

基于差分隐私的非等距直方图发布方法

杨磊^{1,2}, 郑啸^{1,2}, 赵伟^{1,2}

(1. 安徽工业大学计算机科学与技术学院 安徽 马鞍山 243032;

2. 安徽省工业互联网智能应用与安全工程实验室 安徽 马鞍山 243032)

摘要: 已有的基于差分隐私的直方图发布技术在利用直方图反映数据的真实分布特征时可能会出现“重拖尾”和“零桶”现象, 并且在数据量较多处“过于平缓”; 另外, 已有技术对原始直方图进行差分隐私保护时未考虑每个分组所蕴含的信息量大小不同。针对以上问题, 提出一种基于差分隐私的非等距直方图发布方法。首先, 利用经验分布函数根据数据稀疏性合理构建非等距直方图; 然后, 在非等距直方图上应用差分隐私保护技术对原始非等距直方图进行隐私保护; 最后, 根据非等距直方图的组距大小为每组设置隐私预算以提高每组数据的隐私性。实验结果表明, 所提方法在差分隐私下进行直方图发布时充分考虑了数据分布的稀疏性, 有效避免了直方图的“重拖尾”和“零桶”现象, 保证了所发布直方图反映数据分布特征的准确性; 并且为每组添加符合拉普拉斯(Laplace)机制的噪声时, 根据组距为每组设置合理的隐私预算, 在一定程度上提高了不同数据段的隐私性。

关键词: 差分隐私; 非等距; 直方图发布; 拉普拉斯机制; 隐私预算

中图分类号: TP309.2

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020035

Non-equal-width histogram publishing method based on differential privacy

YANG Lei^{1,2}, ZHENG Xiao^{1,2}, ZHAO Wei^{1,2}

1. School of Computer Science and Technology, Anhui University of Technology, Maanshan 243032, China

2. Anhui Engineering Laboratory for Intelligent Applications and Security of Industrial Internet, Maanshan 243032, China

Abstract: Existing histogram publishing technology based on differential privacy may show phenomenon of "retracting" and "zero bucket" when histogram is used to reflect the real distribution characteristics of data, and "too gentle" in the case of large data volume. In addition, the existing technology of the original histogram difference of privacy protection when not considering the amount of information of each group is different. In view of the above problems, a kind of non-equal-width histogram publishing method based on differential privacy was proposed. First of all, a non-isometric histogram based on the sparseness of the data should be

收稿日期: 2019-10-07; 修回日期: 2019-12-23

通信作者: 杨磊, 18255574841@126.com

基金项目: 安徽省重点研究与开发计划基金(201904a05020071)

Foundation Item: The Key R & D Program of Anhui Province, China (201904a05020071)

论文引用格式: 杨磊, 郑啸, 赵伟. 基于差分隐私的非等距直方图发布方法[J]. 网络与信息安全学报, 2020, 6(3): 39-49.
YANG L, ZHENG X, ZHAO W. Non-equal-width histogram publishing method based on differential privacy[J]. Chinese Journal of Network and Information Security, 2020, 6(3): 39-49.

reasonably constructed by empirical distribution function. Secondly, differential privacy protection technology should be applied to non-equal-width histogram to protect the privacy of the original non-equal-width histogram. Finally, the privacy budget should be set for each group according to the class widths of the non-equal-width histogram to improve the privacy of each group of data. The experimental results show that the sparseness of the data distribution is fully taken into account when using the proposed method to perform histogram publishing under differential privacy, effectively avoid the phenomenon of histogram with “retracting” and “zero barrels”, and the accuracy of the published histogram for reflecting the characteristics of the data distribution is guaranteed. Also, when adding noise in line with Laplace mechanism to each group, setting a reasonable privacy budget for each group according to the class widths to some extent increases the privacy of different data segments.

Key words: differential privacy, non-equal-width, histogram publishing, Laplace mechanism, privacy budget

1 引言

随着大数据时代的到来,数据量呈爆炸式增长,大数据以数据量大、数据类型多、处理速度快和价值密度低为特点影响着现代人的决策和生活。著名的咨询机构 IDC (Internet Data Center) 作出估测,人类社会产生数据的速度约每两年增加一倍^[1]。在大数据的时代背景下,对数据进行收集、发布和分析不仅为人们生活带来便利,而且提高了经济和社会效益。各大机构收集各种信息产生了海量数据,并以多种形式发布各种统计结果,其中包括生成能反映数据分布特征的直方图。直方图是一种能直观估计出数据分布特征的技术,直方图的关键在于组距和分组数,大多数的直方图发布技术是从如何选择分组数出发划分直方图,所产生的直方图多是等宽直方图,而一个良好的直方图需要考虑组距和分组数两个方面,因此,需要根据数据的稀疏程度合理划分直方图的组距。例如,医院大多选用电子病历系统,造成了海量的医疗数据。为了反映公民的健康状况,医疗机构会以各种形式发布医疗数据的统计结果。而医疗机构所拥有的这些数据包含许多敏感信息,如果直接发布直方图而不加以隐私保护会泄露个人敏感信息。大数据时代,攻击者拥有愈来愈多的背景知识,使攻击者盗取数据愈加简单。因此,如何在保证个人数据隐私安全性的前提下进行数据发布是当前的重要问题。

表1为某医院病人患 HIV 情况的原始数据库表,表1中的“?”表示除 Bob 的患病信息未知外,其他信息都为攻击者所知,此时称攻击者拥

有关于表1的最大背景知识。为预防疾病和推动社会进步并保护用户个人隐私,数据拥有者将会发布由表1所生成的统计直方图(如图1所示)以供数据分析、科学研究等。但若直接发布图1的统计直方图而不加以隐私保护,攻击者利用查询返回值进行差分攻击^[2]等,可能存在以下个人隐私泄露的情况。

