$ 中的一对 (d, r) 都会给出一个分数, 我们可以理解为分数越高, r 越好。给定一个输入 $d \in D^n$, 该机制的目标是返回一个 $r \in \mathbb{R}$, 使得满足差分隐私的前提下, $q(d, r)$ 的分数越高越好。指数机制如下式所示：当给定一个输入 $d \in D^n$ 和一个参数 ε 时, 指数机制选择 $\mathbb{R}$ 中的值 r 的概率正比于 $\exp(\varepsilon q(d, r))$

$$\Pr[\varepsilon_q^{\varepsilon}(d) = r] \propto \exp(\varepsilon q(d, r))。$$

此外如果函数 $q(d, r)$ 对于只相差一个元素的输入 d_1 和 d_2, 其函数值的最大改变量为 Δ, 即 $\Delta = \max_{d_1, d_2 \in D^n} \|q(d_1, r) - q(d_2, r)\|$, 那么有如下定理成立:

定理 1[23] 指数机制 $\varepsilon_q^{\varepsilon}$ 保持了 $(2\varepsilon\Delta)$-差分隐私。

由于在组合优化中问题的输入和输出大多是离散的, 所以我们还有以下定理成立。

定理 2[19] 对于一个问题的解集合 R, 和解的一个子集合 R_{OPT}, R_{OPT} 其中的元素 r 满足 $q(d, r) = \max_{r \in R} q(d, r)$, 给定参数 t, ε, 当我们运用指数机制从 R 中挑选一个输出 $r' = \varepsilon_q^{\varepsilon}(d)$ 的时候, 其保持 $(2\varepsilon\Delta)$-差分隐私, 且有:

$$\Pr[q(d, \varepsilon_q^{\varepsilon}(d)) \geqslant \max_{r \in R} q(d, r) - \ln(|\ R\ |\ /\ |\ R_{\mathrm{OPT}}\ |)/\varepsilon - t/\varepsilon] > 1 - \exp(-t)。$$

以上定理表达的是, 当使用指数机制选取问题的解时, 以较大的概率指数机制选取的值 r' 的打分 $q(d, r' = \varepsilon_q^{\varepsilon}(d))$ 和最优解的打分 $q(d, r \in R_{\mathrm{OPT}}) = \max_r q(d, r)$ 只差一个常数 $\ln(|\ R\ |\ /\ |\ R_{\mathrm{OPT}}\ |)/\varepsilon + t/\varepsilon$, 且可以根据需要调节参数 t, ε 来控制概率和常数取值。换句话说, 当用指数机制来挑选候选解的时候, 通过指数机制选出来的解和问题最优解的取值相差不大。指数机制由于具有上述性质且其易于使用和分析, 再加上指数机制非常适用于离散问题, 因而在组合优化领域, 大多数隐私算法均使用了指数机制, 如聚类问题[19, 28]、次模问题[17] 等。

2.2 拉普拉斯机制

定义 2[29] 对于一个定义在 D 上的函数 $f: D \to \mathbb{R}$, 用 $\Delta(f) = \max_{d_1, d_2 \in D} \|f(d_1) - f(d_2)\|_1$ 表示函数的最大改变量, 拉普拉斯机制 (Laplace Mechanism) 是指:

$$f(\widehat{D}) = f(D) + \mathrm{Lap}(\Delta(f)/\varepsilon)。$$

$\mathrm{Lap}(\Delta(f)/\varepsilon)$ 表示服从拉普拉斯分布的随机噪声, 其中分布的位置参数为 0, 尺度参数为 $(\Delta(f)/\varepsilon)$。拉普拉斯机制通过向算法结果中加入服从拉普拉斯分布的随机噪声来实现 ε-差分隐私保护。如果把加入服从拉普拉斯分布的随机噪声改成服从高斯分布的随机噪声, 那么这就叫做高斯机制。拉普拉斯机制更加适合连续的问题和算法, 对于算法的中间过程或者结果中加入微小的随机噪声以获得算法性能和隐私保护的平衡。噪声越小, 算法最后的结果就越好, 噪声越大, 用户的隐私保护效果越好, 但算法最后结果的可能受到较大影响从而不太准确。当噪声是 $\mathrm{Lap}(\Delta(f)/\varepsilon)$ 时我们还有以下定理成立。

定理 3[29] 拉普拉斯机制加入的噪声为 $\mathrm{Lap}(\Delta(f)/\varepsilon)$ 时, 其保持了 ε-差分隐私。

指数机制和拉普拉斯机制是最为广泛和常见的两种隐私保护机制, 两者皆可以用来设计各种问题的隐私保护算法。其区别在于指数机制一般用于离散问题和组合算法, 即问题的输入、解空间、算法为离散元素时, 指数机制比较适合。而拉普拉斯机制更适用于连续问题和连续优化算法, 即问题的输入、解空间、算法为连续元素时, 拉普拉斯机制比较适合。

到此我们已经介绍了两种最常用的隐私保护机制, 所有的 k-均值问题的隐私保护算法都是传统算法和以上两种隐私保护机制的结合, 下面的章节我们开始介绍 k-均值问题的差分隐私保护算法。

