基于差分隐私的人口普查关联多属性数据发布

尤菲芙, 蔡剑平, 孙岚

引用本文

尤菲芙, 蔡剑平, 孙岚. 基于差分隐私的人口普查关联多属性数据发布[J]. 计算机科学, 2024, 51(3): 368-377.

YOU Feifu, CAI Jianping, SUN Lan. Census Associated Multiple Attributes Data Release Based on Differential Privacy [J]. Computer Science, 2024, 51(3): 368-377.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230100013

基于差分隐私的人口普查关联多属性数据发布

尤菲芙　蔡剑平　孙　岚
福州大学计算机与大数据学院　福州 350108
(youfeifu97@163.com)

摘　要　发布未经保护的人口普查统计数据有泄露居民个人隐私信息的风险。基于差分隐私的人口普查数据保护方案已经得到研究者的广泛关注。在解决人口普查统计数据的地理区域之间的一致性约束时,具有更复杂层次性、一致性约束的关联多属性数据在现有方法下面临无法在单棵层次树中构建的挑战。文中提出了一种基于差分隐私的人口普查区域内部关联多属性统计数据最优一致发布方法,该方法能够实现复杂一致性约束统计数据的高效发布。首先将复杂的关联多属性之间的一致性约束划分为相对独立且易于求解的多重一致性约束,然后根据人口普查关联多属性数据的结构特性,通过数学分析在现有方法的基础上进行进一步的效率优化,最后结合多重一致性约束问题的逼近方法实现最优一致发布。在真实的人口普查数据集和合成数据集上进行实验,结果表明,所提方法能够在效率表现上优于同类方法1～2个数量级的同时保持与同类方法一致的精度。

关键词:差分隐私;隐私保护;数据发布;一致性约束;人口普查
中图分类号　TP309

Census Associated Multiple Attributes Data Release Based on Differential Privacy

YOU Feifu,CAI Jianping and SUN Lan
College of Computer and Data Science,Fuzhou University,Fuzhou 350108,China

Abstract　The release of unprotected census statistics carries the risk of revealing residents' personal privacy information. Census data protection solutions based on differential privacy have received substantial attention from researchers. Existing methods address the consistency constraint among geographic regions of census statistics,but associated multi-attribute data with more complex hierarchical consistency constraints face the challenge of being unable to build in a single hierarchical tree under existing methods. In this paper,we propose a differentially privacy method for optimally consistent release of associated multiple attributes statistics within census regions,which can achieve efficient release of statistics with complex consistency constraints. Firstly,the consistency constraints among the complex associated multiple attributes are divided into relatively independent and easily solved multiple consistency constraints. Then,based on the structural characteristics of the census associated multiple attributes data,mathematical analysis is used to further optimize the efficiency based on the existing methods. Finally,the optimal consistent release is achieved by combining the approximation method of the multiple consistency constraints problem. Experiments on real census datasets and synthetic datasets show that the proposed method can outperform similar methods in efficiency performance by one to two orders of magnitude while maintaining the same accuracy as similar methods.

Keywords　Differential privacy,Privacy protection,Data release,Consistency constraints,Census

1　引言

人口普查是当今世界各国广泛采用的收集人口资料的一种最基本的科学方法。人口普查作为全国基本人口数据的主要来源,是国情国力调查的重要组成部分[1-2]。人口普查数据是来自各个行业的数据分析师开展数据分析工作的重要统计信息支撑,科学的数据分析有助于完善人口发展战略和政策体系,促进人口长期均衡发展,科学制定国民经济和社会发展规划。人口普查数据发布的实质是发布包含大量个人敏感信息的数据集的统计结果。在人口普查中,采集到的包含大量个人敏感信息的数据交由人口普查相关机构处理和汇总,最后以图表、数据集等多种形式发布统计结果。因此,人口普查数据发布与个人隐私产生了密切关联,现有研究表明,发布未经保护的人口普查数据会导致个人隐私的泄露。以美国的人口普查数据发布为例,在 Dinur 和 Nissim 的重建攻击下[3],美国人口普查局在2000年和2010年的人口普查中使用的发布模型暴露出了严重漏洞。随着个人隐私越来越受重视,有部分安全意识强的居民拒绝调查、回避调查,参与和配合意愿下降[4]。因此,在发布人口普查数据时,必须结合有效的隐私保护措施来保护居民的个人隐私信息。

到稿日期:2023-01-02　返修日期:2023-04-26
通信作者:蔡剑平(jpingcai@163.com)

差分隐私作为数据发布中主流的隐私保护技术，因可以抵御最大背景知识下的差分攻击而得到安全性的认可，主要通过数据干扰来达到隐私保护的目的，并且在严格的隐私定义下，隐私保护水平也能得以量化。在基于差分隐私的数据发布中，层次树结构上的一致性约束问题得到了广泛研究。约束推理[5]通过对树的两次扫描来解决一致性约束问题并提高发布精度，但局限于二叉树结构；PrivTrie[6]提出了一种适用于任意树结构的最优一致发布方法，但复杂的递归函数调用面临大规模发布下计算开销随节点数线性增长的挑战；基于生成矩阵的 GMC[7]可以高效模拟任意树结构上的最优一致递归发布方法，是单棵层次树上一致性约束问题的理想解决方案。

在人口普查数据发布这样的大数据应用场景中，差分隐私模型构建便捷，能获得更高的开发效率，美国人口普查局在 2020 年全国人口普查中就应用了差分隐私技术对数据进行保护[8]。许多在人口普查数据发布中应用差分隐私的方法被陆续提出[9-11]，但这些方法主要聚焦于单一属性或独立多属性的一次或二次统计数据的发布，解决的是基于地理位置的单棵层次树上的一致性约束问题。然而，在实际的人口普查数据发布中，数据分析师们不仅关注某地区单一属性的人口情况（如女性人口数、未成年人口数），也关注该地区复合属性的人口数据（如未成年女性人口数）。单一属性和复合属性之间存在复杂的关联：复合属性必然关联若干个单一属性，单一属性也可以被若干个复合属性关联。我们统称这些有复杂关联的属性为关联多属性，与之对应的人口普查统计数据为关联多属性数据。在“多对多”的关联下，关联多属性数据间的一致性约束呈现的层次性比树结构更复杂，因此我们无法直接为关联多属性数据构建层次树，现有的基于层次树的一致性发布方法也无法直接适用于关联多属性数据的一致性发布。文献[12]提出了多重一致性约束问题的逼近方法，受逼近思想启发，若可以将复杂的层次性一致性约束划分为理想的层次树结构上的一致性约束，复杂的关联多属性数据的一致性发布问题就可以转化为对层次树结构上的子问题的依次、反复求解。但在多重一致性逼近中，反复多次的迭代会产生大量计算开销，单棵层次树上的发布方法 GMC[7]也难以保证计算效率，因此还需要利用关联多属性数据的特定层次结构进行进一步的效率优化。

基于以上分析，本文研究了差分隐私保护下的人口普查关联多属性数据发布问题，考虑了关联多属性数据中隐含的多重一致性约束并对其进行形式化定义，然后基于 GMC 方法提出了子问题的快速求解方法 FGMC，并在真实和合成数据集上进行了验证。

本文的主要贡献如下：

1）将关联多属性数据中复杂的层次性一致性约束问题划分为相对独立且易于求解的层次树结构上的一致性约束子问题，通过对子问题的依次、反复求解来获得原问题的最优解，实现了多重一致性约束问题的逼近方法在人口普查关联多属性数据发布问题中的应用。

2）提出了一种关联多属性数据的层次树结构上的最优一致发布的快速求解方法 FGMC（Fast Generation Matrix-based Optimally Consistent Release），通过分析 GMC 的计算过程和关联多属性数据的特征，利用数学推理得到更高效的求解表达式。

3）在真实的人口普查数据集及合成数据集上全面验证了 FGMC 的性能，并分析了多重一致性逼近下平均迭代轮数的变化情况和效率表现。实验结果表明，FGMC 与同类方法有相同的精度表现，更重要的是，FGMC 在所有情况下都能有更好的效率表现。