表1 病人患 HIV 情况
Table 1 Status of patient with HIV

姓名	年龄	患 HIV 病
Alice	31	Yes
Bob	19	?
Carol	24	No
⋮	⋮	
Cari	57	No
Ben	16	Yes

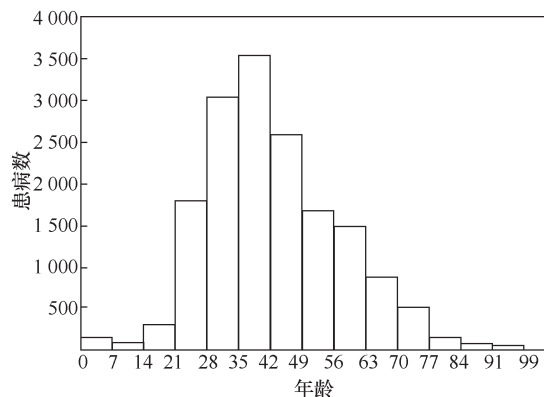


图1 HIV 患病数随年龄分布的原始直方图
Figure 1 The original histogram with the number of HIV distributed by age

假设对图1直方图作如下两种 SQL 统计查询:
SQL1: SELECT COUNT(*) FROM Table1

```
WHERE ( age BETWEEN 14 AND 21 )
AND (HIV = 'Yes')
SQL2: SELECT COUNT(*) FROM Table1
WHERE ( age BETWEEN 14 AND 21 )
AND (HIV = 'Yes')
AND (Name <> 'Bob')
```

SQL1 和 SQL2 查询结果分别为 265、264。

若攻击者只通过其中任一条 SQL 语句的结果, 则无法推断出任何关于个人的隐私信息。但若攻击者结合 SQL1 和 SQL2 的查询结果, 了解到年龄在 14~21 岁患 HIV 的人数为 265, 而除 Bob 外年龄在 14~21 岁患 HIV 的人数为 264, 则可推断 Bob 患 HIV, 由此造成 Bob 的隐私信息被泄露。

针对隐私问题, 所提出的隐私保护模型有很多, 如 k-anonymity^[3]、l-diversity^[4]、t-closeness^[5] 等被相继提出, 以应对拥有不同背景知识的攻击者所发起的各种新型的攻击方式。但这些早期的隐私保护模型一方面需要根据新攻击方法的出现不断改进, 无法应对图 1 所示的最大背景知识攻击; 另一方面它们无法提供一种严格有效的量化评估标准来证明自己的隐私保护水平。基于此, Dwork^[6] 针对统计数据库中数据隐私信息泄露问题提出一种新的隐私保护技术——差分隐私, 它能抵御任意背景知识攻击并发布统计数据时保护数据库中单条记录不泄露。

2 相关工作

自 2006 年 Dwork 等^[7] 提出差分隐私并实现拉普拉斯机制后, 不少学者相继基于差分隐私保护技术提出直方图数据发布技术。针对直方图发布技术, 最早的研究方法是直接向直方图中添加拉普拉斯噪声 (LP^[7] 方法), 但此方法会导致数据可用性低, 长范围查询时噪声过大。随后, 不断有学者对此进行改进, 解决长范围查询造成噪声累加而导致数据可用性低的问题。Xu 等^[8] 提出 NoiseFirst 和 StructureFirst 两种方法, NoiseFirst 方法首先对原始直方图添加拉普拉斯噪声, 再利用 V-Optimal^[9] 方法将直方图分组, StructureFirst 方法首先使用指数机制选择直方图桶的边界将直方图分组, 再对每个分组添加拉普拉斯噪声。这两种方法只考虑了分组造成的重构误差。针对这

些不足, 有学者从直方图结构的角度出发, 使用聚类方法对直方图进行各式地重构, Zhang 等^[10] 提出 OCA 算法、ECA 算法、AHP 算法 3 种聚类方法来合并直方图以减少误差, 但该方法由于对原始直方图加噪后再将直方图排序, 在排序上造成时间复杂度较高。此外, 张啸剑等^[11] 提出了 DiffHR 算法, 这是一种基于马尔可夫链的有效排序方法, 先使用指数机制的抽样技术抽样直方图桶将原始直方图逼近正确排序, 再对此直方图进行自适应贪心聚类以减少由于直方图重构和添加噪声引起的直方图与原始直方图的误差, 实现直方图更精确的发布。Hui 等^[12] 提出 IPHP 算法, 使用二分法分配隐私预算, 先将整体数据分为两部分, 再对这两部分分别继续进行划分, 如此迭代直到返回最终划分结果。唐海霞等^[13] 对以上算法进行了研究并改进, 提出 APB 算法, 使用自适应隐私预算分配对直方图添加两次噪声进行保护, 先对直方图进行排序再使用聚类划分得到最终分组。另外, 就隐私预算分配问题, 国内外专家学者作了较深研究并先后提出多种隐私预算分配方法。Hui 等^[12] 提出使用二分法分配隐私预算, 先将整体数据分为两部分, 再对这两部分分别继续进行划分, 如此迭代直到满足迭代终止条件。但二分法每次迭代取上次隐私预算的一半, 这种分配方法导致每次可使用的预算少从而数据可用性低, 并且未考虑多次加噪过程中的隐私预算分配问题。基于此, 唐海霞等^[13] 设计一种自适应的隐私预算分配方法, 通过权衡噪声误差和重构误差构建隐私预算权重分配模型来为添加噪声和重构直方图分配权重, 并利用贪心算法对直方图进行预分组, 再对直方图进行重构并加噪。虽然以上方法在查询精度上有一定的提升, 但所针对的是某一个数据使用者进行统计数据查询时的隐私预算分配的情况。在实际应用中, 数据的使用者可能有多方, 若让每个使用者任意使用隐私预算会导致某一个数据使用者耗尽隐私预算, 进而导致其他数据使用者无法进行隐私数据查询。针对此问题, Bkakra 等^[14] 就多个数据使用者的隐私预算分配问题进行探讨, 提出一种典型的考虑多参与方的隐私预算分配的差分隐私模型。