3 k-均值问题的差分隐私保护算法

3.1 基于 CDP 的 k-均值问题的差分隐私算法

在集中式差分隐私 (CDP) 模型下, 如图 1 所示存在一个可信赖的数据中心, 其收集所有用户的真实信息, 对这些真实信息进行计算并发布满足差分隐私的信息给公众以供查询。在这个模型下, 数据中心发布的信息是满足差分隐私的因而不会泄露客户的真实信息。在集中式差分隐私 (CDP) 模型下关于隐私 k-均值问题有着不少的工作如文献 [28,30-36] 等, 由于篇幅限制我们挑选其中具有代表性的几个工作进行介绍。在这一章我们设定问题的输入点的集合为 $C \subseteq \mathbb{R}^d$, 输入点的数目为 n, 中心点候选集为 F, 聚类所需中心点的数目为 k。在这个模型下有 k-均值++ 算法、局部搜索技巧、最大覆盖技巧等与隐私保护机制相结合而成的隐私保护算法, 接下来我们一一对其进行介绍。

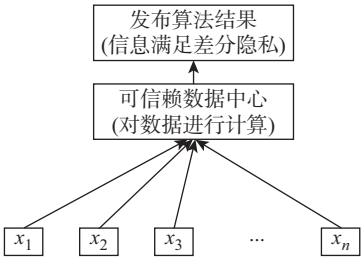

图 1 集中式差分隐私(CDP)模型

3.1.1 基于 k-均值++ 算法的差分隐私算法

Nock 等人[34] 于 2016 年在著名的 k-均值++ 算法[37] 的基础上提出了 k-variates++ 算法, 并证明其满足差分隐私需求。k-variates++ 算法的核心思想是在 k-均值++ 算法的基础上使用拉普拉斯机制增加噪声来保持隐私。令 $\Delta = \max_{c_i,c_j\in C}\|c_i - c_j\|_2$ 为数据集中任意两点间的最大距离。他们算法的近似比为 $O(\log k) \times \mathrm{OPT} + O(n\Delta^2/(\varepsilon + \log n)^2)$。

算法 1[34]

输入： 数据集 $C = \{c_i\}_{i=1}^n \subseteq \mathbb{R}^d$, 一组分布 $\{p(\mu_c, \theta_c), c \in C)\}$, 中心点数量 k。

步骤 1 初始化中心点集合 $S = \varnothing$。

步骤 2 用 t 来表示算法迭代的次数, $t = 1, 2, 3, \cdots, k$。

步骤 2.1 当 $t = 1$ 时, 令 q_1 是 C 上的均匀分布。根据分布 q_1 从 C 中随机取出一个点 c_t。当 $t > 1$ 时, 根据数据点到已有中心点集 S 的距离来构造新的分布 q_t

$$q_t(c) = E_t(c)(\sum_{c^{'}\in C} E_t(c^{'}))^{-1}。$$

其中

$$E_t(c) = \min_{s\in S} \parallel c - s \parallel_2^2$$

表示点到已有中心点集合 S 中离其最近的中心点的距离。根据新的分布 q_t 从 C 中随机取出一个点 c_t。

步骤 2.2 利用新选出的点 c_t 其对应的分布 $p(\mu_{c_t}, \theta_{c_t})$, 从中随机选出新的中心点 s_t。

步骤 2.3 把 s_t 加入到中心点集 S 中更新 S。

输出： 最终的中心点集 S。

该算法与 k-均值++ 算法最大的区别在于, k-均值++ 算法每次选取中心点是根据数据点到已有中心点集合的距离来计算每个数据点被选为下一个中心点的概率, 离已有中心点集合越远的点被选为下一个中心点的概率越大。而 k-variates++ 算法的大体框架和 k-均值++ 算法一致, 只是在根据数据点到已有中心点集合的距离选出新的候选中心点 c_t 时并不直接将其定为下一个中心点, 而是以 c_t 为中心在周围构造出一个拉普拉斯分布, 然后根据这个分布在 c_t 周围随机选择新的点 s_t 作为下一个中心点, 从而保持了算法的隐私性。至于如何选取合适的参数构造出满足拉普拉斯分布的 $p(\mu_{c_t}, \theta_{c_t})$, 详细内容请读者查看文献[34]。

3.1.2 基于局部搜索技巧的差分隐私算法

Balcan 等[28] 于 2017 年提出了局部搜索技巧与指数机制结合而成的 k-均值问题的隐私保护算法。他们算法的主要思想参考了 Gupta 等[19] 于 2010 年提出的 k-中位问题的隐私保护算法。在传统的局部搜索算法中我们每一步选择对当前解的改进量最大的元素对进行交换从而产生新的解, 而在隐私算法中他们把算法每步取最大增量元素对变成了以指数机制概率选择元素对, 将局部搜索技巧与指数机制相结合。他们算法的近似比为 $O(\log^3(n)) \times \mathrm{OPT} + O(\mathrm{poly}(\log(n), d, k))$。

由于 Matoušek[20] 的文章不能直接应用于隐私聚类问题, 即不能简单地应用他们的技巧产生满足差分隐私的近似中心点候选集 F, 所以 Balcan 等[28] 提出了一个满足差分

隐私的近似中心点算法, 其主要思想是在 Matoušek[20] 算法的基础上随机加噪音, 从而达到产生隐私近似中心点候选集的目的。而在产生的近似中心点候选集 F 上寻找 k 个中心点和在 $\mathbb{R}^d$ 上寻找 k 个中心点的两个问题的解的比为 $O(\log^3(n))$, 又因局部搜索技巧是常数近似比技巧, 从而得到了以上的算法近似比。

算法 2[28]

输入: 数据集 $C=\{c_i\}_{i=1}^n \subseteq \mathbb{R}^d$, C 中元素的最大度量 $\|c_i\|_2 \leqslant \Lambda$, 参数 ε, δ, 中心点候选集 F, 中心点数量 k。