2 相关工作

在差分隐私数据发布方面，Hay 等[5]提出了基于层次树结构的约束推理方法，有效提高了发布的精度。Qardaji 等[13]随后分析了约束推理方法，分析表明，将方法拓展到更多维度时，约束推理的优势明显减弱。Ding 等[14]研究了差分隐私下基于多维表的数据立方体，通过对噪声的优化以及一致性的应用提高了数据立方体的精度。Li 等[15]提出了一种采用桶划分的回答差分隐私下范围查询的方法 DAWA。Cormode 等[16]研究了由树结构索引的空间数据发布。随后，Zhang 等[17]也采用了分区的方案将隐私预算划分到不同层次上，进一步提高了隐私保护数据的准确性；Shaham 等[18]提出了一种基于拉普拉斯机制的同质性驱动的空间划分方案，以提高位置数据发布的精度；Li 等[19]提出了一种基于四叉树的混合分解算法 DP-HDAQT，在数据密集区域采用改进的四叉树分解方法来保护空间隐私数据发布。Li 等[20]提出了矩阵机制来解决线性计数查询下的发布问题。Mckenna 等[21]提出了一种差分隐私下高维数据查询的优化方法 HDMM，利用高维矩阵机制有效降低了查询的误差。最近，Cardoso 等[22]提出了应用于连续观测下针对未知限制域和已知无限制域的差分隐私直方图发布算法。Zhu 等[23]提出了一种差分隐私直方图发布方案 CompressTree，该方案结合了分层直方图和直方图压缩，使用粗划分和动态预算分配的压缩方法，有效降低了敏感度，提高了范围查询的准确性，但该方案针对的是单属性数据集。

在一致性后处理方面，Hay 等提出的约束推理方法通过对层次树的两次扫描来恢复其一致性，但约束推理仅支持在满 k 叉树上的计算，限制了其更广泛的应用。Wang 等[6]在有效频繁挖掘工作中提出了适用于任意树的最优一致性发布方法 PrivTrie，但它的实现是基于多个复杂的递归算法，大量的函数调用给算法增加了额外的计算开销。Lee 等[24]将一致性后处理表示为有约束的 L_1 极小化问题，结合噪声分布特性，提出了极大似然后处理算法 ADMM，有效提升了发布精度，但 ADMM 在实际应用中的计算开销大，不适用于大规模数据发布场景。

将差分隐私应用在人口普查数据发布中是美国人口普查局首先提出的[8]，他们设计了基于层次结构的 TopDown 算法并将其应用在 2020 年的人口普查中[10]。该算法首先获取满足隐私要求的噪声计数，然后在地理层次结构中从上到下递归地执行一致性后处理，每次只在相邻的两个层上执行，

直到到达最小的地理区域。Kuo 等[11]研究了基于差分隐私的层次化人口普查数据的二次计数发布问题，提出了新的误差度量标准以及基于等值回归和最优加权匹配的差分隐私解决方案。Fioretto 等[9]对二次计数发布问题做了进一步的研究，利用分层性质、目标函数的结构和动态规划机制，在多项式时间内实现了差分隐私下的发布机制，提高了发布效率。这些方法关注的都是基于地理位置的区域之间的单棵层次树上的一致性约束关系，研究单一属性或独立多属性的一次或二次计数发布，不能直接适用于关联多属性数据发布问题。

3 预备知识

3.1 差分隐私

差分隐私作为一种具备严格形式化定义、强隐私性保证的安全技术，已经被广泛应用于数据发布场景中。

定义 1(ε-差分隐私[25]) 对于至多相差一条记录的邻近数据集 D 和 D'，如果一个随机算法 $\boldsymbol{A}: D \rightarrow Range(\boldsymbol{A})$ 及其所有可能的输出 $O \in Range(\boldsymbol{A})$ 满足以下条件，则称该算法满足 ε-差分隐私：

$$\Pr[\boldsymbol{A}(D)=O] \leqslant \exp(\varepsilon) \times \Pr[\boldsymbol{A}(D')=O] \tag{1}$$

其中，隐私保护程度由隐私预算 ε 决定，较小的隐私预算可以提供更高的安全性，但相对地会削弱数据的可用性。隐私预算的选择需要在两者之间平衡。

在实际应用中，拉普拉斯机制是实现差分隐私的一种基础机制，适用于数值型数据的保护。

定理 1(拉普拉斯机制[26]) 对于 D 上的随机函数 f，若算法 $\boldsymbol{A}$ 满足：

$$\boldsymbol{A}(D)=f(D)+Lap\left(\frac{\Delta q}{\varepsilon}\right) \tag{2}$$

则称算法 $\boldsymbol{A}$ 满足 ε-差分隐私。其中，$Lap(\cdot)$ 表示满足拉普拉斯分布的随机噪声变量，噪声的强度取决于隐私预算 ε 和全局敏感度 Δq。区别于人为设置的隐私预算，全局敏感度由数据集特征定义。

定义 2(全局敏感度[26]) 对于邻近数据集 D 和 D'，随机函数 f 的全局敏感度的定义如下：

$$\Delta q=\max_{D \sim D'} \| f(D)-f(D') \|_1 \tag{3}$$

其中，$\| \cdot \|_1$ 为 L_1 范式。全局敏感度的含义是：在敏感数据表中添加/删除任一条记录，统计表记录随之改变的最大值。

差分隐私的一个重要性质是后处理性质，它为差分隐私算法后处理步骤的合理性提供了依据。

性质 1(后处理性质[27]) 若算法 $\boldsymbol{A}$ 为 D 上的一种 ε-差分隐私算法，则在算法 $\boldsymbol{A}$ 上增加一个后处理步骤的算法 $\boldsymbol{M}(D)=post(\boldsymbol{A}(D))$ 也满足 ε-差分隐私。

3.2 多重一致性约束问题的逼近方法

若复杂的一致性约束问题可以划分为 k 重子问题，设第 i 个一致性子问题的最优解为 $\bar{x}=f_i(\tilde{x})$，则有以下定理。

定理 2(k 重一致性问题的逼近方法[12]) k 重一致性问题的最优解为：

$$\lim_{t \rightarrow \infty}(f_k * \cdots * f_2 * f_1)^t(\tilde{x}) \tag{4}$$

其中，$*$ 为函数的复合运算符，即 $f_j * f_i(x)=f_j(f_i(x))$；t 表示函数的复合运算次数，即 $f^2(x)=f(f(x))$。

根据定理 2，可以将复杂的 k 重一致性问题的最优求解转化为对其一致性子问题最优解的依次、反复求解。

3.3 基于生成矩阵的最优一致发布方法

生成矩阵(Generation Matrix)是一种层次树的可嵌入矩阵表示，利用生成矩阵可以将层次树当作一个整体进行研究，大大降低了研究的复杂性，还可以在不访问局部结构的情况下模拟各种递归算法，并提供各种可解释的矩阵操作来支持层次树的研究。

定义 3(生成矩阵[7]) 设 $\mathcal{T}$ 为高度降序排列的非零加权层次树，节点 i 的权值为 w_i，边 $i \rightarrow f_i$ 的权值为 $w_{i \rightarrow f_i}$，则生成矩阵 $\boldsymbol{G}_{\mathcal{T}} \in \mathbb{R}^{n \times n}$ 第 i 行第 j 列的元素 g_{ij} 由式(5)定义：

$$g_{ij}=\begin{cases} w_i, & i=j \\ -w_{i \rightarrow f_i}, & i \rightarrow f_i \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