尽管以上这些方法通过将直方图合理分组重

构或合理分配隐私预算在发布统计结果时保护了单条数据隐私不被泄露并保证了一定的数据可用性,但其直方图都为等宽直方图,在一定程度上隐藏了数据的一些重要信息,即未能考虑数据的稀疏性,可能造成在数据量较少处出现“重拖尾”和“零桶”现象;而在数据量较多处出现直方图“过于平缓”等现象,导致无法利用直方图的良好性质反映出数据明显的分布特征。在数学领域,杨军等^[15]提出一种通过均匀经验分布函数及其广义逆函数构建非等距直方图的方法,但该方法均匀生成经验分布函数,在医疗健康数据中并不能得到很好的分组结果,甚至与等距直方图分组结果差距不大,并且若直接发布该直方图,会造成严重的数据隐私泄露风险。

本文针对直方图结构和隐私预算分配问题,提出一种基于差分隐私的非等距直方图发布方法。该方法将差分隐私保护技术与非等距直方图相结合,并在差分隐私下进行直方图发布时充分考虑数据分布的稀疏性,保证了所发布直方图对于反映数据分布特征的准确性。另外,利用经验分布函数及其逆函数直接确定分组,并且在应用差分隐私技术对每组添加拉普拉斯噪声时,结合数据分布的稀疏程度,根据组距为每组设置合理的隐私预算,提高数据信息的隐私性,并最大限度保留数据的分布特征,且一定程度上保证长范围查询的精度。

3 定义及相关概念

3.1 差分隐私

差分隐私^[6]与以往隐私保护技术不同的是:一方面,差分隐私对隐私保护进行了严格的数学定义,并为隐私保护程度提供了量化评估方法,使不同参数下的数据集所提供的隐私保护水平具有可比性;另一方面,差分隐私假定恶意攻击者拥有最大的背景知识,即攻击者拥有除目标记录外其他所有记录的信息。因此,差分隐私无须考虑攻击者的背景知识,因为它不能提供比最大背景知识更多的背景知识。

定义 1 ε -差分隐私^[16]。设数据集 D 和 D' 完全相同或最多只相差一条记录(称数据集 D 和 D' 为近邻数据集或兄弟数据集), $\text{Range}(M)$ 是随机算法 M 的输出域。对于任意两个近邻数据集 D

和 D' 及算法输出域的任意子集 $S \in \text{Range}(M)$, 若算法 M 满足

$$P_r[M(D) \in S] \leq e^\varepsilon P_r[M(D') \in S] \quad (1)$$

则称随机算法 M 提供 ε -差分隐私。其中, P_r 表示隐私被公开的可能性,参数 ε 是隐私保护强度参数,也被称为隐私保护预算^[6],它是衡量隐私保护水平的关键因素之一,参数 ε 越小,隐私保护程度越高。衡量隐私保护水平的另一个关键因素是敏感度^[17]。

定义 2 全局敏感度^[6]。对于任意函数 $f: D \rightarrow R^d$, 其全局敏感度 Δf 定义为

$$\Delta f = \max_{D, D'} \|f(D) - f(D')\|_1 \quad (2)$$

其中, D, D' 是近邻数据集, d 是函数 f 的查询维度, $\|\cdot\|_1$ 表示向量的 L_1 范式, R 表示查询函数 f 所映射的实数空间。

3.2 直方图相似度

定义 3 直方图相似度: 设两直方图之间的相似度为 $d(h, h')$, 如式(3)所示。 $\bar{h}$ 为直方图 h

桶高的平均值 $\bar{h} = \frac{\sum_{i=1}^k h_i}{k}$, $\bar{h}'$ 为直方图 h' 桶高的平

均值 $\bar{h}' = \frac{\sum_{i=1}^k h'_i}{k}$, h_i 为直方图 h 的第 i 个桶高, h'_i 为直方图 h' 的第 i 个桶高, $i=1, 2, \dots, k$ (k 为分组数)。

直方图相似度由相关性比较确定。

$$d(h, h') = \frac{\sum_{i=1}^k (h_i - \bar{h})(h'_i - \bar{h}')}{\sqrt{\sum_{i=1}^k (h_i - \bar{h})^2 \sum_{i=1}^k (h'_i - \bar{h}')^2}} \quad (3)$$

4 基于差分隐私的非等距直方图发布方法设计

非等距直方图的关键是分组数和组距大小的确定,分组数由 Moore 公式^[18] ($k \approx C \cdot n^{\frac{2}{5}}$) 确定,其中, C 为取值 1~3 的常数, n 为数据集大小。

4.1 算法设计

本文考虑到传统基于差分隐私保护的直方图发布方法所发布的都是等距直方图,无法精确反映数据分布特征的问题,设计了一种基于差分隐私的非等距直方图发布方法(NHPDP, non-equal-

width histogram publishing method based on differential privacy), 根据原始数据稀疏性合理构建非等距直方图, 并利用组距大小分配隐私预算对直方图添加噪声实现直方图的隐私保护, NHPDP的总体设计如图2所示。

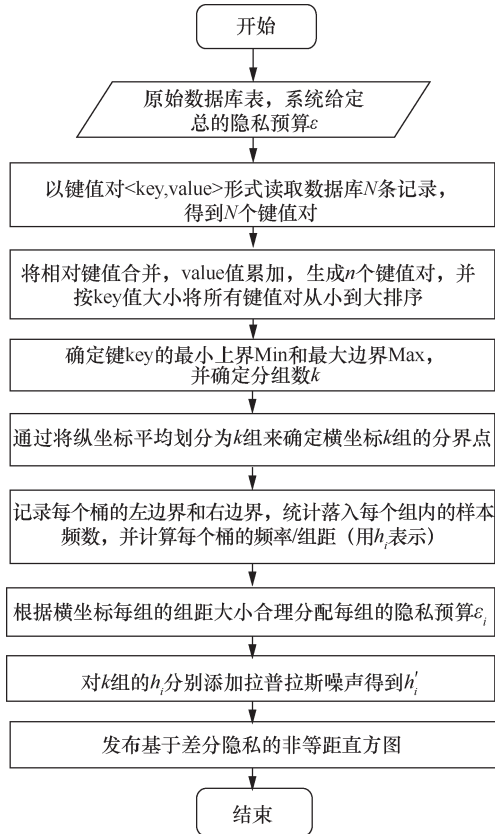


图2 NHPDP的总体设计
Figure 2 Overall design of NHPDP

基于以上分析, 基于差分隐私的非等距直方图发布方法具体实现细节如算法1所示。

算法1 NHPDP

输入 数据键值对 $kv = \{ \langle key_1, value_1 \rangle, \langle key_2, value_2 \rangle, \dots, \langle key_n, value_n \rangle \}$ 、隐私预算 ε 。