步骤 1 均匀地从 F 中独立同分布的采样 k 个中心点构成初始中心点集 Z^0, 令 $T=100k\log(n/\delta)$。

步骤 2 用 t 来表示算法迭代的次数, $t=1,2,3,\cdots,T$。

在每一次迭代中对于任意的两点 ($x\in Z^{t-1}$, $y\in F\backslash Z^{t-1}$), 我们从当前中心点集 Z^{t-1} 中拿出 x, 把 y 加入当前中心点集 Z^{t-1} 的交换后的新中心点集记为 $Z'=Z^{t-1}-\{x\}+\{y\}$。遍历所有的点对 ($x\in Z^{t-1}$, $y\in F\backslash Z^{t-1}$), 以正比于下式的概率选出一对 (x,y)

$$\exp\left(-\varepsilon\frac{\mathrm{cost}(Z')-\mathrm{cost}(Z^{t-1})}{8\Lambda^2(T+1)}\right),$$

即在当前中心点集中把 x 拿出, 把 y 放入时造成的目标函数减小量越大的 (x,y) 对被选中的概率越大。根据选出的 (x,y) 构造下一代的中心点集合 $Z^t \leftarrow Z^{t-1}-\{x\}+\{y\}$。

步骤 3 此时我们得到了 T 个中心点集 $Z^0,Z^1,\cdots,Z^{T-1}$。对于这 T 个中心点集以正比于下式的概率选出一个最终解

$$\exp\left(-\varepsilon\frac{\mathrm{cost}(Z^t)}{8\Lambda^2(T+1)}\right),$$

即目标函数值越小的中心点集被选中的概率越大。

输出: 最终的中心点集 $Z=\{z_1,z_2,\cdots,z_k\}\subseteq F$。

3.1.3 基于集合覆盖技巧的差分隐私算法

之前章节提到的两个算法的近似比并不为常数, 即算法近似比的乘法项误差系数不为 $O(1)$, 这主要是因为 Matoušek[20] 的文章不能直接应用于隐私聚类问题, 即不能简单地应用他们的技巧产生满足差分隐私的近似中心点候选集 F, 所以在 $\mathbb{R}^d$ 上选中心点和在近似中心点候选集 F 上选中心点会导致一个较大的近似比发生。而由于大多数 k-均值问题的差分隐私算法是隐私机制与经典 k-均值问题算法的结合, 且经典 k-均值问题的算法已很难改进且皆为常数近似比。所以后续的工作大多集中在如何产生更好的隐私近似中心点候选集 F 上, 只有找出足够好的隐私近似中心点候选集 F 才可以将算法的近似比进行改进。

在最近的两篇文献[35, 36] 中, 作者通过多种技巧和算法如 JL 降维技巧、集合覆盖问题的一些算法再结合两种隐私机制成功的产生了更好的隐私近似中心点候选集 F, 从而成功把 CDP 下的 k-均值问题的近似比的乘法项误差系数改进到了 $O(1)$ 的程度。接下来我们对 Nguyen 等[36] 的算法做简单介绍, 他们算法的近似比为 $O(1)\times\mathrm{OPT}+O(\mathrm{poly}(\log(n),\sqrt{d},k))$。

算法 3[36]

输入: 数据集 $C=\{c_i\}_{i=1}^n \subseteq \mathbb{R}^d$, C 中元素的最大度量 $\| c_i \|_2 \leqslant \Delta$, 参数 ε, 中心点数量 k。

步骤 1 构造一个合适的JL降维映射 T' 把数据集 C 映射到一个维度为 d' 的空间上, 记降维后的数据集为 C'。

$$C' \leftarrow T'(C),$$
$$d' = O(\frac{\log n}{\varepsilon^2})。$$

步骤 2 设定参数, 令 $r_1 \leftarrow 1/n$, $t_1 \leftarrow \frac{\varepsilon}{n\sqrt{d'}}$。

从 $i=1$ 开始将算法2运行 m 次, $m=\lceil \log_{1+\varepsilon} 2n \rceil$。

$S_i \leftarrow$ 算法2(C', t_i, r_i)。

每次更新参数 $r_{i+1} \leftarrow (1+\varepsilon)r_i$, $t_{i+1} \leftarrow (1+\varepsilon)t_i$。

最后得到候选中心点集合 $S=\bigcup_{i=1}^m S_i$。

步骤 3 把所有 C' 中的点 c 都指派到最近的 S 中的点, 记为 $\mathrm{grid}[c]$。

对于所有的 S 中的点 s, 记指派到 s 上的 C' 中的点的数量为 n_s, 并把他们的周围点的数量估计加一个Laplace噪声得到一个新的周围点的数量估计值 n'_s。

$$n'_s \leftarrow n_s + \mathrm{Lap}(1/\varepsilon_L)。$$

把 S 中所有的点 s 原地复制 n'_s 次, 生成一个新的点集合记为 C''。

步骤 4 对 C'' 使用Lloyd算法得到新的中心点集合 $S''=\{s''_1, \cdots, s''_k\}$。把 C' 中的点指派到最近的 s''_i 上, 根据 s''_i 把 C' 划分成 k 个类, 第 i 类记为 C'_i。

对于这 k 个类分别运行算法3, 得到最后的 k 个中心点 $\widehat{s_i}$。

$$\widehat{s_i} = \text{算法 3 } (C, 1_{C'_i}, \varepsilon_G, \delta_G)。$$
$$\{1_{C'_i}(c) \text{ 表示 } T'(c) \text{ 是否在 } C'_i \text{ 内}\}。$$