生成矩阵 $\boldsymbol{G}_{\mathcal{T}}$ 是一个下三角矩阵，且任意非零加权层次树的生成矩阵表示是唯一的。

4 解决方案

首先，本文使用的符号的含义如表 1 所列。

表 1 符号说明

Table 1 Symbols description

符号	描述
$\boldsymbol{x}$	待发布数据，$\boldsymbol{x}=(t,\boldsymbol{s},\boldsymbol{m})$
t	总数
$\boldsymbol{s}$	分类统计值，$\boldsymbol{s}=(\boldsymbol{s}^1,\boldsymbol{s}^2,\cdots,\boldsymbol{s}^k)$
$\boldsymbol{s}^i$	第 i 个属性的分类统计值，$\boldsymbol{s}^i=(s_1^i,s_2^i,\cdots,s_{n_i}^i)$
s_j^i	第 i 个属性的第 j 个分类的统计值
$\boldsymbol{m}$	微观统计值
$\boldsymbol{x}_i$	第 i 个属性关联的待发布数据，即第 i 棵子树关联的值，$\boldsymbol{x}^i=(t,\boldsymbol{s}^i,\boldsymbol{m})$
$\tilde{\boldsymbol{x}}$	噪声数据，$\tilde{\boldsymbol{x}}=\boldsymbol{x}+Lap()$
$\boldsymbol{H}$	属性分类的组合空间为$[n_1]$到$[n_k]$的笛卡尔乘积，$[n]=\{1,2,\cdots,n\}$
P_j^i	第 i 个属性的第 j 个分类
$\boldsymbol{z}$	k 个属性的分类的一种组合，$\boldsymbol{z}=(z_1,\cdots,z_i,\cdots,z_k)$
$\boldsymbol{m}(\boldsymbol{z})$	k 个属性的分类分别为 $P_{z_1}^1,P_{z_2}^2,\cdots,P_{z_k}^k$ 时的微观统计值，$s_j^i=\sum_{\boldsymbol{z}\in\boldsymbol{H}, z_i=j}\boldsymbol{m}(\boldsymbol{z})$
$\mathcal{T}_i$	第 i 棵子树
u	树 $\mathcal{T}_i$ 上的节点，按层序遍历依次编号为 $0,1,2,\cdots,n+n_i$
f_u	u 的父节点
$v(u)$	节点 u 对应的统计值
$d(u)$	不一致量，为节点 u 与其子节点值之和的差，$d(u)=v(u)-\sum_{f_e=u}v(e)$
d_s	第二层节点不一致量之和，$d_s=\sum_{u=1}^{n_i}d(u)$
$A(u)$	节点 u 的修正量
n_i	第 i 个属性的分类数，$n_i=\lvert\boldsymbol{s}^i\rvert$
n	分类组合数，$n=\prod_{i=1}^{k}n_i=\lvert\boldsymbol{m}\rvert$
k_1^i	子树根节点的出度，$k_1^i=n_i$，简记为 k_1
k_2^i	子树第二层节点的出度，$k_2^i=n/n_i$，简记为 k_2

4.1 问题定义

4.1.1 人口普查关联多属性数据发布

在人口普查中，普查员以户为单位采集包括姓名、年龄、性别等项目的居民个人信息，我们将具有统计意义的项目称作该居民的属性，属性的取值为各个分类，如“性别”属性的取值为“分类1：女性”和“分类2：男性”。属性值包含了大量的个人敏感信息，采集结束后，普查员将其录入数据库。

我们考虑这样一个人口普查数据库，它由一个初始的敏感数据表和若干个统计表组成，记录了普查区域内所有居民的信息。敏感数据表中的每一条记录存储了对应居民的普查信息：能够唯一标识居民的编号 $PersonID$ 和该居民具有统计意义的 k 个属性的属性值($Attr_1$，$Attr_2$，…，$Attr_k$)。如表2所列，属性1取值为分类 a 和分类 b，属性2取值为分类 α 和分类 β，以此类推。敏感数据表包含了大量个人隐私信息，是人口普查关联多属性数据发布中隐私保护的核心，该表并不会被发布以供数据分析师们研究，表中的任何一条与个人隐私相关的记录都不希望被攻击者获取。

表2 k-属性敏感数据

Table 2 k-attribute sensitive data

PersonID	$Attr_1$	$Attr_2$	…	$Attr_k$
1	分类 a	分类 α	…	分类 I
2	分类 b	分类 β	…	分类 I
3	分类 a	分类 β	…	分类 II
…	…	…	…	…
m	分类 b	分类 α	…	分类 II

表2列出了敏感数据表在不同维度统计的结果，这些结果是人口普查关联多属性数据最核心的内容，也是数据分析师们直接分析研究的对象。如图1所示，统计数据包含3类数据，它们相互联系和约束。首先计算所有最细粒度分类的统计数，得到微观统计数据；然后通过累加来计算各个属性下各分类的统计数，得到分类统计数据；最后进行 k 组累加操作，得到 k 个相等的总数数据，总数值相等是统计数据之间一致性约束的其中一个表现。

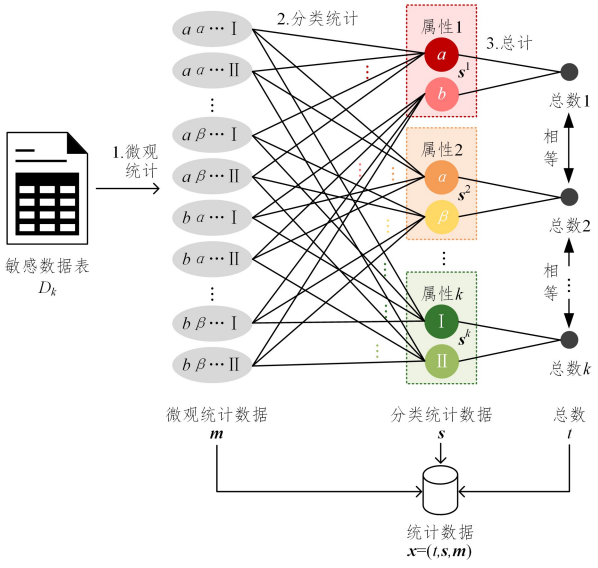


图1 k-属性统计数据构成

Fig. 1 Composition of k-attribute statistics

人口普查关联多属性数据发布的含义即以图表、数据集等形式发布普查区域内基于敏感数据表的统计表，要求发布的统计表具备如下特点：1)安全性，可以抵御包括重建攻击在内的攻击，以保证敏感数据表中的任何一条记录都无法被攻击者以任何手段获取；2)可用性，可以真实地体现敏感数据表的数据分布情况，以便数据分析师得出有效的结论，同时满足统计数据之间的一致性约束，保证不同类型的统计表之间不会互相矛盾。

4.1.2 人口普查关联多属性数据发布问题

考虑一个包含 k 个属性的敏感数据表 D_k，第 i 个属性的分类数为 n_i，总分类数 $n=\prod_{i=1}^{k} n_i$，基于 D_k 的微观统计值为 $\boldsymbol{m}$，分类统计值为 $\boldsymbol{s}$，总数为 t。其中，1)微观统计值 $\boldsymbol{m}$ 为 k 维张量，记$[n]=\{1,2,\cdots,n\}$，$\boldsymbol{m}$ 各维的组合空间 $\boldsymbol{H}$ 定义为$[n_1]$到$[n_k]$的笛卡尔乘积，即 $\boldsymbol{H}=\prod_{i=1}^{k}[n_i]$(以属性数 $k=2$，分类数 $n_1=2$，$n_2=2$ 为例，组合空间 $\boldsymbol{H}=\{(1,1),(1,2),(2,1),(2,2)\}$)，记 $\boldsymbol{H}$ 中的一种组合 $\boldsymbol{z}=(z_1,\cdots,z_i,\cdots,z_k)$，记第 i 个属性的第 j 个分类为 P_j^i，因此定义 $\boldsymbol{m}(\boldsymbol{z})$ 为 k 个属性的分类分别为 $P_{z_1}^1$，$P_{z_2}^2$，…，$P_{z_k}^k$ 时的微观统计值；2)分类统计值 s 细分为与 k 个属性对应的 k 组值，记第 i 个属性的分类统计值为 $\boldsymbol{s}^i$，有 $\boldsymbol{s}=(\boldsymbol{s}^1,\boldsymbol{s}^2,\cdots,\boldsymbol{s}^k)$，每组数据又由该属性下所有分类的统计值构成，记第 i 个属性的第 j 个分类的统计值为 $\boldsymbol{s}_j^i$，有 $s^i=(s_1^i,s_2^i,\cdots,s_{n_i}^i)$；3)$k$ 个相等的总数值简化为 t。

$\boldsymbol{m}$，$\boldsymbol{s}$ 和 t 之间的一致性约束表示为：

$$\begin{cases} t=\sum_{j=1}^{n_i} s_j^i, & \text{for } \forall i\in[1,k] \\ s_j^i=\sum_{\substack{\boldsymbol{z}\in\boldsymbol{H}\\ z_i=j}} m(\boldsymbol{z}), & \text{for } \forall i\in[1,k]\ \forall j\in[1,n_i] \end{cases} \tag{6}$$