输出 满足差分隐私的待发布的非等距直方图 h' 。

1) $kv = \text{sorted}(kv)$; /对所有键值对按键从小到大进行排序/

2) $Max = \max \{ key_i | 1 \leq i \leq n \}$, $Min = \min \{ key_i | 1 \leq i \leq n \}$; /确定最小上界和最大下界/

3) $C = 2$, $k \approx C \cdot n^{\frac{2}{5}}$; /确定较优分组数 k /

4) $B = \text{GetdeltaB}()$; /得到直方图横坐标分组分界点/

5) Calculate $\Delta B_i = B_i - B_{i-1}$, $(B_{iL}, B_{iR}) = (B_{i-1}, B_i)$, $n_i = \sum_{key \in (B_{iL}, B_{iR})} \text{value}_{(key)}$, $f_i = \frac{n_i}{\sum_{j=0}^n \text{value}_j}$; /

得到每组组距 ΔB_i 及组区间 (B_{iL}, B_{iR}) 、每组的频数和 n_i 、每组频率值 f_i /

6) for each (B_i, B_{i-1}) do

7) return $h = \{h_1, h_2, \dots, h_k\}$, where $h_i =$

$\frac{f_i}{\Delta B_i}$; /每组计算频率/组距, 生成非等距直方图/

8) for each $h_i \in h$ do

9) $h'_i = h_i + \text{Laplace}\left(\frac{\Delta f}{\varepsilon_i}\right)$; /对直方图添加 Laplace 噪声/

10) return $h' = \{h'_1, h'_2, \dots, h'_k\}$;

GetdeltaB()方法见 4.2 节, Δf 与 ε_i 值的确定见 4.3 节。

4.2 构建非等距直方图

本文考虑到使用均匀的经验分布函数构建非等距直方图与等距直方图结果无异, 不能很好地反映出非等距直方图效果的问题, 设计一种构建非等距直方图 (CNH, construct non-equal-width histogram) 的方法, 利用数据本身的稀疏性使用不均匀的经验分布函数构建非等距直方图并得到每组的分界点。基于以上分析, 构建非等距直方图方法具体细节如算法2所示。

算法2 CNH 方法

输入 数据键值对 $kv = \{ \langle key_1, value_1 \rangle, \langle key_2, value_2 \rangle, \dots, \langle key_n, value_n \rangle \}$ 。

输出 B

1) function GetdeltaB ()

2) for each $key \in (key_i, key_{i+1})$ do

3) $F = \{0, \dots, \frac{\sum_{j=0}^i \text{value}_j}{\sum_{j=0}^n \text{value}_j}, \dots, 1\}$; /*生成经验分布函数每段的取值*/

4) for $i \leq k$ do

5) if $\frac{\sum_{j=0}^i \text{value}_j}{\sum_{j=0}^n \text{value}_j} \geq \frac{i}{k}$

6) return key_i; /返回 i 对应的 key 值/

7) key_i → B; /将 key 值写入 B 中/

8) else i++;

9) return B; /得到每组分界点/

4.3 隐私预算分配

根据差分隐私中对全局敏感度及隐私预算的定义, 非等距直方图中原始数据集某一条记录的改变对于直方图发布统计计数查询函数 f 的查询结果的最大改变就是 f 的敏感度, 其敏感度为 $\Delta f = \max_i \frac{N - n_i}{N(N-1)\Delta B_i}$, 其中, N 为总数据集的大小。下面给出敏感度的详细推算过程。

对于计算查询 f 的敏感度, 由定义 2 可知, 有

$$\begin{aligned} \Delta f &= \max_{D, D'} f(D) - f(D')_1 = \\ &= \max_i \left| \frac{\frac{n_i}{N} - \frac{n_i - 1}{N-1}}{\Delta B_i} \right| = \max_i \left| \frac{\frac{n_i}{N} - \frac{n_i - 1}{N-1}}{\Delta B_i} \right| = \\ &= \max_i \left| \frac{(N-1)n_i - N(n_i - 1)}{N(N-1)\Delta B_i} \right| = \\ &= \max_i \left| \frac{N - n_i}{N(N-1)\Delta B_i} \right| = \\ &= \max_i \frac{N - n_i}{N(N-1)\Delta B_i}, (N \geq n_i) \end{aligned}$$

在对直方图添加噪声时, 给定总隐私预算 ε , 对每组的桶高添加拉普拉斯^[7]噪声, 即 $h'_i = h_i + \text{Laplace}\left(\frac{f}{\varepsilon_i}\right), i=1, 2, \dots, k$, 其中, ε_i 表示第 i 个分组的隐私预算大小; 而由于每组数据的稀疏程度不同, 包含的隐私信息量也不同, 因此需要对每组数据根据数据的特征进行合理的隐私保护, 提高数据的隐私性。而提高数据隐私性意味着必然要“牺牲”一部分数据可用性, 为保证数据的隐私性和可用性之间的平衡, 本文根据组距大小对每组分别添加噪声。由差分隐私的“并行

组合性”^[17]可知, 若每组数据互不相交, 且保证每组的隐私预算满足 $\varepsilon_i \leq \varepsilon$, 则满足 ε -差分隐私。由于组距反映了数据的稀疏程度, 组距小的直方图桶的数据较密集, 并且在长范围查询时噪声的累加会导致查询结果精度降低, 因此对组距小的桶相较于组距大的桶应添加较少噪声, 而 ε 越小, 添加越多噪声的可能性越大。基于以上分析确定 ε_i 。

$$\varepsilon_i = \varepsilon \cdot \left(1 - \frac{\Delta B_i}{\sum_{i=1}^k \Delta B_i} \right), i=1, 2, \dots, k \quad (4)$$