输出: 最终的中心点集 $\widehat{S}=\{\widehat{s_1}, \widehat{s_2}, \cdots, \widehat{s_k}\}$。

算法的大致可以分为四步: (1) 构造一个合适的 JL 降维映射 T' 把数据集 C 映射到一个维度为 d' 的空间上, 记降维后的数据集为 C'。(2) 对降维后的数据集 C' 运行算法 2 得到近似中心点集合 S。其中算法 2 运用了指数集合和集合覆盖算法来找到合适的近似中心点集合。(3) 用 S 来构造一个新的近似的点集来代表原数据集 C', 构造方法就是把 C' 中的点全部指派到离其最近的近似中心点 s 上, 并把 s 点周围的点进行计数, 把计数加上 Laplace 噪声后记为 n'_s, 并把 S 中的点在原地复制 n'_s 次从而构造出新的点集 C''。这样构造出来的点集 C'' 可以近似的代表原数据集 C', 且保持隐私。(4) 对于构造出的近似数据集 C'' 运用任意的 k-均值问题的算法如经典的 Lloyd 算法得到中心点集合 $S''=\{s''_1, \cdots, s''_k\}$。再使用这些中心点 $S''=\{s''_1, \cdots, s''_k\}$ 把 C' 进行划分得到 k 个类 C'_i, 最后对每个类 C'_i 运行算法 3 得到最后的中心点 $\widehat{s_i}$。算法 3 是 Nissim 等[38] 于 2016 年提出的算法, 可以在带噪音和保持隐私的情况下估计聚类集合的平均值, 从而得到最后的聚类中心。

3.1.4 基于 CDP 的 k-均值问题的其他差分隐私算法

除以上几种算法外, 还有其他几种 CDP 下的 k-均值问题的差分隐私算法。

其中最早的是由 McSherry 等[24,39] 于 2005 年首次提出然后于 2009 年应用于差分隐私上的 DPLloyd 算法, 他们算法的主要思想是在原始的 Lloyd 算法上每一步均添加拉普拉斯噪声, 从而达到满足差分隐私需求的目的, 同时算法的迭代次数需要固定, 以决定每次迭代需要添加多少噪声。

其次有 Nissim 等[40] 于 2007 年提出的采样和聚合框架。在此框架下做 k-均值问题, 首先对于给定的输入数据集 C, 算法把数据集 C 分成 l 块, 记为 $C_1, \cdots, C_l$, 然后算法在每个块 $C_i(1 \leqslant i \leqslant l)$ 上分别计算对应的中心点集合 O_i, 最后它带隐私的将所有块的中心点集合 $\Sigma_1^l O_i$ 求平均和加噪音, 从而输出最后的结果。由于 C 中的任何单个元素都落入一个块中, 添加一个元素最多只能影响一个块的结果, 从而限制了聚合步骤的隐私敏感性。因此可以在最后一步添加更少的噪声以满足差分隐私。

还有 Zhang 等[41] 于 2013 年提出的基于遗传算法的差分隐私模型拟合框架 PrivGene, 给定一个数据集 C 和一个适应度打分函数 $f(C, \theta)$, 目标是找到最优的参数 θ^* 使得打分函数最大。PrivGene 算法初始化一组可能的参数 θ 并通过 GA 算法来不断选取新的 θ。具体来说, 在每次迭代过程中, PrivGene 算法使用指数机制从当前候选集合中选择具有最佳适应度打分的 m' 个参数, 并通过交叉和变异从 m' 个选择的参数中生成新的下一代候选集。他们将此框架应用于 k-均值问题并做了数值实验, 实验结果证明了其框架的优越性。

最后关于该问题最好的近似比的乘法项误差是 Ghazi 等[42] 发表于 2020 年的文章, 他们的文章首先通过 JL 降维技巧[43]、Kirszbraun 定理[44] 等把问题从原空间投影到一个维度为 $O(\log k)$ 的低维空间上, 然后通过一个名为 DensestBall 的技巧在该空间上找到一个较大的近似中心点集合, 接着在该近似中心点集合上通过 Feldman 等[45] 的技巧找到满足差分隐私的近似中心点集合, 最后在满足差分隐私的近似中心点集合上调用任意的非隐私 k-均值问题的算法得到最终结果。该文章已经把近似比的乘法项误差的系数改进到了 6.358，而其加法项误差为 $O(((kd + k^{O(\alpha(1))})/\varepsilon) \cdot \operatorname{poly}\log(n))$。

为了阅读和查询方便, 我们在表 1 中列出了本文提到的一些 CDP 下的 k-均值问题的算法结果对比。其中 α 表示一个很小的常数, η 表示任意的非隐私 k-均值问题的算法近似比, k 是聚类中心数, d 是空间维度, n 是输入数据集大小, ε 是隐私参数。

表 1 CDP 下的 k-均值问题的算法对比

文献	乘法项误差系数	加法项误差
Nock 等[34]	$O(\log(k))$	$O(n\Delta^2/(\varepsilon + \log n)^2)$
Balcan 等[28]	$O(\log^3(n))$	$O(\operatorname{poly}(\log(n), d, k))$
Nguyen 等[36]	$O(1)$	$O(\operatorname{poly}(\log(n), \sqrt{d}, k))$
Ghazi 等[42]	$(1+\alpha)\cdot\eta$	$O(((kd + k^{O(\alpha(1))})/\varepsilon) \cdot \operatorname{poly}\log(n))$