设待发布的关联多属性数据 $\boldsymbol{x}=(t,\boldsymbol{s},\boldsymbol{m})$，对其应用拉普拉斯机制得到噪声数据 $\tilde{\boldsymbol{x}}$：

$$\tilde{\boldsymbol{x}}=\boldsymbol{x}+Lap\left(\frac{\Delta q}{\varepsilon}\right) \tag{7}$$

在人口普查关联多属性数据发布中，全局敏感度 $\Delta q=k+2$，即在敏感数据表中添加/删除任一条记录，最多会引起1项微观统计值的改变、k 项分类统计值的改变和1项总数的改变。

设 $\bar{x}$ 为最优一致性结果，我们定义人口普查关联多属性数据发布问题为如下优化问题：

$$\min_{\bar{\boldsymbol{x}}} \|\bar{\boldsymbol{x}}-\tilde{\boldsymbol{x}}\|_2 \quad \text{s.t. 式(6)} \tag{8}$$

4.2 子问题划分

根据图1和式(6)，人口普查关联多属性数据发布问题中的一致性约束呈现出复杂的层次性，直接求解比较困难。Cai等[12]的工作指出，可以将复杂的一致性问题划分为若干个易于求解的一致性子问题，然后通过逼近方法求得原问题的最优解。通过观察，可以发现每一组分类统计值 $\boldsymbol{s}^i$ 都与微观统计值 $\boldsymbol{m}$ 和总数 t 形成层次树上的一致性约束关系，即：

$$\begin{cases} t=\sum_{j=1}^{n_i} s_j^i \\ s_j^i=\sum_{\substack{\boldsymbol{z}\in\boldsymbol{H}\\ z_i=j}} \boldsymbol{m}(\boldsymbol{z}), \text{for}\ \forall j\in[1,n_i] \end{cases} \tag{9}$$

以属性数 $k=2$ 为例，设属性 1 分类的取值为 $\{a,b\}$，属性 2 分类的取值为 $\{\alpha,\beta,\delta\}$，分类统计值 s^1,s^2 与微观统计值和总数形成如图 2 所示的两棵层次树上的一致性约束关系。

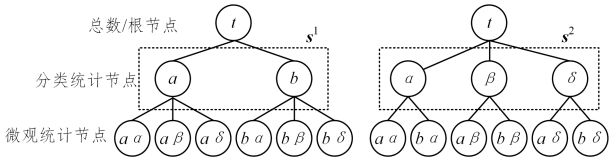


图 2 子问题划分的例子

Fig. 2 Example of subproblem division

通过这种重新构造，可以将 k-属性统计数据的复杂约束关系划分为 k 组子约束关系，每组子约束关系都形成单棵层次树结构，是理想的子问题约束关系。现在，我们给出子问题的形式化定义。

定义 4（人口普查关联多属性数据发布问题的子问题）设待发布的关联多属性数据的噪声版本 $\tilde{\boldsymbol{x}}_i=(t,\boldsymbol{s}^i,\boldsymbol{m})$，则将子问题 $g_i(\tilde{\boldsymbol{x}}_i)$ 定义为如下优化问题：

$$\min_{\bar{\boldsymbol{x}}_i}\|\bar{\boldsymbol{x}}_i-\tilde{\boldsymbol{x}}_i\|_2 \tag{10}$$

s. t. 式(9)

结合定理 2 和定义 4，原问题转化为以下逼近过程：

$$\lim_{t\to\infty}(g_k*\cdots g_2*g_1)^t(\tilde{\boldsymbol{x}}) \tag{11}$$

通过 k 重一致性问题的逼近方法，我们将具有 k 重一致性约束的人口普查关联多属性数据发布问题转化为对其子问题依次、反复的求解，也就是对树结构上的一致性问题进行重复求解。

由文献[7]可知，子问题存在如下闭式解：

$$\bar{\boldsymbol{x}}_i=g_i(\tilde{\boldsymbol{x}}_i)=\tilde{\boldsymbol{x}}_i-\boldsymbol{M}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathrm{T}}\tilde{\boldsymbol{x}}_i \tag{12}$$

其中，$\boldsymbol{M}$ 为满足 $\boldsymbol{M}\boldsymbol{x}_i^{\mathrm{T}}=\boldsymbol{0}$ 的层次树上的一致性约束矩阵，由以下表达式定义：

$$m_{pq}=\begin{cases}1, & p=q\\ -1, & q=f_p\\ 0, & \text{otherwise}\end{cases} \tag{13}$$

但式(12)求解的时间复杂度为 $O(n^3)$，不利于大规模数据发布的求解，因此需要探索一种更加高效的层次树上的一致性发布方法。

4.3 子问题的快速求解方法

在人口普查关联多属性数据发布子问题中，与子问题关联的树 $\mathcal{T}_i$ 是一棵层数恒为 3、根节点出度 k_1 恒为 n_i、内部节点（即第二层节点）出度 k_2 恒为 n/n_i 的特殊结构树。我们的目标是利用特殊结构树 T_i 设计一种高效的算法，用于完成一致性子问题的发布，称之为基于生成矩阵的三层完全树快速计算问题。

GMC 算法指基于生成矩阵的最优一致性发布算法（Generation Matrix-based Optimally Consistent Release），该算法利用生成矩阵的优势，将树结构转化为矩阵表示，在无需访问树的局部结构的前提下完成递归算法的模拟，这是针对目前任意树结构上的一致性约束问题的一种高效解决方案。

定理 3（基于生成矩阵的最优一致发布[7]） 设 $\tilde{\boldsymbol{v}}$ 为待发布噪声数据，$\mathcal{T}$ 为 $\tilde{\boldsymbol{v}}$ 对应的层次树结构，$\boldsymbol{M}_{\mathcal{T}}$ 为层次树的一致性约束矩阵，$\boldsymbol{G}_{\mathcal{T}}$ 为 $\mathcal{T}$ 的生成矩阵，那么 $\tilde{\boldsymbol{v}}$ 的最优一致发布 $\bar{\boldsymbol{v}}$ 如下：

$$\bar{\boldsymbol{v}}=\tilde{\boldsymbol{v}}-\boldsymbol{M}_{\mathcal{T}}(\boldsymbol{G}_{\mathcal{T}}^{-1}(\boldsymbol{G}_{\mathcal{T}}^{-\mathrm{T}}(\boldsymbol{M}_{\mathcal{T}}^{\mathrm{T}}\tilde{\boldsymbol{v}}))) \tag{14}$$

我们基于 GMC 算法展开优化工作。GMC 包含两个主要步骤：生成矩阵 $\boldsymbol{G}_{\mathcal{T}}$ 的构造和依据式(14)的最优解计算。我们将式(14)简记为：

$$\bar{\boldsymbol{v}}=\tilde{\boldsymbol{v}}-\boldsymbol{A} \tag{15}$$

然后将 $\boldsymbol{A}$ 的求解步骤详细拆解为以下 4 个子步骤：$\boldsymbol{\beta}^{(1)}=\boldsymbol{M}^{\mathrm{T}}\tilde{\boldsymbol{v}}$；$\boldsymbol{\beta}^{(2)}=\boldsymbol{G}^{-\mathrm{T}}\boldsymbol{\beta}^{(1)}$；$\boldsymbol{\beta}^{(3)}=\boldsymbol{G}^{-1}\boldsymbol{\beta}^{(2)}$；$\boldsymbol{A}=\boldsymbol{M}\boldsymbol{\beta}^{(3)}$。接下来对生成矩阵 $\boldsymbol{G}_{\mathcal{T}}$ 的构造和以上 4 个核心步骤进行针对 $\mathcal{T}_i$ 结构的优化推导。