5 算法隐私性与可用性分析

算法的隐私性从差分隐私的定义及性质上证明, 可用性利用直方图最终的发布误差及直方图的“相关性比较”进行度量分析。本节主要论证 NHPDP 在根据组距大小设置隐私预算上满足 ε -差分隐私, 并与 DiffHR 和 APB 方法进行可用性对比分析, 通过实验对本文的算法进行验证。

5.1 隐私性分析

定理 1 在隐私预算分配方法中, 利用组距大小按比例分配隐私预算, 满足 ε -差分隐私。

证明 根据差分隐私的并行组合性, 每个分组满足 ε_i -差分隐私, 由于每个分组数据集互不相交, 且对于任意的 i 都有 $\varepsilon_i \leq \varepsilon, i=1, 2, \dots, k$, 因此, NHPDP 满足 $(\max \varepsilon_i)$ -差分隐私, 即 NHPDP 满足 ε -差分隐私。证毕。

5.2 可用性分析

可用性采用直方图发布误差及直方图的“相关性比较”进行度量分析。NHPDP 的总体误差由重构的近似误差和拉普拉斯噪声误差构成。DiffHR、APB 方法与 NHPDP 所发布的直方图的可用性对比分析如表 2 所示。

如表 2 所示, NHPDP 第二项 $\frac{2}{|C_i|\varepsilon^2}$ 中的 ε 表

示总的隐私预算, DiffHR 与 APB 方法都是将隐私预算分为两部分: 第 1 部分是使用隐私预算 ε_1 保护原始直方图结构, 而第 2 部分则使用隐私预算 ε_2 添加噪声。因此, 给定总的隐私预算下, NHPDP 的噪声误差小于 DiffHR 和 APB 方法所产生的噪声误差。

表2 直方图发布过程中的发布误差
Table 2 Errors of grouping on histogram publishing

算法	Error(C_1)
DiffHR	$\sum_{H_j \in C_i} (H_j - \bar{C}_i)^2 + \frac{2}{ C_i \epsilon^2}$
APB	$\sum_{H_j \in C_i} (H_j - \bar{C}_i)^2 + \frac{2}{ C_i \epsilon^2}$
NHPDP	$\sum_{H_j \in C_i} (H_j - \bar{C}_i)^2 + \frac{2}{ C_i \epsilon^2}$

在发布误差的第一项 $\sum_{H_j \in C_i} (H_j - \bar{C}_i)^2$ 中, 由于每种算法该部分表达式相同, 因此计算表达式与具体的算法无关, 而分组时产生的重构误差与算法好坏有直接关联。NHPDP 是通过经验分布函数直接确定直方图分组进行直方图划分。DiffHR 方法首先使用抽样技术和指数机制对直方图进行排序使直方图接近正确排序, 再使用基于懒散下界的聚类划分方法合并桶以重构直方图。APB 方法使用自适应隐私预算分配对直方图添加两次噪声进行保护, 先对直方图进行排序再使用聚类划分得到最终分组。这 3 种方法由于将计数相同的桶放在了同一分组中, 因此很大程度上降低了分组引入的重构误差。而 NHPDP、DiffHR 及 APB 方法发布误差的关键取决于 NHPDP, 由于该方法对“重拖尾”和“零桶”的桶进行了合并, 因此在长范围查询时优于后两者。

6 实验分析

本文实验硬件环境: 实验部署在虚拟机 VMware 上, 操作系统为 Ubuntu 16.04。采用 C++ 语言编程, 使用 Matlab 生成实验图表。

本文实验采用 2010 年中国医疗健康普查结果中 0~99 岁年龄艾滋病病人患病人的数据, 0~99 岁总患病人数为 15982, 根据公共卫生科学数据中心发布的 2010 年艾滋病患病数与年龄段的统计表格, 进行数据的预处理, 在整体大致服从正态分布并保证每个年龄段总人数不变的前提下随机生成每个年龄段的患 HIV 病人人数 (实验中取 $C=2$ 时为最佳分组数量), 合并排序后得到以下顺序统计量, 如表 3 所示。

(1) 不同隐私预算分配方式对比

本节分别取 $\epsilon = 0.01, 0.1, \ln 2, 1, \ln 3, \ln 4$ 对两组直方图进行直方图相关性比较, 为了减少实验

偶然性对实验结果造成的影响, 实验中对每个 ϵ 分别多次添加噪声并取平均值。

表3 预处理后的顺序统计量
Table 3 Order statistics after preprocessing

<0,5>,<1,11>,<2,21>,<3,17>,<4,12>,<5,19>,<6,23>,<7,10>,<8,6>,<9,13>,<10,6>,<11,4>,<12,10>,<13,5>,<14,12>,<15,17>,<16,24>,<17,25>,<18,38>,<19,52>,<20,97>,<21,129>,<22,168>,<23,216>,<24,278>,<25,309>,<26,321>,<27,356>,<28,387>,<29,423>,<30,427>,<31,433>,<32,440>,<33,444>,<34,469>,<35,483>,<36,495>,<37,509>,<38,522>,<39,537>,<40,502>,<41,481>,<42,444>,<43,411>,<44,367>,<45,362>,<46,347>,<47,341>,<48,312>,<49,295>,<50,252>,<51,240>,<52,221>,<53,213>,<54,210>,<55,224>,<56,235>,<57,243>,<58,223>,<59,222>,<60,209>,<61,186>,<62,149>,<63,127>,<64,121>,<65,125>,<66,123>,<67,122>,<68,119>,<69,116>,<70,101>,<71,87>,<72,82>,<73,76>,<74,59>,<75,55>,<76,44>,<77,40>,<78,34>,<79,26>,<80,17>,<81,13>,<82,8>,<83,14>,<84,7>,<85,3>,<86,1>,<87,0>,<88,2>,<89,0>,<90,0>,<91,2>,<92,0>,<93,1>,<94,0>,<95,0>,<96,1>,<97,0>,<98,0>,<99,0>