3.2 基于 LDP 的 k-均值问题的差分隐私算法

在这一节我们介绍局部差分隐私 (LDP) 模型下的 k-均值问题的算法。相比于集中式差分隐私 (CDP) 模型下的 k-均值问题, 近年来研究局部差分隐私 (LDP) 模型下的 k-均值问题的文章数量略少。在 LDP 模型下, 如图 2 所示, 用户在上报信息时就带了噪音, 因

而 LDP 的 k-均值问题的算法相比于 CDP 模型下的算法, 其更注重对于用户统计性能的保护, 需要对用户的真实数据进行一些统计上的估计, 从而在保护隐私的同时保持算法的性能。如果对用户的带噪音数据不加处理和估计直接使用, 可能会导致算法的性能较差。

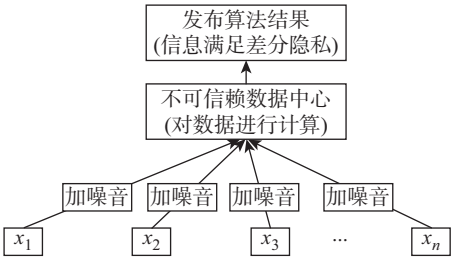

图 2 局部差分隐私(LDP)模型

3.2.1 基于噪声估计和哈希映射的差分隐私算法

这一节我们介绍 Stemmer[46] 于 2020 年提出的基于噪声估计和哈希映射的差分隐私算法。在之前工作如 Nissim 和 Stemmer[47], Kaplan 和 Stemmer[48] 的基础上, Stemmer[46] 的算法将问题的近似比改进为 $O(1)\times \mathrm{OPT}+O(n^{1/2+a}\cdot k\cdot \max\{\sqrt{d},\sqrt{k}\})$, 该算法也是目前 LDP 模型下 k-均值问题最好的算法之一。他们的算法大致分为以下三个部分: 1、在 $\mathbb{R}^d$ 上构造出性质较好的近似中心点候选集 F。2、对于近似中心点候选集 F 中的每一个点 $f\in F$, 用 $\#_C(f)$ 来表示 f 是离他们最近的中心点的那些用户点的数量。用 $\widehat{\#}_C(f)$ 来表示 $\#_C(f)$ 的带噪声的误差估计, 运用一些方法来估计出 $\widehat{\#}_C(f)$ 的值用以满足 LDP 的约束条件。3、对噪音估计 $\widehat{\#}_C(f)$ 和似中心点候选集 F 进行处理, 从中选出 k 个中心点以满足问题要求。我们将大致算法在下面列出, 具体参数和更详细的算法请看 Stemmer[46] 的文章。

算法 4[46]

输入: 失败概率 β, 隐私参数 ε,δ, 用户点输入 $S=(x_1,\cdots,x_n)$, $\|x_i\|_2\leqslant\Lambda$, 参数 t,d。

步骤 1 构造候选中心集: 以参数 t,ε,δ 及参数 $r=\Lambda,\Lambda/2,\Lambda/4,\cdots,\Lambda/n$, 在 S 上运行算法GoodCenters多次, 得到一系列的中心点集合 $Y_1^r,\cdots,Y_M^r$。把所有的 Y_m^r 的集合记为 Y, 即 $Y=\bigcup_{r,m}Y_m^r$。

步骤 2 根据算法把用户点指派到候选中心点 Y 上: 记指派方式为 $a(i,x_i)$。

步骤 3 估计候选中心点集合的权重: 用 $\varepsilon/4$-LDP 算法来对每一个候选中心点 $y\in Y$ 做一个周围点数量的估计 $\hat{a}(y)\approx a(y)\triangleq|\{i:a(i,x_i)=y\}|$。

步骤 4 重新把用户点指派到候选中心点 Y 上: 令 $W=\{y\in Y:\hat{a}(y)\geqslant\Omega(\varepsilon,n,d,\delta,\beta)\}$, 表示在隐私估计下周围用户点较多的候选中心点集合。同时记新的指派方式为 $b(i,x_i)$。对于在 W 中的中心点 $a(i,x_i)$, 令 $b(i,x_i)=a(i,x_i)$, 对于其他的 $a(i,x_i)$, 将 $b(i,x_i)$ 定义为 W 中离 x_i 最近的中心点。

步骤 5 再次估计候选中心点集合的权重: 用 $\varepsilon/4$-LDP 算法来对每一个候选中心点 $y\in W$ 做一个周围点数量的估计 $\hat{b}(y)\approx b(y)\triangleq|\{i:b(i,x_i)=y\}|$。

步骤 6 用指数机制选择下一步的候选中心点集合: 初始化 $Z_0=\varnothing$。从 $i=1$ 直到 $i=100k$ 重复以下步骤:

步骤 6.1 以正比于下式的概率从 W 中选出一点 w:

$$\hat{b}(w)\cdot\min_{z\in Z_{i-1}}\parallel w-z\parallel^2 。$$

步骤 6.2 $Z_i\leftarrow Z_{i-1}\cup\{w\}$。

步骤 7 放大成功概率：重复步骤 6。一共 $O(\log(1/\beta))$ 次, 把最后得到的数量为 $O(100k\cdot\log(1/\beta))$ 的中心点集合记为 Z。