1）生成矩阵 $\boldsymbol{G}_{\mathcal{T}}$ 的构造

首先初始化 $\boldsymbol{\theta}\in\mathbb{R}^{1\times(k_1+1)}$ 为非叶节点的子节点个数，$\boldsymbol{\theta}=(k_1,k_2,k_2,\cdots,k_2)$，然后自底向上遍历所有非根非叶节点 i，更新 f_i，即对根节点对应的 θ_0 进行 k_1 次更新，$\theta_0=k_1-k_1/(k_2+1)$，有：

$$\boldsymbol{\theta}=\left(k_1-\frac{k_1}{k_2+1},k_2,k_2,\cdots,k_2\right) \tag{16}$$

最后，根据 $\boldsymbol{\theta}$ 构造生成矩阵 $\boldsymbol{G}\in\mathbb{R}^{(k_1+1)\times(k_1+1)}$：

$$g_{ij}=\begin{cases}\sqrt{\theta_i+1}, & i=j\\ -\dfrac{1}{\sqrt{\theta_i+1}}, & i=f_i\\ 0, & \text{otherwise}\end{cases} \tag{17}$$

由式(16)、式(17)可得：

$$\boldsymbol{G}=\begin{bmatrix}\sqrt{1+(k_1-\frac{k_1}{k_2+1})} & & & \\ -\dfrac{k_1}{\sqrt{k_2+1}} & \sqrt{k_2+1} & & \\ & & \ddots & \\ -\dfrac{k_1}{\sqrt{k_2+1}} & & & \sqrt{k_2+1}\end{bmatrix} \tag{18}$$

$\mathcal{T}_i$ 的生成矩阵 $\boldsymbol{G}_{\mathcal{T}}$ 只在对角线和第一列有值，且只与 3 个值相关：

$$\begin{cases}G_r=g_{00}=\sqrt{1+\left(k_1-\dfrac{k_1}{k_2+1}\right)}\\ G_d=g_{ii}=\sqrt{k_2+1}\\ G_0=g_{i0}=-\dfrac{1}{\sqrt{k_2+1}}\end{cases} \tag{19}$$

2）$\boldsymbol{\beta}^{(1)}=\boldsymbol{M}^{\mathrm{T}}\tilde{\boldsymbol{v}}$

根据 $\boldsymbol{M}$ 的定义，$\boldsymbol{\beta}^{(1)}$ 为非叶节点的不一致量，我们将结果记为：

$$\boldsymbol{d}=\boldsymbol{\beta}^{(1)}=(d_0,d_1,\cdots,d_{k_1}) \tag{20}$$

并且记 $d_s=\sum_{i=1}^{k_1}d_i$。

3）$\boldsymbol{\beta}^{(2)}=\boldsymbol{G}^{-\mathrm{T}}\boldsymbol{\beta}^{(1)}$

性质 2（生成矩阵的向上传播） 设 g_{ij} 为生成矩阵 $\boldsymbol{G}$ 的第 i 行第 j 列，$\boldsymbol{x}$ 为 $\mathcal{T}$ 对应的待发布向量，那么 $\boldsymbol{z}=\boldsymbol{G}^{\mathrm{T}}\boldsymbol{x}$ 是在 $\boldsymbol{x}$ 上的向上传播，z_i 满足：

$$z_i=\begin{cases}x_i/g_{ii}, & i\text{ 为叶节点}\\ (x_i-\sum\limits_{j\in C_i}g_{ji}z_j)/g_{ii}, & \text{otherwise}\end{cases}\tag{21}$$

根据性质 2，i 为内部节点时：

$$\beta_i^{(2)}=\frac{d_i}{g_{ii}}=\frac{d_i}{G_d}\tag{22}$$

i 为根节点时：

$$\begin{aligned}\beta_i^{(2)}&=\frac{d_i-\sum\limits_{j\in C_i}g_{ji}\beta_j^{(2)}}{g_{ii}}=\frac{d_0-\sum\limits_{j\in C_0}g_{j0}\beta_j^{(2)}}{g_{00}}\\&=\frac{d_0-\sum\limits_{j\in C_0}G_0\frac{d_j}{G_d}}{G_r}=\frac{d_0-\frac{G_0}{G_d}\sum\limits_{j\in C_0}d_j}{G_r}\\&=\frac{d_0-\frac{G_0}{G_d}d_s}{G_r}=\frac{d_0G_d-G_0d_s}{G_rG_d}\\&=\frac{d_0G_d^2+d_s}{G_rG_d^2}\end{aligned}\tag{23}$$

即：

$$\beta_i^{(2)}=\begin{cases}\frac{d_i}{G_d}, & i\text{ 为内部节点}\\ \frac{d_0G_d^2+d_s}{G_rG_d^2}, & i\text{ 为根节点}\end{cases}\tag{24}$$

4)$\boldsymbol{\beta}^{(3)}=\boldsymbol{G}^{-1}\boldsymbol{\beta}^{(2)}$

性质 3（生成矩阵的向下传播） 设生成矩阵 $\boldsymbol{G}$ 和待发布向量 $\boldsymbol{x}$ 与性质 2 有相同定义，那么 $\boldsymbol{z}=\boldsymbol{G}^{-1}\boldsymbol{x}$ 是在 $\boldsymbol{x}$ 上的向下传播，z_i 满足：

$$z_i=\begin{cases}x_i/g_{ii}, & i\text{ 为根节点}\\ (x_i-g_{if_i}z_{f_i})/g_{ii}, & \text{otherwise}\end{cases}\tag{25}$$

根据性质 3，i 为根节点时：

$$\beta_i^{(3)}=\frac{\beta_i^{(2)}}{g_{ii}}=\frac{\frac{d_0G_d^2+d_s}{G_rG_d^2}}{G_r}=\frac{d_0G_d^2+d_s}{G_r^2G_d^2}\tag{26}$$

i 为内部节点时：

$$\begin{aligned}\beta_i^{(3)}&=\frac{\beta_i^{(2)}-g_{if_i}\beta_{f_i}^{(3)}}{g_{ii}}=\frac{\frac{d_i}{G_d}-G_0\frac{d_0G_d^2+d_s}{G_r^2G_d^2}}{G_d}\\&=\frac{d_i}{G_d^2}+\frac{d_0}{G_r^2G_d^2}+\frac{d_s}{G_r^2G_d^4}\end{aligned}\tag{27}$$

即：

$$\beta_i^{(3)}=\begin{cases}\frac{d_i}{G_d^2}+\frac{d_0}{G_r^2G_d^2}+\frac{d_s}{G_r^2G_d^4}, & i\text{ 为内部节点}\\ \frac{d_0G_d^2+d_s}{G_r^2G_d^2}, & i\text{ 为根节点}\end{cases}\tag{28}$$

5)$\boldsymbol{A}=\boldsymbol{M}\boldsymbol{\beta}^{(3)}$

根据 $\boldsymbol{M}$ 的定义，i 为根节点时：

$$A_i=\beta_i^{(3)}=\frac{d_0G_d^2+d_s}{G_r^2G_d^2}\tag{29}$$

i 为内部节点时：

$$\begin{aligned}A_i&=-\beta_0^{(3)}+\beta_i^{(3)}\\&=\frac{d_0G_d^2+d_s}{G_r^2G_d^2}+\left(\frac{d_i}{G_d^2}+\frac{d_0}{G_r^2G_d^2}+\frac{d_s}{G_r^2G_d^4}\right)\end{aligned}\tag{30}$$

i 为叶节点时：

$$A_i=-\beta_{f_i}^{(3)}=-\left(\frac{d_i}{G_d^2}+\frac{d_0}{G_r^2G_d^2}+\frac{d_s}{G_r^2G_d^4}\right)\tag{31}$$

我们将式(19)代回到式(31)，得到 A_i 关于 k_1，k_2 和 $\boldsymbol{d}$ 的表达式：

$$A_i=\begin{cases}\frac{d_0t+d_s}{r}, & i=0\\ \frac{d_i}{w}-k_2\times\left(\frac{d_0}{r}+\frac{d_s}{wr}\right), & 1\leqslant i\leqslant k_1\\ -\frac{d_{f_i}}{w}-\left(\frac{d_0}{r}+\frac{d_s}{wr}\right), & k_1+1\leqslant i\leqslant k_1k_2+k_1\end{cases}\tag{32}$$

其中，

$$\begin{cases}w=1+k_2\\ r=1+k_2+k_1k_2\end{cases}\tag{33}$$

至此，我们将 GMC 在 $\mathcal{T}_i$ 上的计算过程优化为一个仅关于 k_1，k_2 和 $\boldsymbol{d}$ 的表达式。基于式(32)和式(33)，我们提出了一种人口普查关联多属性数据发布子问题中优化的 GMC 算法 FGMC。

算法 1 基于生成矩阵的最优一致快速发布(FGMC)

输入：$\tilde{\mathbf{x}}_i$，k_1，k_2

输出：$\bar{\mathbf{x}}_i$

1. 初始化树节点集合 **U** 和节点值 **v**(u)；
2. //计算不一致量 **d** 和 d_s
3. for u$\notin$leaf do
4. **E**←child(u)；
5. d(u)←v(u)$-\sum$v(e)；//e$\in$**E**
6. end for
7. **U**′←**U**−leaf-root；
8. d_s←$\sum$d(u′)；//u′$\in$**U**′为内部节点
9. 根据式(33)计算 w，r；
10. 根据式(32)计算 A；
11. return $\bar{\mathbf{x}}_i=\tilde{\mathbf{x}}_i-\mathbf{A}$.