由于数据稀疏性不同, 所蕴藏的隐私信息量也不同, NHPDP 根据组距大小对直方图添加噪声以提高数据的隐私性, 而提高隐私性可能意味着“牺牲”一部分的数据可用性。为了说明算法的隐私性与可用性, 实验在不同隐私预算 (epsilon) 下对比分析两种隐私预算分配方式产生的误差, 用添加噪声后的直方图与原始直方图之间的均方误差 (MSE, mean squared error) 来衡量, 结果如图 3 所示。另外, 为了说明 NHPDP 在提高数据隐私性的同时很好地保证了数据的可用性, 实验将由原始直方图 h 根据不同隐私预算添加噪声得到的直方图 h' 和使用相同隐私预算添加噪声得到的直方图 h'' 分别与原始直方图进行相关性比较, 衡量噪声不同隐私预算下直方图与原始直方图之间的相似度 (similarity), 结果如图 4 所示。

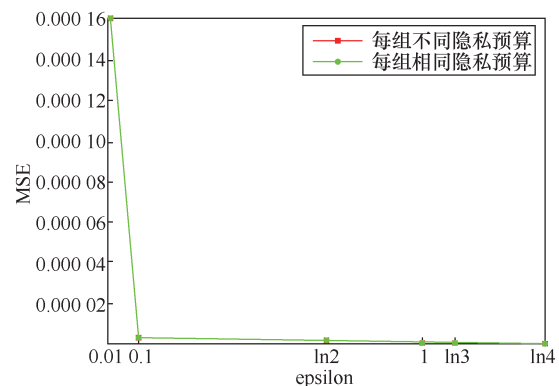


图3 不同隐私预算分配下产生的噪声误差
Figure 3 Noise errors with different privacy budget allocations

从图3可看出, 尽管实验过程中减少了分配给每组的隐私预算, 但与未减少隐私预算对比发现, 所产生的噪声误差并未增大, 说明即使在理论上减少每组的隐私预算以提高隐私性可能会“损失”可用性, 但这种损失微乎其微。

从图4的实验比较结果, 可以明显看出 NHPDP 在每个分组上根据组距大小使用不同隐私预算添加噪声所产生的直方图和使用相同隐私预算添加噪声所产生的直方图, 两者与添加噪声前的直方图的相似度差距不大, 甚至大部分结果中, 前者都优于后者, 这进一步说明了 NHPDP 保护了不同稀疏度分组数据的隐私性。

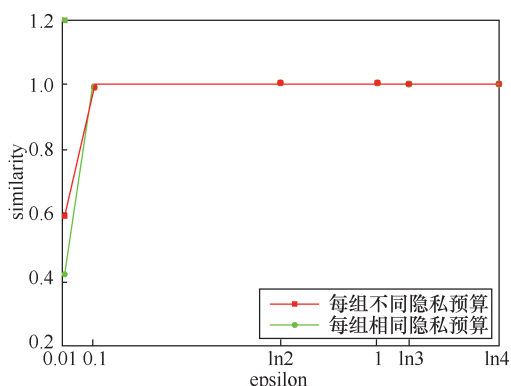


图4 噪声直方图与原始直方图之间的相似度

Figure 4 Similarity between histograms with noise and without noise

(2) 直方图分布特征对比

为反映 NHPDP 对数据总体分布特征的保护, 实验将非等距噪声直方图与等距噪声直方图进行对比分析。生成的非等距噪声直方图如图5所示, 传统方法生成的等距噪声直方图如图6所示。

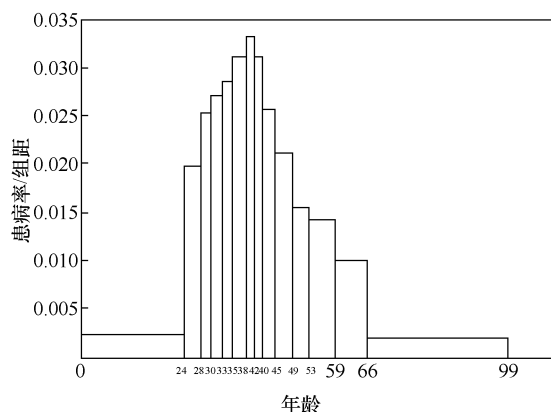


图5 2010年中国 HIV 患病率随年龄分布的非等距噪声直方图
Figure 5 The non-equal-width histogram with noise of the prevalence of HIV distributed by age in China in 2010

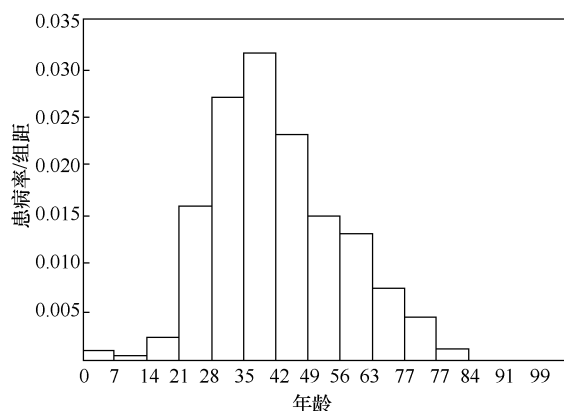


图6 2010年中国 HIV 患病数随年龄分布的等距噪声直方图
Figure 6 The equal-width histogram with noise of the prevalence of HIV distributed by age in China in 2010

从图5与图6可以看出, 传统的等距直方图可能导致“重拖尾”和“零桶”现象, 并且生成的直方图在数据量较多处过于平坦。实验结果中, 年龄在66~99岁的患病人数相较总患病人数较少, 表现在图6的等距直方图中, 导致出现了84~91岁、91~99岁这样的样本频数为0的区间, 并且存在多个这样的“零桶”。图5相较于图6直方图划分更合理, 更能反映出数据的分布特征。例如, 实验结果中, 年龄在24~49岁的患病人数居多, 每个年龄点的人数分布都不同, 并且通过图5可以看出年龄在38~40岁患病人数最多, 且最大值出现在38~40岁, 符合原始数据的分布情况, 而在图6中, 年龄在24~48岁直方图过于平缓, 无法反映出24~48岁患病人数的变化情况; 因此, NHPDP 能在保护隐私的前提下更好、更精确地反映数据变化趋势及分布特征。