步骤 8 再次估计候选中心点集合的权重：用 $\varepsilon/4$-LDP 算法来对每一个候选中心点 $z\in Z$ 做一个周围点数量的估计 $\hat{\zeta}(z)\approx\zeta(z)\triangleq|\{i:Z(x_i)=z\}|$。$Z(x_i)$ 表示 Z 中离 x_i 最近的点。

输出: $(Z,\hat{\zeta})$。

从大致算法可以看出算法的主要工作就在于构造候选中心集和对带噪声的用户点的位置进行反复的评估, 由于初始数据带噪音, 因而每一步都需要对用户点的位置进行估计以计算候选中心点附近的用户点数量。其中构造候选中心集的算法 GoodCenters, 它的思想是使用局部敏感的哈希函数对输入点进行哈希处理, 使靠的近的输入点大概率映射到同一个哈希值, 而离得远的点映射到其他的哈希值, 根据哈希值把周围拥有大量点的地方作为候选中心点。其中 $\varepsilon/4$-LDP 算法是对带噪音的点的位置进行估计, 其用到的思想是来源于处理带隐私的计数问题的相关算法[49], 其可以在带隐私的情况下估计出原输入的近似分布。

3.2.2 基于 LDP 的 k-均值问题的其他差分隐私算法

由于 Stemmer[46] 的算法包含了很多轮, 每一轮都需要用到上一轮的带隐私的输出作为输入, 同时其算法的近似比为 $O(1)\times\mathrm{OPT}+O(n^{1/2+a}\cdot k\cdot\max\{\sqrt{d},\sqrt{k}\})$, 而近似比中的乘法项常数也在 100 以上。为了解决这个问题在 2021 年的 ICML 上 Chang 等[50] 提出了一个 LDP 下的 k-均值问题的差分隐私算法, 他们的算法的近似比的乘法项误差系数为 6.358, 把之前 Stemmer[46] 的算法的乘法项误差降低到了一个很小的常数, 同时他们算法近似比的加法项误差为 $O(k^{O_\alpha(1)}\cdot\sqrt{nd}\cdot\mathrm{poly}(\log(\mathrm{n})/\varepsilon))$, 且他们的算法是一个一轮的非交互式算法。他们算法大致分为两步: (1) 使用 Har-Peled 等[51] 提出的名为 net tree 的分层数据结构来构造一个带隐私的近似候选中心点集合。(2) 在这个带隐私的候选中心点集合上运行经典的不带隐私的 k-均值问题的算法如 Ahmadian 等[2] 给出的近似比为 6.357 的近似算法来得到最终结果。文章中用来构造带隐私的近似候选中心点集合的这个树的数据结构中, 树的每个叶子代表问题的输入点, 树越深越能精准的代表整个输入集合, 同时为了隐私, 还需向树中注入噪声。对于这棵树, 树的叶子太少时对输入集合的拟合就会不精准, 树的叶子太多时噪音带来的误差就会很大。文章的主要贡献是提供了一种构建树的方法, 可以在以上两个方面取得平衡且不会对 k-均值问题的目标产生太大影响。同时他们还在使用混合高斯分布生成的数据集上对自己的算法进行了实验, 实验结果符合算法预期效果。

另有 Chaturvedi 等人[52] 于 2021 年提出了两个算法, 对于任意的 $\alpha>0$ 和 $c>\sqrt{2}$, 还有任意的非隐私 k-均值问题算法的近似比 η。他们文章的第一个算法的近似比为 $(1+\alpha)\cdot\eta\times\mathrm{OPT}+O(n^{1/2}\cdot\sqrt{d}\cdot k^{O(1/\alpha^2)})$。我们可以看出, 如果在这里使用 Ahmadian 等人[2] 给出的近似比为 6.357 的近似算法, 那么第一个算法的乘法项误差的系数和 Chang 等人[50]

的一样也会是 6.358, 同时他们第一个算法的近似比的加法项误差为 $O(n^{1/2}\cdot\sqrt{d}\cdot k^{O(1/\alpha^2)})$, 与 Chang 等人[50] 的工作相比差距不大。他们文章中的第二个算法把 Stemmer[46] 的结果中加法项误差 $O(n^{1/2+a}\cdot k\cdot\max\{\sqrt{d},\sqrt{k}\})$ 中的参数 n 的指数项从 $(1/2+a)$ 给降到了 $1/2$, 即他们第二个算法的加法项误差为 $O(n^{1/2}\cdot\sqrt{d}\cdot k^{1+O(1/(2c^2-1))})$, 乘法项误差系数仍是一个常数 $O(c^2)$。在这篇文章中他们主要使用的技术是 JL 降维技巧, 局部敏感哈希映射技巧, 以及 Bassily 等人[53] 最新提出的 Bitstogram 算法。Bitstogram 算法是在 LDP 情况下解决 heavy-hitters 问题的一个方法, 它和我们前面提到的 $\varepsilon/4$-LDP 算法类似, 其可以在带隐私的情况下估计出原输入的近似分布。

最后为了阅读和查询方便, 我们在表 2 中列出了本文提到的一些 LDP 下的 k-均值问题的算法结果对比。其中 α 表示一个很小的常数, η 表示任意的非隐私 k-均值问题的算法近似比, k 是聚类中心数, d 是空间维度, n 是输入数据集大小, ε 是隐私参数。