根据定理 1(拉普拉斯机制)，对关联多属性数据 $\boldsymbol{x}$ 应用拉普拉斯机制(见式(7))所得的噪声数据 $\tilde{\boldsymbol{x}}$ 满足 ε-差分隐私。FGMC 在 $\tilde{\boldsymbol{x}}$ 的基础上进行了一致性后处理，即 $FGMC(\tilde{\boldsymbol{x}}_i)=post(\tilde{\boldsymbol{x}})$。根据性质 1(后处理性质)可得：FGMC 也满足 ε-差分隐私。

5 实验结果与分析

5.1 实验环境与数据集

实验使用的硬件环境为 Intel(R) Core(TM) i7-6700 CPU @ 3.41GHz，运行内存为 16GB，在 Windows 操作系统上用 Python 编程语言实现所有方法。在多重一致性逼近中，收敛条件为10^{-6}。

实验使用了两个数据集：美国 2010 年人口普查数据集 Census2010 和随机生成的合成数据集 SynData。

Census2010 是美国人口普查局发布的关于 2010 年全美人口普查统计结果的数据集，统计粒度最小可至街区级别，数据集的列记录了每一个区域的基本信息以及在该区域范围内各属性各分类的统计结果。本文实验选取了 Census2010 中总人口数大于零的统计粒度为普查区的 449814 条数据，然后截取了与种族、性别和年龄 3 个属性相关的列。对于没有分类统计结果的属性，通过对微观统计数据的求和得到。Census2010 中，种族有 7 个分类，性别有 2 个分类，年龄有 23 个分类。实验使用的 4 个子集的详细信息如表 3 所列。

表 3 Census2010 子集的详细情况

Table 3 Details of Census2010 subset

子集名称	属性	总分类数	数据集规模
Census_k2_n14	性别、种族	2×7=14	449814
Census_k2_n46	性别、年龄	2×23=46	449814
Census_k2_n161	种族、年龄	7×23=161	449814
Census_k3_n322	性别、种族、年龄	2×7×23=322	449814

为了更全面地探究算法的性能，还随机生成了一个合成数据集 SynData。SynData 由 8 个子数据集构成，属性数 k 为 2～8，总分类数 n 设计为 2 的指数幂，为 $2^2 \sim 2^{13}$，最后一组特殊子数据集的总分类数 n 设计为 10 的指数幂，用于单独分析算法 FGMC 的性能。所有数据集的微观统计数据值在 0 到 max 间随机生成（max 在 1 到 500 间随机生成），然后通过求和得到分类统计数据和总数统计数据。详细信息如表 4 所列。

表 4 SynData 的详细情况

Table 4 Details of SynData

子集名称	属性数	总分类数	数据集规模
SynData_k2	2	$2^2 \sim 2^{13}$	12×10000
SynData_k3	3	$2^3 \sim 2^{13}$	11×10000
SynData_k4	4	$2^4 \sim 2^{13}$	10×10000
SynData_k5	5	$2^5 \sim 2^{13}$	9×10000
SynData_k6	6	$2^6 \sim 2^{13}$	8×10000
SynData_k7	7	$2^7 \sim 2^{13}$	7×10000
SynData_k8	8	$2^8 \sim 2^{13}$	6×10000
SynData_k2s	2	$10^1 \sim 10^7$	7×100

5.2 评价指标

本文采用算法的运行时间、多重一致性逼近过程中的迭代轮数和均方根误差 3 个指标来评价实验结果。

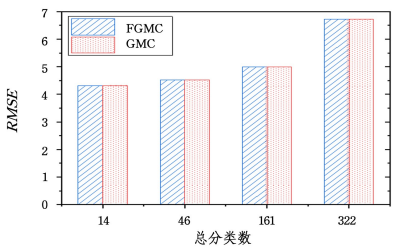
均方根误差（Root Mean Square Error）可以衡量隐私保护前后数据的差异程度，计算式如下：

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\bar{x}_i-x_i)^2} \tag{34}$$

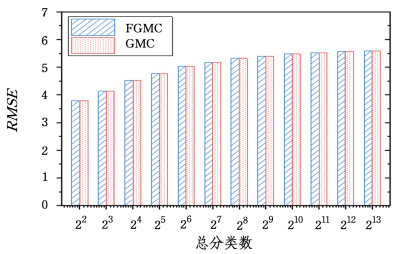
5.3 实验结果与分析

5.3.1 算法 FGMC 的正确性验证

给定相同的噪声数据 $\tilde{\boldsymbol{x}}$，在多重一致性逼近中，分别使用 FGMC 和 GMC 进行一致性处理，然后对比 RMSE 和迭代轮数，实验结果如图 3 和图 4 所示。



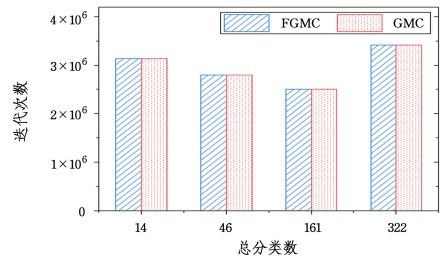
(a) Census2010



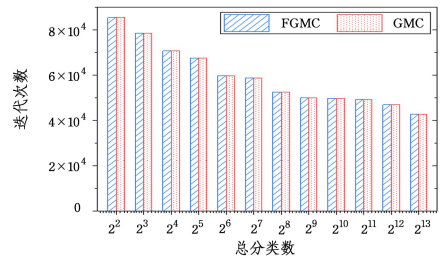
(b) SynData

图 3 FGMC 和 GMC 的 RMSE 对比

Fig. 3 Comparison of FGMC and GMC's RMSE



(a) Census2010



(b) SynData

图 4 FGMC 和 GMC 的迭代轮数对比

Fig. 4 Comparison of FGMC and GMC's iteration rounds

在 Census2010 和 SynData 中，RMSE 和迭代轮数都保持一致，由此可见，FGMC 是 GMC 的等效实现。

5.3.2 性能分析

首先，探究算法 FGMC 完成一次一致性处理的能力，受 Census2010 属性数和分类数的限制，实验在 SynData_k2s 上完成，隐私预算设置为 1，实验结果如图 5 所示。首先对比 GMC 和 PrivTrie，在总分类数较小时，GMC 构造矩阵所消耗的时间占比高，因此 PrivTrie 的表现更好。但随着总分类数的增加，PrivTrie 的运行时间随之线性增加，这时 GMC 的优越性得以体现，效率提升至 20 倍以上。然后分析 FGMC，得益于直接应用数学推导结论，FGMC 节省了矩阵初始化构造的时间开销，在 GMC 的基础上将运行时间在总体上优化了一个数量级，因此在所有总分类数的设置中，FGMC 都能够有同时超越 GMC 和 PrivTrie 的效率表现。