(3) 均方误差对比

不同隐私预算 ϵ 下 NHPDP 与 APB 方法的发布误差是衡量可用性的关键。为便于比较, 实验过程中将 APB 方法生成的数据转换成 NHPDP 相同格式数据。由于 APB 方法建立在 DiffHR 方法的基础上, 因此下面对不同 ϵ 下 NHPDP 与 APB 方法的均方误差进行对比, 实验结果如图7所示。

从图7可以看出, NHPDP 方法较 APB 方法, 在一定程度上减少了误差, 提高了数据可用性。尤其是在隐私预算小于0.1时, APB 方法的变化较为明显, 这是由于 APB 方法在确定 k 值时将 ϵ

划分成了两部分，一部分用于第一次对直方图桶添加拉普拉斯噪声进行保护，另一部分对直方图每组的均值添加拉普拉斯噪声， ϵ 越小，划分之后的 ϵ 越小，添加的噪声越大，噪声受 k 值影响越大。但当隐私预算越来越大时，NHPDP 与 APB 方法的误差逐渐稳定，这是由于随着隐私预算增大，APB 方法的隐私预算分配使其添加的噪声更小，越来越接近 NHPDP，而 NHPDP 直接使用经验分布函数确定分组，噪声误差只在直方图添加拉普拉斯过程中产生，因此其均方差误差较为稳定。

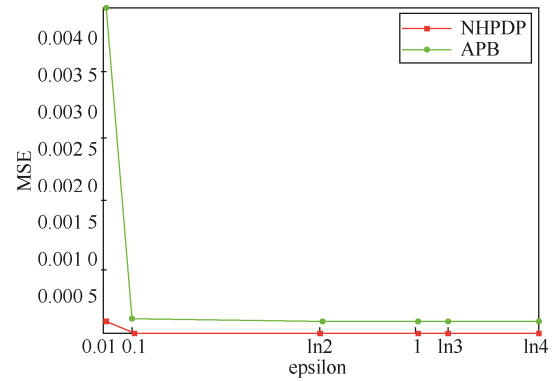


图7 ϵ 变化时 MSE 的变化
Figure 7 MSE under different ϵ

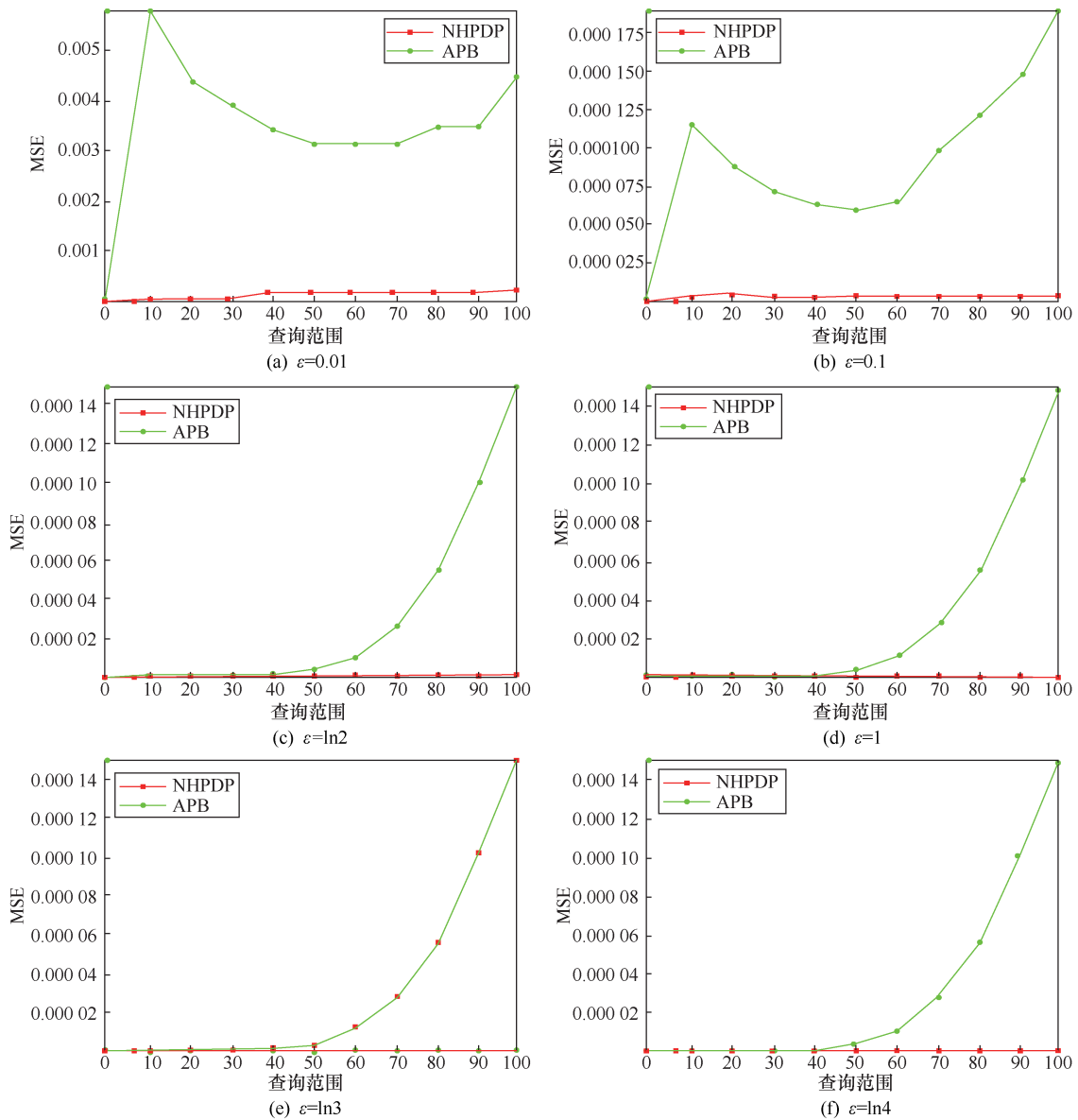


图8 不同 ϵ 下查询范围均方差
Figure 8 MSE of range size under different ϵ

(4) 查询范围均方差

不同查询范围下 NHPDP 和 APB 方法响应查询范围的均方差误差映了方法所发布的直方图的可用性。为便于比较,实验过程中将 APB 方法生成的数据转换成 NHPDP 相同格式数据。实验结果如图 8 所示。从图 8 可看出, NHPDP 在 ε 较小时,长范围查询优势较为明显,而在 ε 稍大时前半段范围查询时无明显优势,但随着查询范围增长,差距逐渐拉大,优势逐渐明显。因此, NHPDP 能在一定程度上保证长范围查询的精度。