表 2 LDP 下的 k-均值问题的算法对比

文献	乘法项误差系数	加法项误差
Nissim 和 Stemmer[47]	$O(k)$	$O(n^{2/3+a}\cdot d^{1/3}\cdot k^{1/2})$
Kaplan 和 Stemmer[48]	$O(1)$	$O(n^{2/3+a}\cdot d^{1/3}\cdot k^{2})$
Stemmer[46]	$O(1)$	$O(n^{1/2+a}\cdot k\cdot\max\{\sqrt{d},\sqrt{k}\})$
Chaturvedi 等[52], 算法 1	$(1+\alpha)\cdot\eta$	$O(n^{1/2}\cdot d^{1/2}\cdot k^{O(1/\alpha^2)})$
Chaturvedi 等[52], 算法 2	$O(c^2)$	$O(n^{1/2}\cdot d^{1/2}\cdot k^{1+O(1/(2c^2-1))})$
Chang 等[50]	$(1+\alpha)\cdot\eta$	$O(n^{1/2}\cdot d^{1/2}\cdot k^{O_\alpha(1)}\cdot\mathrm{poly}(\log(n)/\varepsilon))$

4 讨 论

本文主要介绍了国际上目前对于 k-均值问题的差分隐私算法的研究现状, 对 k-均值问题、差分隐私概念和机制、差分隐私模型及不同模型下的差分隐私算法进行了较为详细的介绍。从目前结果可以看出, 无论是集中式差分隐私 (CDP) 模型下还是局部差分隐私 (LDP) 模型下的 k-均值问题, 相应的差分隐私算法近似比的乘法项误差系数均和不带隐私的 k-均值问题的结果十分接近了, 即为 6.358。后续值得进一步研究的方向有:

(1) 由于两种差分隐私模型下的算法近似比的乘法项误差均和不带隐私情况下的 k-均值问题的算法近似比几乎一致, 所以接下来在近似比方面还能改进的是算法的加法项误差。

(2) 目前大多数文献成果的算法时间复杂度较高, 还有一小部分没有算法的时间复杂度分析, 对于这方面我们可以继续研究的是如何在不影响性能的情况下尽量减少算法的运行时间。

(3) 目前关于 k-均值问题的变形问题, 如带惩罚、带异常点等问题的差分隐私算法研究还较少, 下一步可以研究这些问题的差分隐私算法。

参 考 文 献

[1] Dasgupta S. The hardness of k-means clustering [R]. San Diego: Department of Computer Science and Engineering, University of California, 2008: CS2008-0916.

[2] Ahmadian S, Norouzi-Fard A, Svensson O, et al. Better guarantees for k-means and euclidean k-median by primal-dual algorithms [J]. *SIAM Journal on Computing*, 2019, **49**(4): FOCS17-97-FOCS17-156.

[3] 张冬梅, 李敏, 徐大川, 等. k-均值问题的理论与算法综述 [J]. 中国科学: 数学, 2020, **50**(9): 1387-1404.

[4] Lu R, Zhu H, Liu X, et al. Toward efficient and privacy-preserving computing in big data era [J]. *IEEE Network*, 2014, **28**(4): 46-50.

[5] Wang T, Zheng Z, Rehmani M H, et al. Privacy preservation in big data from the communication perspective—A survey [J]. *IEEE Communications Surveys Tutorials*, 2018, **21**(1): 753-778.

[6] Yao X, Zhou X, Ma J. Differential privacy of big data: An overview [C]//*Proceedings of the 2nd IEEE International Conference on Big Data Security on Cloud, High Performance and Smart Computing and Intelligent Data and Security*, 2016: 7-12.

[7] Jain P, Gyanchandani M, Khare N. Differential privacy: its technological prescriptive using big data [J]. *Journal of Big Data*, 2018, **5**(1): 1-24.

[8] Skarkala M E, Maragoudakis M, Gritzalis S, et al. Privacy preservation by k-anonymization of weighted social networks [C]//*Proceedings of the 4th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012: 423-428.

[9] Zheleva E, Getoor L. Privacy in social networks: A survey [M]//*Social Network Data Analytics*, Boston: Springer, 2011: 277-306.

[10] Task C, Clifton C. What should we protect? Defining differential privacy for social network analysis [M]//*State of the Art Applications of Social Network Analysis*. Cham: Springer, 2014: 139-161.

[11] Gowtham M, Ahila S S. Privacy enhanced data communication protocol for wireless body area network [C]//*Proceedings of the 4th International Conference on Advanced Computing and Communication Systems*, 2017: 1-5.

[12] Li M, Lou W, Ren K. Data security and privacy in wireless body area networks [J]. *IEEE Wireless communications*, 2010, **17**(1): 51-58.

[13] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy [C]//*Proceedings of the 16th ACM SIGSAC Conference on Computer and Communications Security*, 2016: 308-318.

[14] Kaissis G A, Makowski M R, Rückert D, et al. Secure, privacy-preserving and federated machine learning in medical imaging [J]. *Nature Machine Intelligence*, 2020, **2**(6): 305-311.

[15] Kaissis G, Ziller A, Passerat-Palmbach J, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging [J]. *Nature Machine Intelligence*, 2021, **3**(6): 473-484.

[16] Chaturvedi A, Nguyễn H L, Zakynthinou L. Differentially private decomposable submodular maximization [C]//*Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021: 6984-6992.

[17] Mitrovic M, Bun M, Krause A, et al. Differentially private submodular maximization: data summarization in disguise [C]//*Proceedings of the 34th International Conference on Machine Learning*, 2017: 2478-2487.

[18] Rafiey A, Yoshida Y. Fast and private submodular and k-submodular functions maximization with matroid constraints [C]//*Proceedings of the 37th International Conference on Machine Learning*, 2020: 7887-7897.