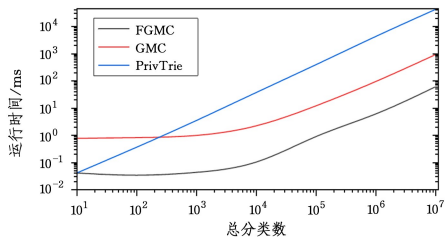
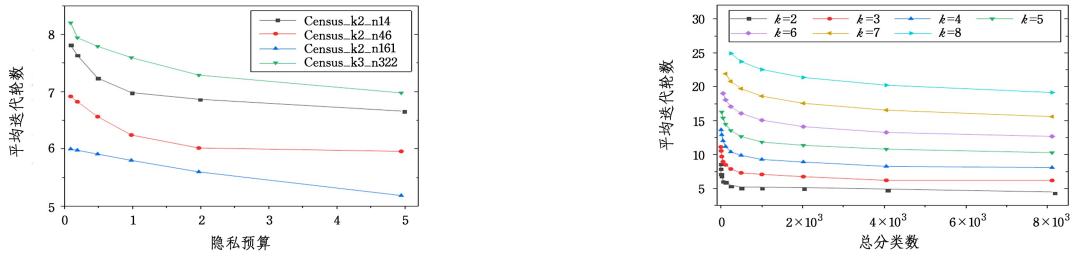


图 5 FGMC 与同类算法在 SynData_k2s 上的效率比较

Fig. 5 Efficiency comparison of FGMC with similar algorithms on SynData_k2s

接着，我们探究影响运行时间的另外一个因素，即多重一致性逼近过程中的迭代轮数。分别在 Census2010 和 SynData 上执行多重一致性逼近，隐私预算设置为 1，实验结果如图 6 所示。与直觉相反，随着总分类数的增加，平均迭代轮数呈现相反的变化趋势，这是因为 FGMC 的一致性处理的本质是将父子节点间的“不一致量”平均地分摊到子节点上，总分类数越多意味着每一个代表微观统计数据的节点分摊的“不一致量”越少，即变化量越小，在进入下一个属性的一致性处理时

“不一致量”变小，因此多重一致性逼近更快地达成了在所有 k 个属性上的一致，在实验中则体现为平均迭代轮数的减少。在总分类相同时，由于每一轮更新都要在 k 个属性上依次迭代，因此平均迭代轮数随着属性数的增长而增长。

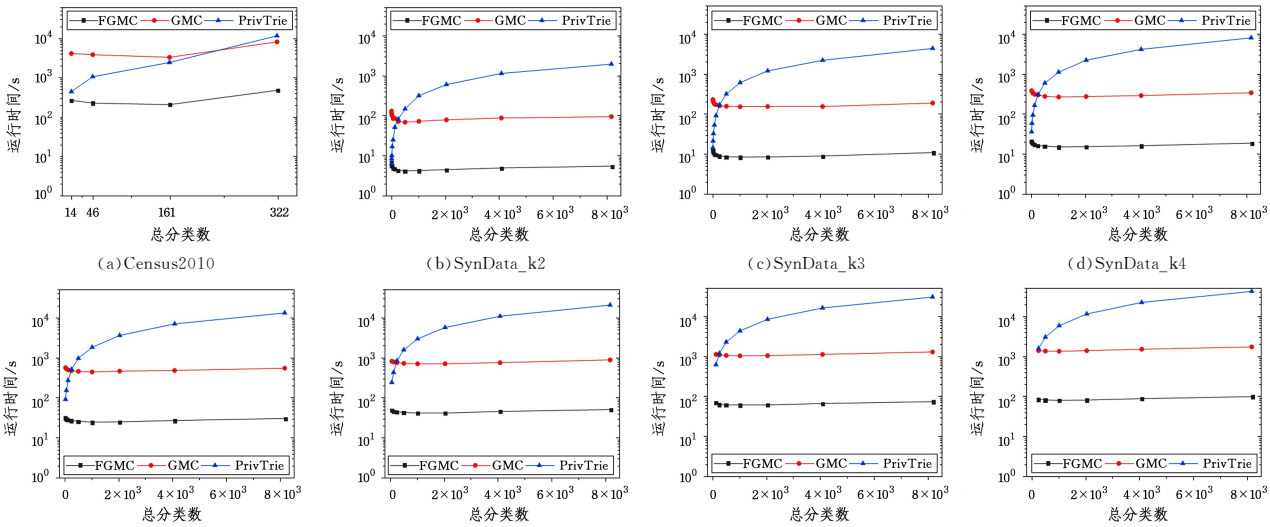


(a)Census2010　(b)SynData

图 6　多重一致性逼近下平均迭代轮数的变化情况

Fig. 6　Variation of the average iteration rounds with multiple consistency approximation

现在，我们分别将一致性处理算法设置为 FGMC，GMC 和 PrivTrie 来执行多重一致性逼近，记录运行时间，实验结果如图 7 所示。由 SynData 上的结果可知，得益于迭代轮数的降低，FGMC 和 GMC 曲线存在最小值，即在达到最小值以前，总分类数增加，算法效率反而更高；Census2010 上的结果也体现了这一点，当属性数为 2 时，总分类数由 14 增长至 46 和 161，程序运行时间逐步缩短，因为参考 SynData_k2 上的结果，极小值在总分类数为 512 时才出现。还可以观察到，FGMC 的运行时间比 GMC 快 16～19 倍，与 PrivTrie 相比，除了在属性数和总分类数都很小的情况下效率相当，在其他情况下 FGMC 总是比 PrivTrie 更高效，总分类数为 8 192 时，两者的效率相差 400 倍以上。



(a)Census2010　(b)SynData_k2　(c)SynData_k3　(d)SynData_k4

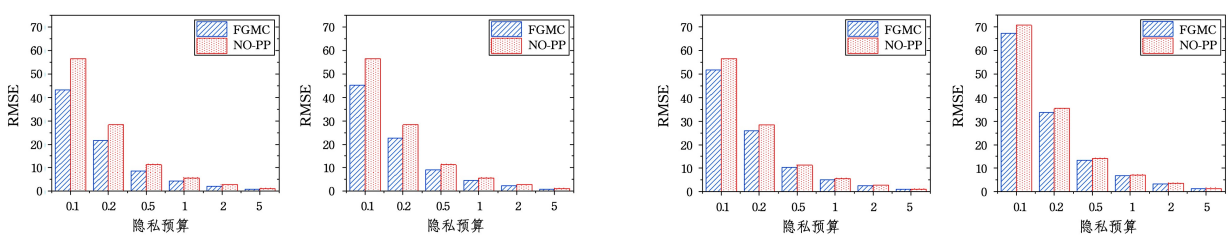
(e)SynData_k5　(f)SynData_k6　(g)SynData_k7　(h)SynData_k8

图 7　多重一致性逼近下 FGMC 与同类算法的效率比较

Fig. 7　Comparison of the efficiency of FGMC with similar algorithms with multiple consistency approximation