7 结束语

本文针对差分隐私下的非等距直方图发布问题,主要完成了以下工作。

针对等距直方图不能反映出数据真实分布特征,并且利用均匀经验分布函数设置分组点时与等距直方图无明显区别问题,本文设计利用数据稀疏性生成不均匀经验分布函数构建非等距直方图,并将差分隐私保护技术与非等距直方图相结合保护数据隐私。

针对保护数据隐私性中,不同数据段所蕴含的隐私信息量大小不同所需进行隐私保护的程度也不同的问题,本文设计根据组距大小进行隐私预算分配方法合理分配隐私预算以保护不同数据段的隐私信息。

本文方法针对非等距直方图的隐私保护技术,在一定程度上较为合理地保护了不同稀疏度的分组数据的隐私性,并且较好地反映了数据的总体分布特征。但本文方法由于直接对非等距直方图的每个分组添加噪声而导致的噪声误差仍有改进余地。并且现在的数据采集大多从各种物联网终端采集数据,在物联网环境下,数据以动态流数据形式存在,下一步将对差分隐私下动态流数据的非等距直方图展开研究。

参考文献:

- [1] 林子雨. 大数据技术原理与应用[M]. 北京: 人民邮电出版社, 2015.
- [2] LIN Z Y. BigData technology and application[M]. Beijing: Posts and Telecom Press, 2015.
- [3] 李森有, 季新生, 游伟. 基于置信度分析的差分隐私保护参数配置方法研究[J]. 网络与信息安全学报, 2019(4): 29-39.
- [4] LI S Y, JI X S, YOU W. Research on differential privacy protection parameter configuration method based on confidence level[J]. Chinese Journal of Network and Information Security, 2019, 5(4): 29-39.
- [5] SWEENEY L. K-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.
- [6] MACHANAVAJJHALA A, GEHRKE J, KIFER D, et al. L-diversity: privacy beyond k-anonymity[C]//22nd International Conference on Data Engineering. 2006: 3.
- [7] LI N, LI T, VENKATASUBRAMANIAN S. T-Closeness: privacy beyond k-anonymity and l-diversity[C]//2007 IEEE 23rd International Conference on Data Engineering. 2007.
- [8] DWORK C. Differential privacy[C]//International Colloquium on Automata, Languages, and Programming. Berlin: Springer, 2006.
- [9] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography Conference. Berlin: Springer, 2006.
- [10] XU J, ZHANG Z, XIAO X, et al. Differentially private histogram publication[C]//IEEE International Conference on Data Engineering. 2012.
- [11] JAGADISH H V, KOUDAS N, MUTHUKRISHNAN S, et al. Optimal histograms with quality guarantees[C]//VLDB. 1998: 275-286.
- [12] ZHANG X, CHEN R, XU J, et al. Towards accurate histogram publication under differential privacy[M]//Proceedings of the 2014 SIAM International Conference on Data Mining. 2014.
- [13] 张啸剑, 邵超, 孟小峰. 差分隐私下一种精确直方图发布方法[J]. 计算机研究与发展, 2016, 53(5): 1106-1117.
- [14] ZHANG X J, SHAO C, MENG X F. Accurate histogram release under differential privacy[J]. Journal of Computer Research and Development, 2016, 53(5): 1106-1117.
- [15] HUI L, CUI J, LIN X, et al. Improving the utility in differential private histogram publishing: theoretical study and practice[C]//IEEE International Conference on Big Data. 2017.
- [16] 唐海霞, 杨庚, 白云璐. 一种自适应差分隐私预算分配策略的直方图发布算法[J]. 计算机应用研究, 2020(1): 1-8.
- [17] TANG H X, YANG G, BAI Y L. Histogram publishing algorithm based on adaptive privacy budget allocation strategy under differential privacy [J]. Application Research of Computers, 2020(1): 1-8.
- [18] BKAKRIA A, TASIDOU A, CUPPENS-BOULAHIA N, et al. Optimal distribution of privacy budget in differential privacy[C]//Risks and Security of Internet and Systems(CRISIS) 2018. 2018.
- [19] 杨军. 一种基于经验分布函数的不等组距直方图的绘制方法 [P]. 2015.
- [20] YANG J. A method for drawing unequal spacing histograms based on empirical distribution functions [P]. 2015.
- [21] ZHU T, LI G, ZHOU W, et al. Differentially private data publishing and analysis: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2017: 1-1.
- [22] DWORK C, ROTH A. The algorithmic foundations of differential

privacy[J]. Theoretical Computer Science, 2013, 9(3-4): 211-407.

- [18] 张建方, 王秀祥. 直方图理论与最优直方图制作[J]. 应用概率统计, 2009, 25(2): 201-214.

ZHANG J F, WANG X X. Histogram theories and optimal histogram construction algorithms[J]. Chinese Journal of Applied Probability and Statistics, 2009, 25(2): 201-214.

[作者简介]



杨磊（1995-），女，安徽池州人，安徽工业大学硕士生，主要研究方向为差分隐私与大数据。



郑嘯（1975-），男，福建莆田人，博士，安徽工业大学教授、硕士生导师，主要研究方向为计算机网络、服务计算与云计算、安全与隐私保护。



赵伟（1988-），男，安徽六安人，博士，安徽工业大学副教授、硕士生导师，主要研究方向为无线网状网络、移动自组织网络、边缘计算。