[19] Gupta A, Ligett K, McSherry F, et al. Differentially private combinatorial optimization [C]//*Proceedings of 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, 2010: 1106-1125.

[20] Matoušek J. On approximate geometric k-clustering [J]. *Discrete and Computational Geometry*, 2000, **24**: 61-84.

[21] Dwork C. Differential privacy [C]//*Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming*, 2006: 1-12.

[22] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C]//*Proceedings of the 3rd Theory of Cryptography Conference*, 2006: 265-284.

[23] Mcsherry F, Talwar K. Mechanism design via differential privacy [C]//*Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, 2007: 94-103.

[24] McSherry F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis [C]//*Proceedings of the 35th ACM SIGMOD International Conference on Management of Data*, 2009: 19-30.

[25] Fisher C. Over 267 million facebook users reportedly had data exposed online [EB/OL]. [2021-12-20]. https://www.engadget.com/2019/12/19/facebook-data-exposed-online/.

[26] Daniel V, Kershner I. Personal data of all 6.5 million israeli voters is exposed [EB/OL]. [2021-12-20]. https://www.nytimes.com/2020/02/10/world/middleeast/israeli-voters-leak.html.

[27] Bassily R, Smith A, Thakurta A. Private empirical risk minimization: Efficient algorithms and tight error bounds [C]//*Proceedings of the IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014: 464-473.

[28] Balcan M F, Dick T, Liang Y, et al. Differentially private clustering in high-dimensional euclidean spaces [C]//*Proceedings of the 34th International Conference on Machine Learning*, 2017: 322-331.

[29] Dwork C, Roth A. The algorithmic foundations of differential privacy [J]. *Foundations and Trends in Theoretical Computer Science*, 2014, **9**(3/4): 211-407.

[30] Wang Y, Wang Y X, Singh A. Differentially private subspace clustering [C]//*Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015: 1000-1008.

[31] Su D, Cao J, Li N, et al. Differentially private k-means clustering [C]//*Proceedings of the 6th ACM Conference on Data and Application Security and Privacy*, 2016: 26-37.

[32] Feldman D, Xiang C, Zhu R, et al. Coresets for differentially private k-means clustering and applications to privacy in mobile sensor networks [C]//*Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2017: 3-16.

[33] Huang Z, Liu J. Optimal differentially private algorithms for k-means clustering [C]//*Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2018: 395-408.

[34] Nock R, Canyasse R, Boreli R, et al. k-variates++: more pluses in the k-means++ [C]//*Proceedings of the 33rd International Conference on Machine Learning*, 2016: 145-154.

[35] Jones M, Nguyen H L, Nguyen T D. Differentially private clustering via maximum coverage [C]//*Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021: 11555-11563.

[36] Nguyen H L, Chaturvedi A, Xu E Z. Differentially private k-means via exponential mechanism and max cover [C]//*Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021: 9101-9108.

[37] Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding [C]//*Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007: 1027-1035.

[38] Nissim K, Stemmer U, Vadhan S P. Locating a small cluster privately [C]//*Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2016: 413-427.

[39] Blum A, Dwork C, McSherry F, et al. Practical privacy: the SuLQ framework [C]//*Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2005: 128-138.

[40] Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis [C]//*Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, 2007: 75-84.

[41] Zhang J, Xiao X, Yang Y, et al. PrivGene: differentially private model fitting using genetic algorithms [C]//*Proceedings of the 13th ACM SIGMOD International Conference on Management of Data*, 2013: 665-676.

[42] Ghazi B, Kumar R, Manurangsi P. Differentially private clustering: Tight approximation ratios [C]//*Proceedings of the 34th Conference on Neural Information Processing Systems*, 2020: 33-93.

[43] Makarychev K, Makarychev Y, Razenshteyn I. Performance of johnson-lindenstrauss transform for k-means and k-medians clustering [C]//*Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019: 1027-1038.

[44] Kirszbraun M. ü ber die zusammenziehende und lipschitzsche transformationen [J]. *Fundamenta Mathematicae*, 1934, **22**(1): 77-108.

[45] Feldman D, Fiat A, Kaplan H, et al. Private coresets [C]//*Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, 2009: 361-370.

[46] Stemmer U. Locally private k-means clustering [C]//*Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2020: 548-559.

[47] Nissim K, Stemmer U. Clustering algorithms for the centralized and local models [C]//*Proceedings of the 29th Algorithmic Learning Theory*, 2018: 619-653.

[48] Stemmer U, Kaplan H. Differentially private k-means with constant multiplicative error [C]//*Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, 2018: 5431-5441.

[49] Hsu J, Khanna S, Roth A. Distributed private heavy hitters [C]//*Proceedings of the 39th International Colloquium on Automata, Languages, and Programming*, 2012: 461-472.

[50] Chang A, Ghazi B, Kumar R, et al. Locally private k-means in one round [C]//*IProceedings of the 38th International Conference on Machine Learning*, 2021: 1441-1451.

[51] Har-Peled S, Mendel M. Fast construction of nets in low-dimensional metrics and their applications [J]. *SIAM Journal on Computing*, 2006, **35**(5): 1148-1184.

[52] Chaturvedi A, Jones M, Nguyen H L. Locally private k-means clustering with constant multiplicative approximation and near-optimal additive error [EB/OL]. [2021-12-19]. https://doi.org/10.48550/arXiv.2105.15007.

[53] Bassily R, Nissim K, Stemmer U, et al. Practical locally private heavy hitters [J]. *Journal of Machine Learning Research*, 2020, **21**(16): 1-42.