5.3.3　算法精度

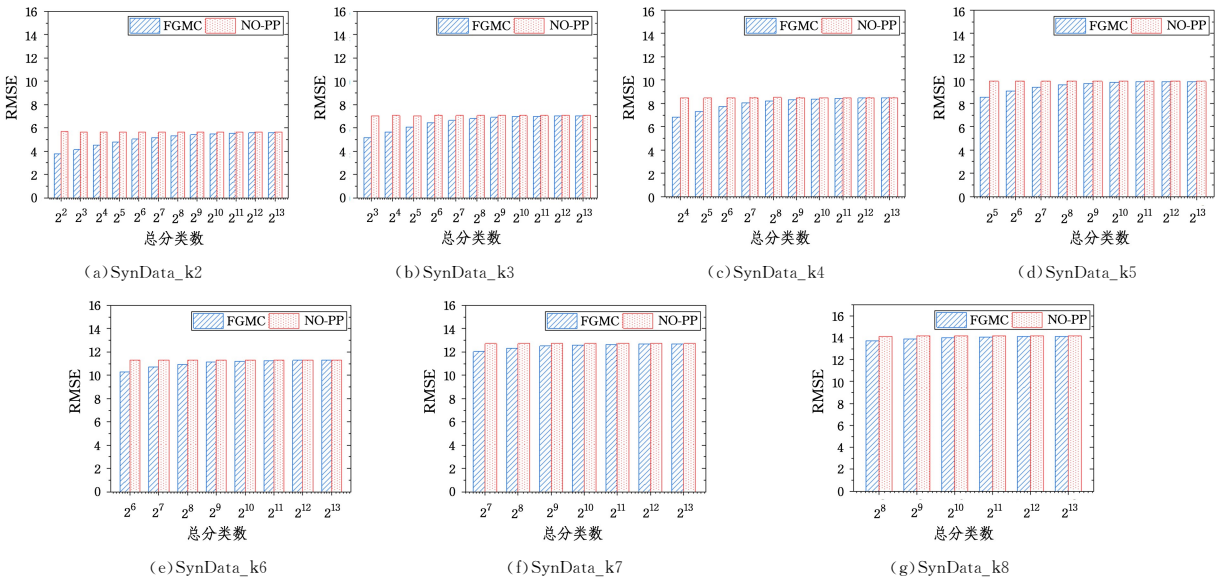
本文设计了消融实验来衡量算法对精度的提升。本实验引入了 NO-PP(NO Postprocessing)作为对比。该方法仅应用式(7)所示的拉普拉斯机制来保证隐私，而不执行一致性优化后处理。实验中，根据隐私预算对 Census2010 分别加入不同强度的噪声，SynData 部分的隐私预算统一设置为 1。实验结果如图 8 和图 9 所示，可以看出，首先，在 Census2010 和 SynData 的所有子集中，经过 FGMC 处理的数据总是拥有更低的 RMSE；其次，Census2010 上的结果表明 RMSE 随着隐私预算的增加而减小；最后，从 SynData 上的结果来看，RMSE 还会随着属性数的增加而变大，这是因为属性数的增加会提高发布的敏感度，因此需要加入更多的噪声以保证隐私。



(a)Census_k2_n14　(b)Census_k2_n46　(c)Census_k2_n161　(d)Census_k2_n322

图 8　多重一致性逼近下 FGMC 的精度表现(Census)

Fig. 8　Accuracy performance of FGMC with multiple consistency approximation(Census)



(a) SynData_k2 (b) SynData_k3 (c) SynData_k4 (d) SynData_k5

(e) SynData_k6 (f) SynData_k7 (g) SynData_k8

图 9 多重一致性逼近下 FGMC 的精度表现(SynData)

Fig. 9 Accuracy performance of FGMC with multiple consistency approximation(SynData)

结束语 针对更为复杂的关联多属性之间的一致性约束问题,本文结合多重一致性约束问题的逼近思想,划分出了多重一致性约束,利用数学推理优化子问题的求解方法,提出了基于差分隐私的多重一致性逼近下 FGMC 算法,目的在于解决人口普查数据发布中关联多属性一致性约束问题。通过真实人口普查数据集和合成数据集与现有的方法进行运行时间和 RMSE 的对比分析,表明了 FGMC 在保证相同精度的前提下能够显著提升发布效率。

在未来的工作中,我们将研究关联属性进一步扩展的人口普查关联多属性统计数据,并扩展约束集类型,以适应实际应用中的区域内任意范围查询的精度要求和其他实用性需求。

参 考 文 献

[1] NING J Z. Major figures on 2020 population census of China [J]. China Statistics, 2021(5): 4-5.

[2] HU Y, LI R. Method discussion on 2020 population census of China — based on the small census in 2015[J]. The World of Survey and Research, 2017(7): 51-54.

[3] DINURI, NISSIMK. Revealinginformationwhilepreservingprivacy [C] // Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. New York: Association for Computing Machinery, 2003: 202-210.

[4] CHEN W Q. Reform and development of the Chinese population census[J]. The World of Survey and Research, 2012(11): 48-52.

[5] HAY M, RASTOGI V, MIKLAU G, et al. Boosting the accuracy of differentially private histograms through consistency[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 1021-1032.

[6] WANG N, XIAO X, YANG Y, et al. PrivTrie: Effective Frequent Term Discovery under Local Differential Privacy[C] // 2018 IEEE 34th International Conference on Data Engineering (ICDE). Paris: IEEE, 2018: 821-832.

[7] CAI J P, LIU X M, LI J Y, et al. Generation Matrix: An Embeddable Matrix Representation for Hierarchical Trees[J]. arXiv: 2201.11297, 2022.

[8] ABOWD J M. The US Census Bureau adopts differential privacy [C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: Association for Computing Machinery, 2018: 2867-2867.

[9] FIORETTO F, VAN HENTENRYCK P, ZHU K. Differential privacy of hierarchical census data: An optimization approach [J]. Artificial Intelligence, 2021, 296: 103475.

[10] ABOWD J, ASHMEAD R, SIMSON G, et al. Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge [R/OL]. Washington: US Census Bureau, 2019. https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf.

[11] KUO Y H, CUIU C C, KIFER D, et al. Differentially private hierarchical count-of-counts histograms [J]. Proceedings of the VLDB Endowment, 2018, 11(11): 1509-1521.

[12] CAI J P, LIU X M, XIONG J B, et al. Approximation method of multiple consistency constraint under differential privacy [J]. Journal on Communications, 2021, 42(6): 107-117.

[13] QARDAJI W, YANG W, LI N. Understanding hierarchical methods for differentially private histograms[J]. Proceedings of the VLDB Endowment, 2013, 6(14): 1954-1965.

[14] DING B, WINSLETT M, HAN J, et al. Differentially private data cubes: optimizing noise sources and consistency[C] // Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: Association for Computing Machinery, 2011: 217-228.

[15] LI C, HAY M, MIKLAU G, et al. A Data-and Workload-Aware Algorithm for Range Queries Under Differential Privacy[J]. Proceedings of the VLDB Endowment, 2014, 7(5): 341-352.

[16] CORMODE G, PROCOPIUC C, SRIVASTAVA D, et al. Diffe-

rentially private spatial decompositions[C]// 2012 IEEE 28th International Conference on Data Engineering. Arlington:IEEE, 2012:20-31.

[17] ZHANG J,XIAO X,XIE X. Privtree:A differentially private algorithm for hierarchical decompositions[C]// Proceedings of the 2016 International Conference on Management of Data. New York:Association for Computing Machinery,2016:155-170.

[18] SHAHAM S,GHINITA G,AHUJA R,et al. HTF:Homogeneous Tree Framework for Differentially-Private Release of Location Data[C]// Proceedings of the 29th International Conference on Advances in Geographic Information Systems. New York: Association for Computing Machinery,2021:184-194.

[19] LI S,GENG Y,LI Y. A Differentially private hybrid decomposition algorithm based on quad-tree[J]. Computers & Security, 2021,109:102384.

[20] LI C,MIKLAU G,HAY M,et al. The matrix mechanism:optimizing linear counting queries under differential privacy[J]. The VLDB Journal,2015,24:757-781.

[21] MCKENNA R,MIKLAU G,HAY M,et al. HDMM:Optimizing error of high-dimensional statistical queries under differential privacy[J]. arXiv:2106.12118,2021.

[22] CARDOSO A R,ROGERS R. Differentially private histograms under continual observation:Streaming selection into the unknown[C]// International Conference on Artificial Intelligence and Statistics. New York:PMLR,2022:2397-2419.

[23] ZHU H,YIN F,PENG S,et al. Differentially private hierarchical tree with high efficiency[J]. Computers & Security,2022, 118:102727.

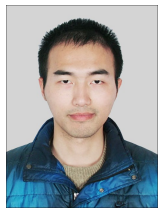
[24] LEE J,WANG Y,KIFER D. Maximum likelihood postprocessing for differential privacy under consistency constraints[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:Association for Computing Machinery,2015:635-644.

[25] DWORK C. Differential privacy[C]// Automata,Languages and Programming:33rd International Colloquium. Berlin:Springer, 2006:1-12.

[26] DWORK C,MCSHERRY F,NISSIM K,et al. Calibrating noise to sensitivity in private data analysis[C]// Theory of Cryptography:Third Theory of Cryptography Conference. Berlin:Springer,2006:265-284.

[27] DWORK C,ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends® in Theoretical Computer Science,2014,9(3/4):211-407.

YOU Feifu, born in 1997, postgraduate. Her main research interests include privacy protection and data security.



CAI Jianping, born in 1990, Ph.D. His main research interests include differential privacy, federated learning, machine learning and optimization theory.

(责任编辑:喻藜)